

UC Riverside

UC Riverside Electronic Theses and Dissertations

Title

Robust Multimodal Learning With Heterogeneous and Missing Modalities

Permalink

<https://escholarship.org/uc/item/5kj741gh>

ISBN

9798180113283

Author

Reza, Md Kaykobad

Publication Date

2026-05-01

Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NonCommercial-ShareAlike License, available at <https://creativecommons.org/licenses/by-nc-sa/4.0/>

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
RIVERSIDE

Robust Multimodal Learning With Heterogeneous and Missing Modalities

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Computer Science

by

Md Kaykobad Reza

June 2026

Dissertation Committee:

Dr. M. Salman Asif, Chairperson
Dr. Eamonn J. Keogh
Dr. Vassilis Tsotras
Dr. Ahmed Eldawy

The Dissertation of Md Kaykobad Reza is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

First, I would like to express my sincere gratitude to my advisor, Dr. M. Salman Asif, for his guidance, support, and patience throughout my doctoral journey. Beyond teaching me the fundamentals of research, his mentorship established the intellectual standards that continue to guide how I frame problems, evaluate evidence, and conduct scientific inquiry. I am deeply grateful for the opportunity to have been his student. I also thank my dissertation committee members: Dr. Eamonn J. Keogh, Dr. Vassilis Tsotras, and Dr. Ahmed Eldawy, for their insightful feedback, constructive critiques, and encouragement throughout my evaluations and defense, which were instrumental in refining the quality of this work.

My sincere thanks go to Mashhour Solh, Edward Ayrapietian, and Ameya Patil for their guidance and mentorship during my internship at Amazon. The opportunity to translate academic ideas into real-world systems under their supervision broadened my perspective and strengthened my approach to applied research. I also thank my collaborators, colleagues, and teachers at the University of California, Riverside, and beyond, whose discussions, shared ideas, and collaborations significantly contributed to both this work and my academic journey.

Finally, I owe a profound debt of gratitude to my family. I thank my parents for their lifelong encouragement, my wife for her unwavering support and sacrifices, and my daughter for the joy and inspiration she brings each day. This achievement would not have been possible without their love and belief in me.

To my parents, my wife, and my daughter,
for their unwavering support, encouragement, and inspiration.

ABSTRACT OF THE DISSERTATION

Robust Multimodal Learning With Heterogeneous and Missing Modalities

by

Md Kaykobad Reza

Doctor of Philosophy, Graduate Program in Computer Science
University of California, Riverside, June 2026
Dr. M. Salman Asif, Chairperson

Multimodal learning seeks to develop models capable of reasoning over heterogeneous inputs such as vision, language, audio, and sensor data. Despite significant progress in recent years, three key challenges remain: (i) supporting arbitrary combinations of input modalities, (ii) maintaining robustness when modalities are missing at inference time, and (iii) integrating capabilities from multiple expert models without costly joint retraining on large paired multimodal datasets. This dissertation addresses these challenges through a series of principled methods that progressively relax assumptions on modality availability, supervision, and data requirements.

First, we study multimodal fusion under heterogeneous modality combinations. We propose MMSFormer, a Multimodal Segmentation Transformer with a novel fusion strategy that effectively integrates information from diverse modalities. Unlike prior approaches limited to fixed modality sets, MMSFormer supports arbitrary modality combinations and exhibits monotonic performance improvements as additional modalities are introduced. Extensive experiments demonstrate state-of-the-art performance on multimodal material and semantic segmentation tasks, with ablation studies highlighting the contribution of each modality.

Second, we address robustness to missing modalities. We introduce two complementary approaches: (i) a parameter-efficient adaptation mechanism that modulates intermediate representations of pretrained models using fewer than 1% additional parameters, and (ii) Cross-Modal Proxy Tokens (CMPTs), which approximate missing modality representations by attending to available modalities without explicit modality generation. Both methods significantly reduce performance degradation across diverse tasks and settings while preserving strong performance when all modalities are present during inference.

Finally, we propose Singular Subspace Alignment and Merging (SSAM), a training-free framework for combining independently trained multimodal large language models into a unified model. SSAM aligns modality-specific representations within a shared low-rank subspace and merges them in parameter space to preserve complementary knowledge while minimizing modality interference. Without using any paired multimodal training data, SSAM achieves state-of-the-art performance across multiple benchmarks, surpassing prior training-free merging methods and, in some cases, jointly trained multimodal models.

Together, these contributions advance multimodal learning toward flexible, robust, and scalable systems capable of operating under real-world constraints.

Contents

List of Figures	xii
List of Tables	xvi
1 Introduction	1
1.1 Multimodal Learning	2
1.2 Problems with Multimodal Learning	3
1.2.1 Supporting Arbitrary Combinations of Input Modalities	3
1.2.2 Missing Modality Robustness	4
1.2.3 Merging Capabilities from Multiple Multimodal Models	4
1.3 Organization of The Book	5
2 MMSFormer: Multimodal Transformer for Material and Semantic Segmentation	7
2.1 Introduction	7
2.2 Related Work	11
2.3 Proposed Model	13
2.3.1 Overall Model Architecture	13
2.3.2 Modality Specific Encoder	14
2.3.3 Multimodal Fusion Block	15
2.3.4 Shared MLP Decoder	17
2.4 Experiments and Results	17
2.4.1 Datasets	18
2.4.2 Implementation Details	18
2.4.3 Performance Comparison with Existing Methods	20
2.4.4 Performance Comparison for Incremental Modality Integration	23
2.4.5 Qualitative Analysis of the Predictions	24
2.4.6 Ablation Study on the Fusion Block	26
2.4.7 Ablation Study on Different Modality Combinations	27
2.4.8 Computational Complexity of the Fusion Block	29
2.5 Conclusion	30

3	Robust Multimodal Learning with Missing Modalities via Parameter-Efficient Adaptation	31
3.1	Introduction	31
3.2	Related Work	34
3.3	Proposed Method	37
3.3.1	Network Adaptation for Missing Modalities	38
3.3.2	Parameter-Efficient Adaptation for Robust MML	40
3.4	Experiments and Results	45
3.4.1	Datasets and Tasks	45
3.4.2	Implementation Details	47
3.4.3	Experiments on Multimodal Segmentation	48
3.4.4	Experiments on Multimodal Sentiment Analysis	54
3.4.5	Experiments on Multimodal Action Recognition	56
3.4.6	Experiments on Multimodal Classification	57
3.4.7	Feature and Parameter Analysis	58
3.5	Limitations and Future Directions	60
3.6	Conclusion	61
4	Robust Multimodal Learning via Cross-Modal Proxy Tokens	62
4.1	Introduction	62
4.2	Related Work	65
4.3	Proposed Approach	67
4.3.1	Multimodal Embedding Tokens	68
4.3.2	Cross-Modal Proxy Tokens (CMPTs)	69
4.3.3	Learn CMPTs by Alignment and Adaptation	69
4.3.4	Gating Module and Fusion	70
4.3.5	Handling Missing Modalities	71
4.4	Experiments and Results	72
4.4.1	Datasets	72
4.4.2	Implementation Details	73
4.4.3	Performance when Modalities are Missing	74
4.4.4	Performance when All Modalities are Available	77
4.4.5	Generalization to Different Missing Rates	78
4.4.6	Ablation Studies	79
4.5	Limitations and Future Directions	84
4.6	Conclusion	85
5	SSAM: Singular Subspace Alignment for Merging Multimodal Large Language Models	86
5.1	Introduction	86
5.2	Related Work	90
5.3	Proposed Method	92
5.3.1	Problem Formulation	93
5.3.2	Language Vectors	94
5.3.3	Shared Consensus Subspace Construction	95

5.3.4	Language Vector Merging	97
5.4	Experiments and Results	98
5.4.1	Datasets and Benchmarks	98
5.4.2	Implementation Details	99
5.4.3	Results on AVQA Benchmarks	101
5.4.4	Results on MCUB Benchmarks	103
5.4.5	Ablation Studies	104
5.5	Discussion and Future Directions	107
5.6	Conclusion	108
6	Conclusion	110
6.1	Summary of Contributions	110
6.1.1	Flexible Multimodal Fusion	111
6.1.2	Robustness to Missing Modalities via Adaptation	111
6.1.3	Cross-Modal Approximation with Proxy Tokens	111
6.1.4	Training-Free Multimodal Model Merging	112
6.1.5	Unified Perspective	112
6.2	Open Problems and Future Directions	113
6.2.1	Generalization to New and Unseen Modality Combinations	113
6.2.2	Robustness Beyond Missing Modalities	113
6.2.3	Learning Without Paired Multimodal Training Data	114
6.2.4	Continual and Lifelong Multimodal Learning	114
6.2.5	Interpretability and Explainability	115
6.2.6	Toward General Multimodal Intelligence	115
6.3	Final Remarks	116
	Bibliography	117
	Appendices	138
A1	Robust Multimodal Learning with Missing Modalities via Parameter-Efficient Adaptation	138
A1.1	Datasets	138
A1.2	Implementation Details	140
A1.3	Number of Learnable Parameters	142
A1.4	Performance Comparison with Parameter Efficient Model Adaption Techniques	143
A1.5	Per Class IoU Comparison	147
A1.6	Cosine Similarity Analysis	149
A1.7	Effectiveness of Our Adaptation Method on Different Backbones	151
A1.8	Why Dedicated Model Performs Better on MFNet Dataset?	153
A1.9	Visualization of Predicted Segmentation Maps	154
A2	Robust Multimodal Learning via Cross-Modal Proxy Tokens	157
A2.1	Related Work on Multimodal Learning	157
A2.2	Detailed Dataset Description	158
A2.3	Implementation Details	160

A2.4	Determining the Optimal LoRA Rank	160
A2.5	Different Modality Ratios During Training and Inference	162
A2.6	Effectiveness of CMPTs	164
A2.7	Do CMPTs Benefit All Classes?	165
A2.8	Limitations and Failure Case Analysis	167
A2.9	Analysis of Modality Dependence	169
A2.10	Generalizability to Other Parameter-Efficient Adaptation Methods	173
A2.11	Applicability to SigLIP-Based Models	174
A2.12	Scalability with Larger Vision and Text Encoders	175
A3	SSAM: Singular Subspace Alignment for Merging Multimodal Large Language Models	177
A3.1	Datasets and Benchmarks	177
A3.2	Implementation Details	179
A3.3	Proposed SSAM Algorithm	181
A3.4	Selecting Optimal Subspace Rank (k)	182
A3.5	Qualitative Results	183

List of Figures

2.1	(a) Overall architecture of MMSFormer model. Each image passes through a modality-specific encoder where we extract hierarchical features. Then we fuse the extracted features using the proposed fusion block and pass the fused features to the decoder for predicting the segmentation map. (b) Illustration of the mix transformer [195] block. (c) Proposed multimodal fusion block.	8
2.2	Visualization of predictions on MCubeS and PST900 datasets. For brevity, we only show the RGB image and ground truth material segmentation maps along with the predictions in Figure 2.2a. Our model shows better predictions on both of the datasets.	22
2.3	Visualization of predicted segmentation maps for different modality combinations on MCubeS and FMB datasets. Both figures show that prediction accuracy increases as we incrementally add new modalities. They also illustrate the fusion block’s ability to effectively combine information from different modality combinations.	25
3.1	a) Overview of our model adaptation approach. A model pretrained on all the modalities is adapted using a small number of learnable parameters to handle different modality combinations. We insert adaptable layers after each layer of the encoders and the fusion block to learn the modulation as a function of the available input modalities to compensate for the missing modalities. The grayed-out branch (missing modality) is inactive and does not contribute to the output. b) Low-rank model adaption computes features using frozen weights and low-rank weight updates and combine them. c) Scale and shift adaptation transforms input by element-wise multiplication and addition.	32
3.2	Examples of predicted segmentation maps for the Pretrained and Adapted models. Title above each subimage shows method name (available modalities). Segmentation quality improves significantly after model adaptation for all input modality combinations. Green boxes highlight areas with salient differences in results. For MCubeS dataset, we only show RGB input images for brevity. A, D and N denote AoLP, DoLP, and NIR, respectively.	53
3.3	Cosine similarity between complete and missing modality features of pretrained and adapted models on MCubeS and NTU RGB+D datasets.	58

4.1	(a) We introduce Cross-Modal Proxy Tokens (CMPTs), a novel approach to address missing modality challenges. CMPTs effectively learn to approximate missing modality class tokens by adapting pretrained encoders through a joint optimization of alignment and task-specific objectives. Our approach accommodates both complete and missing modalities during training and inference, thereby enhancing robustness across varying missing modality scenarios. (b) CMPTs achieve state-of-the-art performance, consistently outperforming recent baseline methods in both complete and missing modality scenarios. Following the experimental setup of [94, 89], the radar plot illustrates F1-macro scores on the MM-IMDb dataset across varying modality availability.	64
4.2	Generalization to varying missing rates during inference. Models are trained with 100% image + 100% text and evaluated with 100% image + $x\%$ text following [123, 89]. Our approach demonstrates better generalization, particularly under severe modality loss.	77
4.3	Effectiveness of CMPTs on the MM-IMDb dataset. All models are trained with 100% image and 100% text data, then evaluated under varying amount of missing modalities. CMPTs achieve significant performance improvement, especially under severe missing modality scenarios.	78
4.4	CMPTs improve performance across most of the classes when modalities are missing. However, it fails to accurately predict certain classes (e.g., short/documentary) in the absence of text, likely due to the image modality’s insufficient information for those classes. Classes are sorted based on the amount of performance improvement.	80
4.5	t-SNE plots of the fused feature tokens $\mathcal{T}$ for ten classes under different modality settings. The t-SNE plots show that fused features without CMPTs (red) deviate significantly from the complete multimodal features (green). Incorporating CMPTs (blue) provides embeddings that closely align with the complete multimodal features, indicating improved semantic alignment and robustness in different missing modality scenarios.	82
4.6	Attention map visualizations from the last layer of the vision encoder for the CLS token and CMPTs in the text-missing scenario on the UPMC Food-101 dataset. The attention maps show that CLS tokens attend to semantically-irrelevant regions, which likely leads to incorrect predictions. In contrast, CMPTs attend to semantically-relevant regions for the respective classes and lead to accurate predictions.	83
5.1	Comparison of multimodal learning paradigms. (a) Conventional methods train a single MLLM on paired multimodal datasets. (b) ImageBind-LLM [58] aligns all modalities to shared image embeddings. (c) SSAM offers a training-free method to merge independently trained specialist MLLMs into a single, unified model via subspace projection.	87

5.2	Overview of the proposed SSAM method. (a) Subspace projection: each language vector Δ_i is projected onto a shared low-rank subspace prior to merging. (b) SSAM constructs covariance matrices A and B from all Δ_i , extracts the shared left and right bases U_c and V_c , projects language vectors, and merge them into Δ_{merged}	88
5.3	Qualitative examples on commonality understanding on the MCUB-4 dataset. . .	104
5.4	Qualitative examples on the AVQA dataset.	106
5.5	Comparison with simple model-merging baselines. SSAM consistently outperforms them across all four datasets.	106
5.6	Effect of subspace rank (k) on average performance. Empirically $k = 128$ offers slightly better performance.	106
A1.1	Cosine similarity between complete and missing modality features of the pre-trained model and complete and missing modality features of the adapted model under different missing modality scenarios on MCubeS dataset. The adapted model shows higher similarity to the complete modality features indicating less deviation and better handling of missing modalities.	149
A1.2	Cosine similarity between complete and missing modality features of the pre-trained model and the adapted model under different missing modality scenarios on NTU RGB+D dataset.	150
A1.3	Visualization of predicted segmentation maps for pretrained and adapted models on MFNet, NYUDv2, and MCubes datasets for multimodal semantic and material segmentation.	155
A2.4	Effectiveness of CMPTs in handling missing modalities on UPMC Food-101 dataset. CMPTs significantly improves performance when modalities are severely missing. Models are trained on 100% image + 100% text and tested with varying levels of missing modalities.	164
A2.5	CMPTs improve performance across most of the classes when <i>text</i> modality is missing on UPMC Food-101 dataset. However, it fails to accurately predict certain classes (e.g., bibimbap/sashimi), likely due to the image modality’s insufficient information for those classes. Classes are sorted based on the amount of performance improvement.	166
A2.6	CMPTs improve performance across most of the classes when <i>image</i> modality is missing on UPMC Food-101 dataset. However, it fails to accurately predict certain classes (e.g., lobster_roll_sandwich/cheese_plate), likely due to the text modality’s insufficient information for those classes. Classes are sorted based on the amount of performance improvement.	167
A2.7	t-SNE plots of the fused feature tokens $\mathcal{T}$ for ten classes where the performance gains from CMPTs are limited. In these cases, the fused feature embeddings with CMPTs (blue) remain diffuse and overlapping, and exhibit weak alignment with the complete multimodal features (green). Although CMPTs show improvement over the baseline without CMPTs (red), the alignment remains insufficient to form consistently discriminative clusters, indicating limited or no improvement in missing modality robustness in these classes.	168

A2.8	Attention map visualizations from the last layer of the vision encoder for the CLS token and CMPTs in the text-missing scenario on the UPMC Food-101 dataset. Left: Cases where the fused features with and without CMPTs provide correct predictions. Here, CLS tokens already attend to semantically-relevant regions and CMPTs provide further refinement, but do not change the prediction. Right: Cases where CMPTs attend to semantically-relevant regions but lead to incorrect predictions. The last example illustrates a scenario where both the fused features with and without CMPTs result in incorrect predictions. These examples highlight limitations of CMPTs when the available modality lacks sufficient discriminative information or when learned cross-modal associations are imprecise.	170
A2.9	Per class performance analysis on MM-IMDb dataset. Certain classes show strong dependence on a specific modality.	170
A2.10	Per class performance analysis on UPMC Food-101 dataset. Certain classes show strong dependence on a specific modality.	171
A3.11	Qualitative results on the MCUB-4 [14] dataset with four input modalities (image, video, audio, and point cloud) paired with natural-language queries. SSAM demonstrates strong cross-modal reasoning and provides coherent answers in complex multimodal settings, even though none of the specialist models were trained on this modality combination.	182
A3.12	Qualitative results on the MUSIC-AVQA [96] and AVQA [201] datasets with video and audio as the input modalities paired with natural-language queries. SSAM provides consistent responses, demonstrating stable performance on multimodal queries despite the specialist models not being trained on paired video-audio-language data.	183
A3.13	Qualitative results on the AVQA [201] dataset with video and audio as input modalities paired with natural-language queries. SSAM effectively combines information from both audio and video to produce more accurate and consistent responses, demonstrating stable performance on multimodal queries compared to individual specialist models.	184

List of Tables

2.1	Performance comparison on Multimodal Material Segmentation (MCubeS) dataset. Here A, D, and N represent AoLP, DoLP, and NIR respectively.	19
2.2	Performance comparison on FBM dataset. We show performance for different methods from already published works.	19
2.3	Per-class % IoU comparison on MCubeS dataset. MMSFormer shows better performance in detecting most of the classes compared to the current state-of-the-art models. * indicates that the code and pretrained model from the authors were used to generate the results.	20
2.4	Per-class % IoU comparison on FMB dataset. We show comparison for 8 classes (out of 14) that are published. MMSFormer outperforms existing methods for all the classes except truck.	21
2.5	Per-class % IoU and % mIoU comparison on PST900 dataset.	23
2.6	Performance comparison (% mIoU) on MCubeS dataset for different modality combinations. Here A, D, and N represent AoLP, DoLP, and NIR respectively. . .	24
2.7	Ablation study of the Fusion Block on FMB dataset with RGB and infrared modalities. The table shows the contribution of different modules in fusion block in overall model performance.	26
2.8	Per class % IoU comparison on MCubeS dataset for different modality combinations. As we add modalities incrementally, overall performance increases gradually. This table also shows that specific modality combinations assist in identifying specific types of materials better.	27
2.9	Comparison of number of parameters in the fusion block and GFLOPs for MCubeS dataset having 4 input modalities with an input shape of $(3 \times 512 \times 512)$ for each modality. Our fusion block shows significantly lower complexity compared to existing methods.	29
3.1	Performance comparison with different baseline methods for multimodal semantic segmentation on MFNet and NYUDv2 datasets and multimodal material segmentation on MCubeS dataset. We use CMNeXt as the base model. Bold letters represent best results.	48

3.2	Performance comparison with existing robust methods for MFNet dataset. RGB and Thermal columns report performance when only RGB and only Thermal are available. Average column reports average performance when one of the two modalities gets missing. ‘-’ indicates that results for those cells are not published. * indicates that available code and pretrained models from the authors were used to generate the results.	50
3.3	Performance comparison with existing robust methods for NYUDv2 dataset. RGB and Depth columns report performance when only RGB and only Depth are available. Average column indicates average performance when one of the two modalities gets missing. * indicates that available code and pretrained models from the authors were used to generate the results.	51
3.4	Performance comparison with existing methods for multimodal sentiment analysis.	54
3.5	Performance (top-1 accuracy) comparison with existing methods on NTU RGB+D action recognition dataset. RGB and Depth columns report performance when only RGB and only Depth are available. * indicates that available code and pretrained models were used to generate the results.	55
3.6	Performance (accuracy) comparison with prompting based approach for multimodal classification on UPMC Food-101 dataset. Image and text columns indicate the amount of image and text modality available during both training and testing. † indicates that those values are approximated from the plots published in [94].	56
4.1	Performance comparison when modality gets missing during both training and inference. We train and evaluate our model with same modality ratio following existing baselines [94, 81, 89] and report the mean and standard deviation (SD) of three independent runs with different seeds.	74
4.2	Performance comparison when a complete modality gets missing during inference. We report % accuracy for all the datasets. Our method shows overall better performance compared to recent state-of-the-art methods.	75
4.3	Performance comparison with both modalities available during training and inference on Image-Text datasets. We report % accuracy for UPMC Food-101 and F1-Macro/F1-Micro score for MM-IMDb dataset. Here, ‘-’ indicate results that are not reported in corresponding papers.	76
4.4	Performance (% accuracy) comparison with both modalities available during training and inference on Audio-Video datasets. Here, ‘*’ denotes results generated using existing code and ‘-’ indicates that results for those cells are not published.	76
4.5	Ablation study on the alignment loss weight λ . We report performance on the UPMC Food-101 and AVE datasets across both complete and missing modality scenarios.	81
5.1	Performance comparison on MUSIC-AVQA [96] and AVQA [201] datasets. SSAM achieves the best average performance across all modality combinations and consistently outperforms both training-free model merging baselines and models fine-tuned on paired multimodal data.	100

5.2	Performance comparison on MCUB benchmark [14]. SSAM achieves the best average performance across all modality combinations and consistently outperforms both training-free merging baselines and models finetuned on paired multimodal data.	102
5.3	Comparison of specialist vs. merged MLLMs on the MUSIC-AVQA and AVQA datasets. The merged model consistently outperforms individual specialists across modality pairs and generalizes well to the unseen four-modality (V+I+A+L) setting.	105
5.4	Comparison of specialist vs. merged MLLMs across three domain-specific benchmarks. VLM, Vid-LM, and ALM denote Vision-Language, Video-Language, and Audio-Language models, respectively. SSAM (ours) consistently outperforms individual specialists.	105
A1.1	Hyperparameters for the experiments on CMU-MOSI and CMU-MOSEI datasets for multimodal sentiment analysis.	140
A1.2	Learnable parameter counts for different parameter efficient model adaptation methods. As seen from the table, scale and shift introduce less than 0.7% of the total model parameters.	142
A1.3	Performance comparison (% mIoU) of different parameter-efficient adaptation techniques for MFNet, NYUDv2, and MCubeS datasets. Each column reports mIoU of the Adapted model with the corresponding modalities, and Avg indicates average performance. A and D denote Angle and Degree of Linear Polarization. .	143
A1.4	Performance comparison with parameter efficient model adaptation techniques on CMNeXt model for MFNet dataset. Average column indicates average performance when one of the two modalities gets missing. Mean accuracy, F1 score and % mIoU are shown for all experiments.	144
A1.5	Performance comparison with parameter efficient model adaptation techniques on CMNeXt model for NYUDv2 dataset. Average column indicates average performance when one of the two modalities gets missing. Mean accuracy, F1 score and % mIoU are shown for all experiments.	145
A1.6	Performance comparison with different parameter efficient model adaptation techniques on CMNeXt model for MCubeS dataset. Average column indicates the average performance. Mean accuracy, F1 score and % mIoU are shown for all experiments.	146
A1.7	Per class % IoU comparison between pretrained and adapted CMNeXt model on MFNet dataset. Adapted model show better performance for most of the classes leading to overall performance improvement.	147
A1.8	Per class % IoU comparison between pretrained and adapted CMNeXt model on MCubeS dataset. Adapted model show better performance for most of the classes leading to overall performance improvement. Here A, D, and N stand for Angle of Linear Polarization (AoLP), Degree of Linear Polarization (DoLP) and Near-Infrared (NIR) respectively.	148
A1.9	Missing modality performance comparison of base CMNeXt and Adapted CMNeXt model with Resnet-34 [66], Resnet-50 [66], and Swin-S [117] backbones on MFNet dataset.	151

A1.10	Per-class IoU comparison among Dedicated, Pretrained, and Adapted model on MFNet Dataset.	152
A1.11	Per-class IoU comparison among Dedicated, Pretrained, and Adapted model on MCubeS Dataset.	153
A2.12	List of all the hyperparameters used in our experiments.	159
A2.13	Performance comparison with different LoRA ranks (r). The models are trained on complete multimodal data and tested on both complete and missing modality scenarios. The best Accuracy (Acc.) for each scenario is highlighted in bold . Params (M) indicates the number of learnable parameters in millions.	161
A2.14	Performance comparison on MM-IMDb and UPMC Food-101 datasets when modality gets missing during both training and inference following setup from existing baselines [81, 89].	163
A2.15	Performance comparison with three different parameter efficient adaptation methods. We used rank=1 for all the methods. The best Accuracy (Acc.) for each scenario is highlighted in bold . Params (M) indicates the number of learnable parameters in millions.	172
A2.16	Performance comparison of our method with a ViT versus a SigLIP model as the vision backbone. The models are trained on complete multimodal data and tested on both complete and missing modality scenarios. We report % accuracy as the performance metric, and the highest score for each scenario is highlighted in bold	174
A2.17	Performance comparison of our method using <i>Base</i> versus <i>Large</i> ViT and BERT models. The models are trained on complete multimodal data and tested on both complete and missing modality scenarios. We report % accuracy for UPMC Food-101 and F1-Macro score for MM-IMDb dataset. Params (M) indicates the number of learnable parameters in millions. The best performance for each scenario is highlighted in bold	176
A3.18	Performance of SSAM with different subspace ranks (k) on the MCUB-3 and MCUB-4 datasets. I, A, V, P, and L denote Image, Audio, Video, Point Cloud, and Language, respectively. Rank $k = 128$ empirically yields overall better performance.	181

Chapter 1

Introduction

Multimodal learning (MML) has emerged as a powerful paradigm for building intelligent systems that can reason over heterogeneous input modalities such as vision, language, audio, point-cloud, and different sensor data [8, 197]. By integrating complementary information from multiple modalities, multimodal systems have demonstrated significant improvements across a wide range of applications, including image segmentation [131, 18], visual question answering [126, 201], captioning [220, 205], classification [51, 152], and autonomous driving [194, 150].

Despite recent advances in MML, current multimodal models are often constrained by assumptions that limit their applicability in real-world settings. In particular, many approaches rely on fixed modality combinations, require all modalities to be available during inference, or depend on large-scale paired datasets for training and evaluating multimodal models. These limitations hinder the deployment of multimodal systems in practical scenarios, where available modality combinations can vary widely, data collection is costly, training models require huge compute resources, and models must evolve to incorporate new capabilities over time.

This dissertation addresses these challenges by developing methods that enable flexible, robust, and scalable multimodal learning. The proposed approaches progressively relax constraints on modality combinations, modality availability, and training requirements, moving toward multimodal systems that can operate effectively under real-world conditions.

1.1 Multimodal Learning

Multimodal learning focuses on designing models that can process, align, and fuse information from multiple heterogeneous input sources [138, 197]. A key motivation behind MML is that different input modalities often provide complementary perspectives of the same underlying phenomenon. For example, visual data like image and video capture spatial structure, language encodes semantic meaning, and audio conveys temporal dynamics. Combining these complementary information enables richer feature representations and more reliable predictions.

A central component of multimodal systems is the fusion mechanism, which integrates modality-specific information into a unified, comprehensive representation to enhance model performance, robustness, and prediction. Existing fusion strategies include early fusion [133], where raw inputs are combined prior to feature extraction; late fusion [35], where predictions from unimodal models are aggregated; and intermediate fusion techniques such as attention-based or token-level fusion [178, 134]. Recent advances in transformer-based architectures [90, 99] have further enabled flexible and scalable multimodal reasoning across diverse tasks.

While these methods have achieved strong performance, they are often designed for a fixed set of input modalities which are difficult to extend to newer set of modalities and rely on complete and well-aligned paired multimodal data during both training and inference which is costly

and time-consuming to prepare. Such assumptions are rarely satisfied in practical applications, motivating the need for more adaptive and robust multimodal frameworks.

1.2 Problems with Multimodal Learning

Despite the success of existing approaches, several fundamental challenges remain that limit the real-world application of multimodal systems. This dissertation addresses three key problems to enhance the scalability and robustness of these systems.

1.2.1 Supporting Arbitrary Combinations of Input Modalities

Most existing multimodal models are designed for specific modality combinations, such as RGB-Depth, RGB-Thermal, or Vision-Language pairs [62, 164, 110]. As a result, their architectures and fusion strategies are tightly coupled to the specific set of modalities. Extending these models to new modality combinations often requires redesigning the architecture, fusion strategy or retraining the whole model from scratch.

However, real-world applications frequently involve dynamic and heterogeneous modality combinations. A practical multimodal system should be able to flexibly incorporate any subset of available modalities and effectively leverage additional modalities when they are present. Designing such models requires fusion methods that are both scalable and modality-agnostic, while maintaining computational efficiency and performance.

1.2.2 Missing Modality Robustness

In many real-world scenarios, input modalities may be missing, corrupted, or unavailable due to sensor failures, privacy constraints, or data acquisition limitations [123]. Unfortunately, most multimodal models are not designed to be inherently robust to such conditions and experience significant performance degradation when one or more modalities are absent at inference time.

Existing approaches to address missing modalities include modality dropout [136], masking strategies [5], and modality imputation using generative models [203]. While these methods provide partial solutions, they often require specialized training procedures, training additional networks, or assumptions about modality distributions. Furthermore, most of these approaches do not generalize well to arbitrary missing modality scenarios.

Developing multimodal models that can adapt to varying modality availability while maintaining strong performance in both complete and incomplete modality scenarios remains a critical challenge for robust multimodal learning.

1.2.3 Merging Capabilities from Multiple Multimodal Models

Recent advances in multimodal large language models (MLLMs) have demonstrated impressive capabilities by aligning multiple modalities within a shared embedding space to build a unified model [202, 225]. However, training such models typically requires large-scale paired multimodal datasets and substantial computational resources.

An alternative paradigm is to combine independently trained specialist models, each optimized for a specific modality combination or task to build a unified multimodal model. While training-free model merging methods have shown promise in homogeneous settings [78, 199], ex-

tending these approaches to heterogeneous multimodal models introduces new challenges. In particular, differences in representation spaces and parameter structures can lead to task and modality interference and degraded performance when models are naively combined.

Enabling effective integration of capabilities from multiple expert models without retraining or access to paired multimodal data is essential for scalable multimodal systems that can continuously incorporate new modalities and capabilities without costly retraining or data collection.

1.3 Organization of The Book

This dissertation is organized into four main chapters. Each chapter addresses one key challenge in multimodal learning.

Chapter 2 focuses on multimodal fusion under heterogeneous modality combinations. We introduce MMSFormer, a transformer-based architecture with a novel fusion block designed to integrate information from arbitrary combinations of input modalities. The proposed method demonstrates strong performance across multiple multimodal segmentation tasks and exhibits consistent performance improvements as additional modalities are introduced.

Chapter 3 addresses robustness to missing modalities through parameter-efficient adaptation. We propose a method that modulates intermediate feature representations of a pretrained multimodal model using lightweight transformations, enabling the model to adapt dynamically to different missing modality scenarios. The method requires minimal additional learnable parameters while providing substantial performance improvements in different missing modality scenarios.

Chapter 4 further investigates missing modality robustness by introducing Cross-Modal Proxy Tokens (CMPTs), which approximate representations of missing modalities using available

ones. This approach enables effective handling of diverse missing modality conditions while maintaining strong performance in modality complete settings.

Chapter 5 explores training-free multimodal model merging to combine capabilities from multiple multimodal models. We propose Singular Subspace Alignment and Merging (SSAM), a framework that aligns and merges independently trained MLLMs within a shared low-rank subspace. This method enables the construction of unified multimodal systems without requiring paired training data or joint retraining.

Finally, Chapter 6 discusses the overall contribution and some open problems that are crucial for advancing multimodal learning toward more general, reliable, and efficient systems. Together, these chapters present a comprehensive framework for advancing multimodal learning toward flexible, robust, and scalable systems capable of operating under real-world constraints.

Chapter 2

MMSFormer: Multimodal Transformer for Material and Semantic Segmentation

2.1 Introduction

Image segmentation [21, 131] methods assign one class label to each pixel in an image. The segmentation map can be used for holistic understanding of objects or context of the scene. Image segmentation can be further divided into different types; examples include semantic segmentation [52, 174], instance segmentation [50, 56], panoptic segmentation [38, 91] and material segmentation [106, 169]. Each of these segmentation tasks are designed to address specific challenges and applications. Multimodal image segmentation [53, 219] aims to enhance the accuracy and completeness of the task by leveraging diverse sources of information, and potentially leading to a more robust understanding of complex scenes. In contrast to single-modal segmentation [131], the multimodal approach [219] is more complex due to the necessity of effectively integrating het-

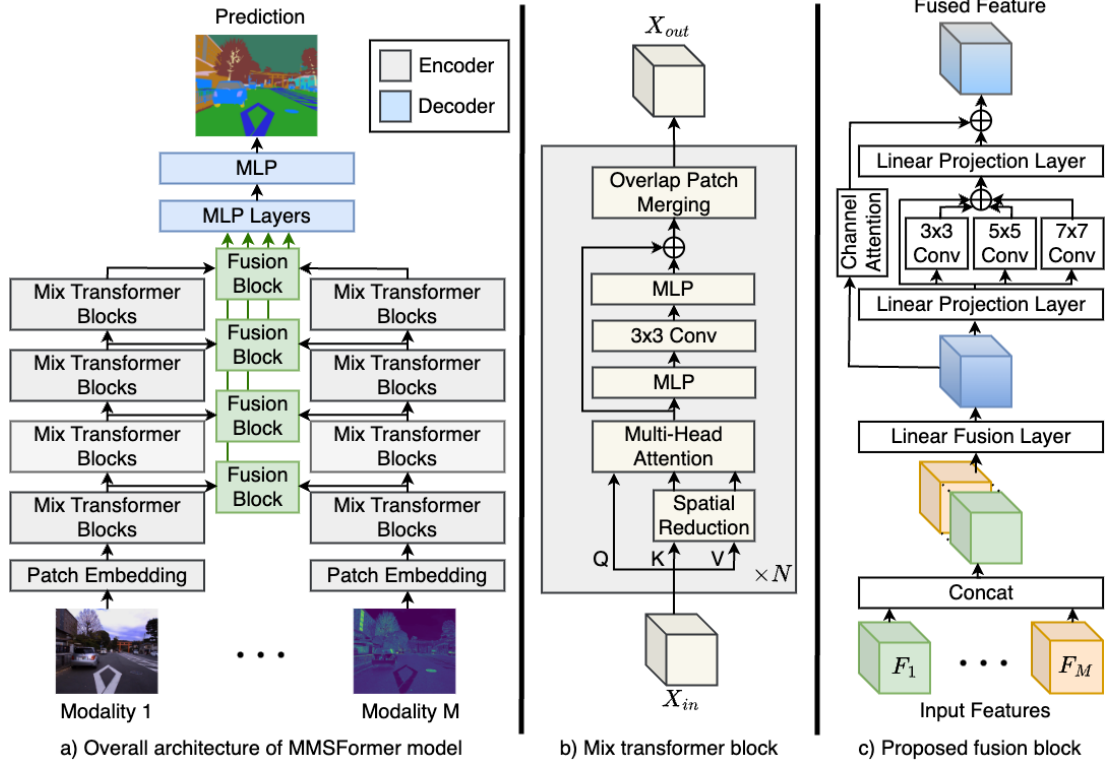


Figure 2.1: (a) Overall architecture of MMSFormer model. Each image passes through a modality-specific encoder where we extract hierarchical features. Then we fuse the extracted features using the proposed fusion block and pass the fused features to the decoder for predicting the segmentation map. (b) Illustration of the mix transformer [195] block. (c) Proposed multimodal fusion block.

erogeneous data from different modalities. Key challenges arise from variations in data quality and attributes, distinct traits of each modality, and need to create models capable of accurately and coherently segmenting with the fused information.

Most of the existing multimodal segmentation methods are designed to work with specific modality pairs, such as RGB-Depth [18, 62, 74], RGB-Thermal [100, 105, 164], and RGB-Lidar [97, 143, 222]. As they are designed for specific modality combinations, most of them generally do not work well with modality combinations different from the ones used in the original design. Recently, CMX [214] introduced a technique to fuse information from RGB and one other supple-

mentary modality, but it is incapable of fusing more than two modalities at the same time. Some recent models have proposed techniques to fuse more than two modalities [11, 106, 215]. However, they either use very complex fusion strategies [11, 214] or require additional information like semantic labels [106] for performing underlying tasks.

In this work, we propose a novel fusion block that can fuse information from diverse combination of modalities. We also propose a new model for multimodal material and semantic segmentation tasks that we call MMSFormer. Our model uses transformer based encoders [195] to capture hierarchical features from different modalities, fuses the extracted features with our novel fusion block and utilizes MLP decoder to perform multimodal material and semantic segmentation. In particular, our proposed fusion block uses parallel convolutions to capture multi-scale features, channel attention to re-calibrate features along the channel dimension and linear layer to combine information across multiple modalities. Such a design provides a simple and computationally efficient fusion block that can handle an arbitrary number of input modalities and combine information effectively from different modality combinations. We compare our fusion block with some of the existing fusion methods in terms of number of parameters and GFLOPs in Table 2.9.

To evaluate our proposed MMSFormer and fusion block, we focus on multimodal material segmentation on MCubeS [106] dataset and multimodal semantic segmentation on FMB [114] and PST900 [158] datasets. MCubeS dataset consists of four different modalities: RGB, angle of linear polarization (AoLP), degree of linear polarization (DoLP) and near-infrared (NIR). FMB dataset includes RGB and infrared modalities, while PST900 dataset comprises RGB and thermal modalities. We show the overall and per-class performance comparison in Table 2.1-2.5 for these datasets. A series of experiments highlight the ability of the proposed fusion block to effectively combine fea-

tures from different modality combinations, resulting in superior performance compared to current state-of-the-art methods. Ablation studies show that different input modalities assist in identifying different types of material classes as shown in Table 2.8. Furthermore, as we add new input modalities, overall performance increases gradually highlighting the ability of the fusion block to incorporate useful information from new modalities. We summarize the results in Table 2.4 and 2.6 for FMB and MCubeS datasets respectively. Main contributions of this work are:

- We propose MMSFormer, a new multimodal segmentation model that incorporates a novel fusion block to fuse information from arbitrary (heterogeneous) combinations of modalities.
- Our model achieves state-of-the-art performance on three different datasets. Furthermore, our method achieves better performance for all modality combinations compared to the current leading models.
- A series of ablation studies show that each module on the fusion block has an important contribution towards the overall model performance and each input modality assists in identifying specific material classes.

Rest of the chapter is structured as follows. Section 2.2 presents a brief review of related work. We describe our model and fusion block in detail in Section 2.3. Section 2.4 presents experimental results and ablation studies on multimodal material and semantic segmentation tasks with qualitative and quantitative analysis.

2.2 Related Work

Image segmentation has witnessed significant evolution, spurred by advancements in machine learning and computational capabilities. A significant improvement in this evolution came with the inception of fully convolutional networks (FCNs) [119, 190], which enabled pixel-wise predictions through the utilization of hierarchical features within convolutional neural networks (CNNs). This led to the development of a variety of CNN based models for different images segmentation tasks. U-Net [151] is one such model that utilizes skip-connections between the lower-resolution and corresponding higher-resolution feature maps. DeepLabV3+ [16] introduced dilated convolutions (atrous convolutions) into the encoder allowing the expansion of the receptive field without increasing computational complexity significantly. PSPNet [221] introduced global context modules that enable the model to gather information from a wide range of spatial scales, essentially integrating both local and global context into the segmentation process.

Recently, Transformer based models have proven to be very effective in handling complex image segmentation tasks. Some of the notable transformer-based models are Pyramid Vision Transformer (PVT) [176], SegFormer [195], and Mask2Former [20]. PVT [176] utilizes transformer based design for various computer vision tasks. SegFormer [195] utilizes efficient self-attention and lightweight MLP decoder for simple and efficient semantic segmentation. Mask2Former [20] uses masked-attention along with pixel decoder and transformer decoder for any segmentation task. Their success demonstrates the capacity of these models to provide state-of-the-art solutions in various segmentation tasks.

In the context of multimodal image segmentation, fusion of data from diverse sources [219] has gained traction as a means to extract richer information and improve accuracy. A variety

of models and fusion strategies have been proposed for RGB-Depth segmentation tasks. FuseNet [62] model integrates depth feature maps into RGB feature maps, while SA-Gate [18] employs Separation-and-Aggregation Gating to mutually filter and recalibrate RGB and depth modalities before fusion. Attention Complementary Module has been proposed by ACNet [74] that extracts weighted RGB and depth features for fusion. The domain of RGB-Thermal image segmentation has also gained prominence. Recent models include RTFNet [164] that achieves fusion through elementwise addition of thermal features with RGB, RSFNet [100] proposing Residual Spatial Fusion module to blend RGB and Thermal modalities, and EAEFNet [105] utilizing attention interaction and attention complement mechanisms to merge RGB and Thermal features. A number of methods also focus on fusing RGB-Lidar data that include TransFuser [143] that employs Transformer blocks, whereas LIF-Seg [222] relies on coarse feature extraction, offset learning, and refinement for effective fusion.

While the previously mentioned studies focus on specific pairs of modalities, some recent research has demonstrated promising results in the fusion of arbitrary modalities. CMX [214] introduces cross-modal feature rectification and fusion modules to merge RGB features with supplementary modalities. For multimodal material segmentation, MCubeSNet [106] model is proposed, which can seamlessly integrate four different modalities to enhance segmentation accuracy. In the context of arbitrary modal semantic segmentation, CMNeXt [215] introduces Self-Query Hub and Parallel Pooling Mixer modules, offering a versatile approach for fusing diverse modalities. Additionally, HRFuser [11] employs multi-window cross-attention to fuse different modalities at various resolutions, thereby enriching model performance.

Though some of these models can fuse different modalities, they either use very complex fusion strategies [11, 214, 215] or requires additional information [106] to perform underlying task. We aim to design a simple fusion module that can handle arbitrary number of input modalities and able to effectively fuse information from diverse modality combinations.

2.3 Proposed Model

The overall architecture of our proposed MMSFormer model and the fusion block is shown in Figure 2.1. The model has three modules: (1) Modality specific encoder; (2) Multimodal fusion block; and (3) Shared MLP decoder. We use mix transformer [195] as the encoder of our model. We choose mix transformer for various reasons. First, it can provide hierarchical features without positional encoding. Second, it uses spatial reduction before attention that reduces the number of parameters significantly [176, 195]. Third, it also works well with simple and lightweight MLP decoder [195].

2.3.1 Overall Model Architecture

Our overall model architecture is shown in Figure 2.1a. Assume we have M distinct modalities. Given a set of modalities as input, each modality-specific encoder captures distinctive features from each input modality by mapping the corresponding image into modality-specific hierarchical features as

$$F_m = \text{Encoder}_m(I_m), \quad (2.1)$$

where $I_m \in \mathbb{R}^{H \times W \times 3}$ represents input image for modality $m \in \{1, 2, \dots, M\}$ and $\text{Encoder}_m(\cdot)$ denotes the encoder for that modality. The encoder generates four feature maps at $\{\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}\}$

of the input image resolution. We represent them as $F_m = \{F_m^1, F_m^2, F_m^3, F_m^4\}$. For simplicity we denote the shape of the feature map for the i^{th} encoder stage as $(H_i \times W_i \times C_i)$ where $i \in \{1, 2, 3, 4\}$. We use four separate fusion blocks, one corresponding to each encoder stage, to fuse the features from each stage of the encoder. We pass the extracted features F_m^i for all modalities to the i^{th} fusion block as

$$F^i = \text{FusionBlock}^i(\{F_m^i\}_m). \quad (2.2)$$

Each fusion block fuses the features extracted from all the modalities to generate a combined feature representation $F = \{F^1, F^2, F^3, F^4\}$, where F^i denotes the fused feature at i^{th} stage. Finally, we pass the combined features F to the MLP decoder [195] to predict the segmentation labels.

2.3.2 Modality Specific Encoder

We use mix transformer encoder [195] to capture hierarchical features from the input modalities. Each input image I_m goes through patch embedding layer where it is divided into 4×4 patches following [195] and then fed to the mix transformer blocks. The design of mix transformer block is shown in Figure 2.1b. We denote the input to any mix transformer block as $X_{in} \in \mathbb{R}^{H_i \times W_i \times C_i}$ that is reshaped to $N_i \times C_i$ (with $N_i = H_i W_i$) and used as query Q , key K , and value V . To reduce the computational overhead, spatial reduction is applied following [176] using a reduction ratio R . K and V are first reshaped into $\frac{N_i}{R} \times C_i R$ matrices and then mapped to $\frac{N_i}{R} \times C_i$ matrices via linear projection. A standard multi-head self-attention (MHSA) maps Q, K, V to intermediate features as

$$\text{MHSA}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h) W^O, \quad (2.3)$$

$$\text{head}_j = \text{Attention}(Q W_j^Q, K W_j^K, V W_j^V),$$

where h represents the number of attention heads, $W_j^Q \in \mathbb{R}^{C_i \times d_K}$, $W_j^K \in \mathbb{R}^{C_i \times d_K}$, $W_j^V \in \mathbb{R}^{C_i \times d_V}$, and $W^O \in \mathbb{R}^{hd_V \times C_i}$ are the projection matrices, d_K, d_V represent dimensions of K, V , respectively. We can formulate the Attention function as

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_K}}\right)V, \quad (2.4)$$

where Q, K , and V are the input query, key, and value matrices. MHSA is followed by a mixer layer (with two MLP and one 3×3 convolution layer). The convolution layer provides sufficient positional encoding into the transformer encoder for optimal segmentation performance [195]. This layer can be written as

$$\begin{aligned} \hat{X}_{in} &= \text{MHSA}(Q, K, V), \\ X_{out} &= \text{MLP}(\text{GELU}(\text{Conv}_{3 \times 3}(\text{MLP}(\hat{X}_{in})))) + \hat{X}_{in}, \end{aligned} \quad (2.5)$$

Finally, overlap patch merging is applied to X_{out} following [195] to generate the final output.

2.3.3 Multimodal Fusion Block

After extracting hierarchical features, we fuse them using our proposed fusion block. The fusion block shown in Figure 2.1c is responsible for fusing the features extracted from the modality specific encoders. We have one fusion block for each of the four encoder stages. For the i^{th} fusion block, let us assume the input feature maps are given as $F_m^i \in \mathbb{R}^{H_i \times W_i \times C_i} \forall m \in \{1, 2, \dots, M\}$. First, we concatenate the feature maps from M modalities along the channel dimension to get the combined feature map $F^i \in \mathbb{R}^{H_i \times W_i \times MC_i}$. Then we pass the features through a linear fusion layer that combines the features and reduces the channel dimension to C_i . Let us denote the resulting features as $\hat{F}^i \in \mathbb{R}^{H_i \times W_i \times C_i}$. We represent the operation as

$$\hat{F}^i = \text{Linear}(F_1^i || \dots || F_M^i). \quad (2.6)$$

Here $\parallel$ represents concatenation of features along the channel dimension and the linear layer takes an MC_i dimensional input and generates a C_i dimensional output.

After the linear fusion layer, we added a module for capturing and mixing multi-scale features. The module consists of two linear projection layers having parallel convolution layers in between them. First we apply a linear transformation on $\hat{F}^i$ along the channel dimension by passing it through the first linear projection layer. It refines and tunes the features from different channels. Then we apply 3×3 , 5×5 , and 7×7 convolutions to effectively capture diverse features across different spatial contexts. By employing convolutions with different sizes, the fusion block can attend to local patterns as well as capture larger spatial structures, thereby enhancing its ability to extract meaningful features from the input data. Finally we apply another linear transformation along the channel dimension using the second linear project layer to consolidate the information captured by the parallel convolutions, promoting feature consistency and enhancing the discriminative power of the fused features. These steps can be performed as

$$\tilde{F}^i = \text{Linear}(\hat{F}^i), \quad (2.7)$$

$$F^i = \text{Linear}(\tilde{F}^i + \sum_{k \in \{3,5,7\}} \text{Conv}_{k \times k}(\tilde{F}^i)). \quad (2.8)$$

We found that using 3 parallel convolution layers with sizes 3×3 , 5×5 , and 7×7 provide optimal performance. Increasing the convolution kernel size reduces performance which we show in Table 2.7. As larger kernels reduce performance, we did not add more than 3 parallel convolutions in our model.

We apply Squeeze-and-Excitation block [73] as channel attention in the residual connection. The final fused feature can be represented as

$$F^i = \text{ChannelAttention}(\hat{F}^i) + F^i. \quad (2.9)$$

Channel attention re-calibrates interdependence between channels and allows the model to select the most relevant features or channels while suppressing less important ones [73]. This leads to more effective feature representations and thus better performance on the underlying task.

2.3.4 Shared MLP Decoder

The fused features generated from all the 4 fusion blocks are sent to the shared MLP decoder. We use the decoder design proposed in [195]. The decoder shown in Figure 2.1a can be represented as the following equations:

$$\begin{aligned}
 F^i &= \text{Linear}(F^i), \quad \forall i \in \{1, 2, 3, 4\} \\
 F^i &= \text{Upsample}(F^i), \quad \forall i \in \{1, 2, 3, 4\} \\
 F &= \text{Linear}(F^1 || \dots || F^4), \\
 P &= \text{Linear}(F).
 \end{aligned} \tag{2.10}$$

The first linear layers take the fused features of different shapes and generate features having the same channel dimension. Then the features are up-sampled to $\frac{1}{4}^{th}$ of the original input shape, concatenated along the channel dimension and passed through another linear layer to generate the final fused feature F . Finally F is passed through the last linear layer to generate the predicted segmentation map P .

2.4 Experiments and Results

We evaluated our model and proposed fusion block on multiple datasets and with different modality combinations for multimodal semantic and material segmentation tasks. We also compared our methods with existing baseline methods both qualitatively and quantitatively. We

report results from already published works whenever possible. * indicates that we have used the code and pretrained models from the papers to generate the results.

2.4.1 Datasets

Multimodal material segmentation (MCubeS) dataset [106] contains 500 sets of images from 42 street scenes having four modalities: RGB, AoLP, DoLP, and NIR. It provides annotated ground truth labels for both material and semantic segmentation and divided into training set with 302 image sets, validation set with 96 image sets, and test set with 102 image sets. This dataset has 20 class labels corresponding to different materials.

FMB dataset [114] is a new and challenging dataset with 1500 pairs of calibrated RGB-Infrared image pairs. The training and test set contains 1220 and 280 image pairs respectively. The dataset covers a wide range of scenes under different lighting and weather conditions (Tyndall effect, rain, fog, and strong light). It also provides per pixel ground truth annotation for 14 different classes.

PST900 dataset [158] contains 894 pairs of synchronized RGB-Thermal image pairs. The dataset is divided into training and test sets with per pixel ground truth annotation for five different classes.

2.4.2 Implementation Details

To ensure a fair comparison with prior models, we followed the same data preprocessing and augmentation strategies employed in previous studies [106, 215, 114]. We used the Mix-Transformer (MiT) [195] encoder pretrained on the ImageNet [27] dataset as the backbone for our model to extract features from different modalities. Each modality has a separate encoder. We used a shared MLP decoder introduced in SegFormer [195] and used random initialization for it.

Table 2.1: Performance comparison on Multimodal Material Segmentation (MCubeS) dataset. Here A, D, and N represent AoLP, DoLP, and NIR respectively.

Method	Modalities	% mIoU
DRConv [15]	RGB-A-D-N	34.63
DDF [232]	RGB-A-D-N	36.16
TransFuser [143]	RGB-A-D-N	37.66
DeepLabv3+ [16]	RGB-A-D-N	38.13
MMTM [82]	RGB-A-D-N	39.71
FuseNet [62]	RGB-A-D-N	40.58
MCubeSNet [106]	RGB-A-D-N	42.46
CBAM [185]	RGB-A-D-N	51.32
CMNeXt [215]	RGB-A-D-N	51.54
MMSFormer	RGB-A-D-N	53.11

Table 2.2: Performance comparison on FBM dataset. We show performance for different methods from already published works.

Methods	Modalities	% mIoU
CBAM [185]	RGB-Infrared	50.1
GMNet [237]	RGB-Infrared	49.2
LASNet [95]	RGB-Infrared	42.5
EGFNet [32]	RGB-Infrared	47.3
FEANet [26]	RGB-Infrared	46.8
DIDFuse [227]	RGB-Infrared	50.6
ReCoNet [75]	RGB-Infrared	50.9
U2Fusion [196]	RGB-Infrared	47.9
TarDAL [113]	RGB-Infrared	48.1
SegMiF [114]	RGB-Infrared	54.8
MMSFormer	RGB-Infrared	61.7

We trained and evaluated all our models using two NVIDIA RTX 2080Ti GPUs using PyTorch. We utilized a polynomial learning rate scheduler with a power of 0.9 to dynamically adjust the learning rate during training. The first 10 epochs were designated as warm-up epochs with a learning rate of 0.1 times the original rate. For loss computation, we used the cross-entropy loss function. Optimization was performed using the AdamW [120] optimizer with an epsilon value of 10^{-8} and weight decay set to 0.01. For CBAM [185], we use the same encoder, decoder and hyperparameters used in our experiments and replace our fusion block with CBAM¹ module. To be specific, after extracting the feature maps from each input modality using the modality specific encoders, we add them (sum them up) and pass the combined feature map to the CBAM module.

¹<https://github.com/luuuyi/CBAM.PyTorch>

Table 2.3: Per-class % IoU comparison on MCubeS dataset. MMSFormer shows better performance in detecting most of the classes compared to the current state-of-the-art models. * indicates that the code and pretrained model from the authors were used to generate the results.

Methods	Asphalt	Concrete	Metal	Road marking	Fabric	Glass	Plaster	Plastic	Rubber	Sand	Gravel	Ceramic	Cobblestone	Brick	Grass	Wood	Leaf	Water	Human	Sky	Mean
MCubeSNet [106]	85.7	42.6	47.0	59.2	12.5	44.3	3.0	10.6	12.7	66.8	67.1	27.8	65.8	36.8	54.8	39.4	73.0	13.3	0.0	94.8	42.9
CBAM [185]	85.7	47.7	55.4	70.4	27.6	54.7	0.9	30.9	26.5	61.6	63.0	28.0	71.1	41.8	58.6	47.4	76.7	56.3	25.9	96.5	51.3
CMNeXt [215] *	84.3	44.9	53.9	74.5	32.3	54.0	0.8	28.3	29.7	67.7	66.5	27.7	68.5	42.9	58.7	49.7	75.4	55.7	18.9	96.5	51.5
MMSFormer	88.0	48.3	56.2	72.2	35.4	54.9	0.5	34.6	29.4	67.2	69.0	29.9	73.4	44.7	59.5	47.8	77.1	50.5	26.9	96.6	53.1

2.4.3 Performance Comparison with Existing Methods

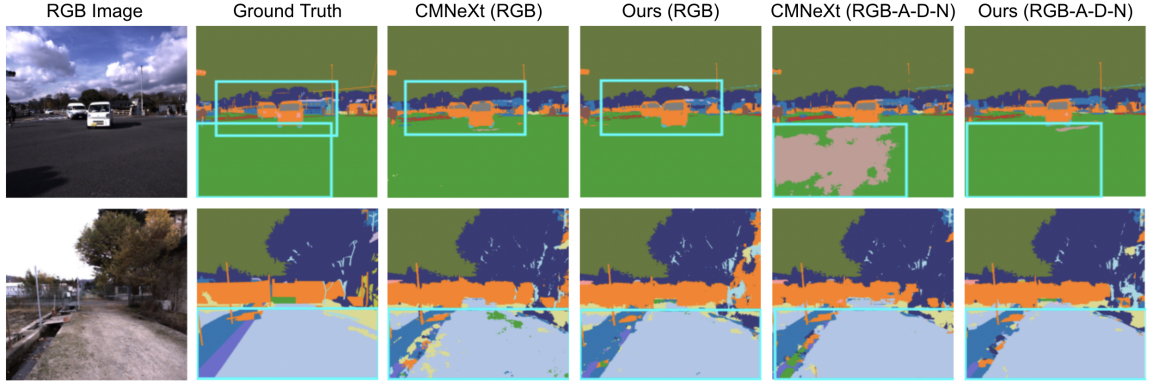
We conducted a rigorous performance evaluation of our model compared to established baseline models for three datasets. The comprehensive results are summarized in Tables 2.1–2.6. We report results for CBAM from our experiments. Other results are taken from published literature.

Results on MCubeS Dataset. Table 2.1 shows the overall performance comparison between our model and existing baseline models for MCubeS dataset. Our model achieves a mean intersection-over-union (mIoU) of 53.11%, surpassing the current state-of-the-art model. It shows 1.57% improvement over CMNeXt [215], 1.79% improvement over CBAM [185] and 10.65% improvement over MCubeSNet [106] models. To further analyze the performance of our model, we conducted a per-class IoU analysis and presented in Table 2.3. Our model performs better in detecting most of the material classes compared to the current best models. Notably, our model demonstrates a substantial improvement in detecting plastic (+3.7%), fabric (+3.1%), asphalt (+2.3%), and cobblestone (2.3%) classes while maintaining competitive or better performance in other classes. This led to the overall better performance and sets new state-of-the-art for this dataset.

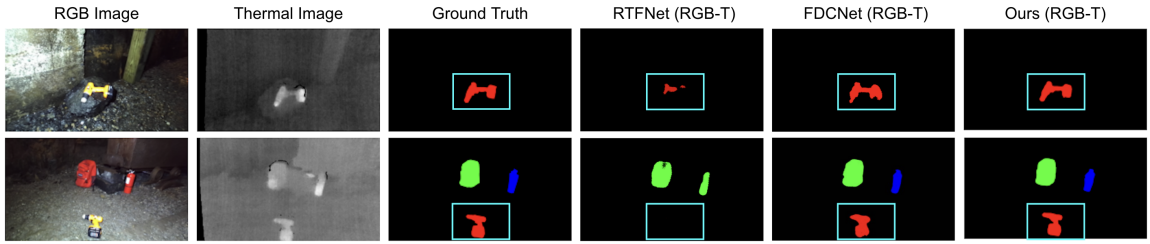
Table 2.4: Per-class % IoU comparison on FMB dataset. We show comparison for 8 classes (out of 14) that are published. MMSFormer outperforms existing methods for all the classes except truck.

Methods	Modalities	Car	Person	Truck	T-Lamp	T-Sign	Building	Vegetation	Pole	% mIoU
SegMiF [114]	RGB	78.3	46.6	43.4	23.7	64.0	77.8	82.1	41.8	50.5
MMSFormer	RGB	80.3	56.7	42.1	31.6	77.8	77.9	85.4	48.1	57.2
CBAM [185]	RGB-Infrared	71.9	49.3	20.9	25.8	67.1	75.8	80.9	19.7	50.1
GMNet [237]	RGB-Infrared	79.3	60.1	22.2	21.6	69.0	79.1	83.8	39.8	49.2
LASNet [95]	RGB-Infrared	72.6	48.6	14.8	2.9	59.0	75.4	81.6	36.7	42.5
EGFNet [32]	RGB-Infrared	77.4	63.0	17.1	25.2	66.6	77.2	83.5	41.5	47.3
FEANet [26]	RGB-Infrared	73.9	60.7	32.3	13.5	55.6	79.4	81.2	36.8	46.8
DIDFuse [227]	RGB-Infrared	77.7	64.4	28.8	29.2	64.4	78.4	82.4	41.8	50.6
ReCoNet [75]	RGB-Infrared	75.9	65.8	14.9	34.7	66.6	79.2	81.3	44.9	50.9
U2Fusion [196]	RGB-Infrared	76.6	61.9	14.4	28.3	68.9	78.8	82.2	42.2	47.9
TarDAL [113]	RGB-Infrared	74.2	56.0	18.8	29.6	66.5	79.1	81.7	41.9	48.1
SegMiF [114]	RGB-Infrared	78.3	65.4	47.3	43.1	74.8	82.0	85.0	49.8	54.8
MMSFormer	RGB-Infrared	82.6	69.8	44.6	45.2	79.7	83.0	87.3	51.4	61.7

Results on FMB Dataset. Performance Comparison for FMB dataset is shown on Table 2.2. Our model shows a significant improvement of 6.9% mIoU compared to the current state-of-the-art model. Per-class IoU analysis for this dataset is shown on Table 2.4. For a fair comparison, we only compare the performance on 8 classes (out of 14) that are published in literature. T-Lamp and T-Sign represent Traffic Lamp and Traffic Sign, respectively. Our model shows an overall performance improvement of 6.7% mIoU for RGB only predictions compared to the most recent SegMiF [114] model. Alongside this, our model also shows superior performance in detecting all of the classes except for the truck class for both RGB only and RGB-Infrared semantic segmentation



(a) Visualization of predictions on MCubeS dataset



(b) Visualization of predictions on PST900 dataset

Figure 2.2: Visualization of predictions on MCubeS and PST900 datasets. For brevity, we only show the RGB image and ground truth material segmentation maps along with the predictions in Figure 2.2a. Our model shows better predictions on both of the datasets.

tasks. Performance on RGB-Infrared input modalities is much better than RGB only performance for all the classes, which demonstrates the ability of the fusion block to effectively fuse information from the input modalities.

Results on PST900 Dataset. We also tested our model on PST900 [158] dataset and summarized the result in Table 2.5. Experiments show that our model outperforms existing baseline models for RGB-Thermal semantic segmentation on this dataset. It outperforms the most recent CACFNet [233] model by 0.89% mIoU. Our model also shows better performance in detecting 3

Table 2.5: Per-class % IoU and % mIoU comparison on PST900 dataset.

Methods	Modalities	Background	Fire-Extinguisher	Backpack	Hand-Drill	Survivor	% mIoU
ACNet [74]	RGB-Thermal	99.25	59.95	83.19	51.46	65.19	71.81
CCNet [76]	RGB-Thermal	99.05	51.84	66.42	32.27	57.50	61.42
Efficient FCN [111]	RGB-Thermal	98.63	39.96	58.15	30.12	28.00	50.98
RTFNet [164]	RGB-Thermal	99.02	51.93	74.17	7.07	70.11	60.46
PSTNet [158]	RGB-Thermal	98.85	70.12	69.20	53.60	50.03	68.36
EGFNet [235]	RGB-Thermal	99.26	71.29	83.05	64.67	74.30	78.51
MTANet [234]	RGB-Thermal	99.33	64.95	87.50	62.05	79.14	78.60
MFENet [236]	RGB-Thermal	99.40	66.38	81.02	72.50	75.60	78.98
GMNet [237]	RGB-Thermal	99.44	73.79	83.82	85.17	78.36	84.12
CGFNet [172]	RGB-Thermal	99.30	71.71	82.00	59.72	77.42	78.03
GCNet [112]	RGB-Thermal	99.35	77.68	79.37	82.92	73.58	82.58
GEBNet [31]	RGB-Thermal	99.39	73.07	85.93	67.14	80.21	81.15
GCGLNet [47]	RGB-Thermal	99.39	77.57	81.01	81.90	76.31	83.24
DHFNet [12]	RGB-Thermal	99.44	78.15	87.38	71.18	74.81	82.19
MDRNet+ [223]	RGB-Thermal	99.07	63.04	76.27	63.47	71.26	74.62
FDCNet [224]	RGB-Thermal	99.15	71.52	72.17	70.36	72.36	77.11
CBAM [185]	RGB-Thermal	99.43	73.81	82.75	80.00	69.60	81.12
EGFNet [32]	RGB-Thermal	99.55	79.97	90.62	76.08	80.88	85.42
CACFNet [233]	RGB-Thermal	99.57	82.08	89.49	80.90	80.76	86.56
MMSFormer	RGB-Thermal	99.60	81.45	89.86	89.65	76.68	87.45

out of the 5 classes available in the dataset and competitive performance in other two classes.

2.4.4 Performance Comparison for Incremental Modality Integration

A critical aspect of this work involves evaluating the effectiveness of our proposed fusion block in combining valuable information from diverse modalities. To analyze this effect, we trained

Table 2.6: Performance comparison (% mIoU) on MCubeS dataset for different modality combinations. Here A, D, and N represent AoLP, DoLP, and NIR respectively.

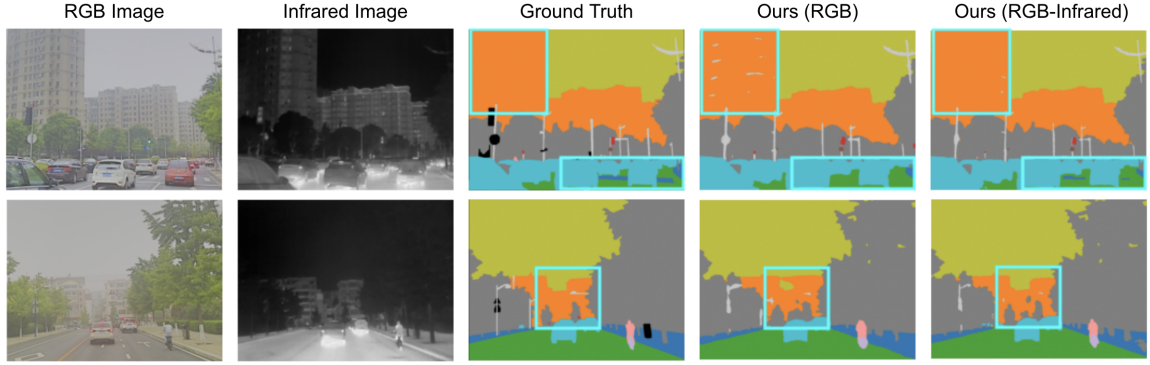
Modalities	MCubeSNet [106]	CMNeXt [215]	MMSFormer
RGB	33.70	48.16	50.44
RGB-A	39.10	48.42	51.30
RGB-A-D	42.00	49.48	52.03
RGB-A-D-N	42.86	51.54	53.11

our model with various combinations of modalities on the MCubeS dataset. The results are presented in Table 2.6. Our model exclusively trained on RGB data provided an mIoU score of 50.44%, which is 2.28% greater than the current state-of-the-art model. We observe progressive improvement in performance as we incorporated additional modalities: AoLP, DoLP, and NIR. The integration led to incremental performance gains, with the mIoU increasing from 50.44% to 51.30%, then to 52.03%, and ultimately reaching to 53.11%. These findings serve as a compelling evidence that our fusion approach effectively leverages and fuses valuable information from different combination of modalities, resulting in a notable enhancement in segmentation performance.

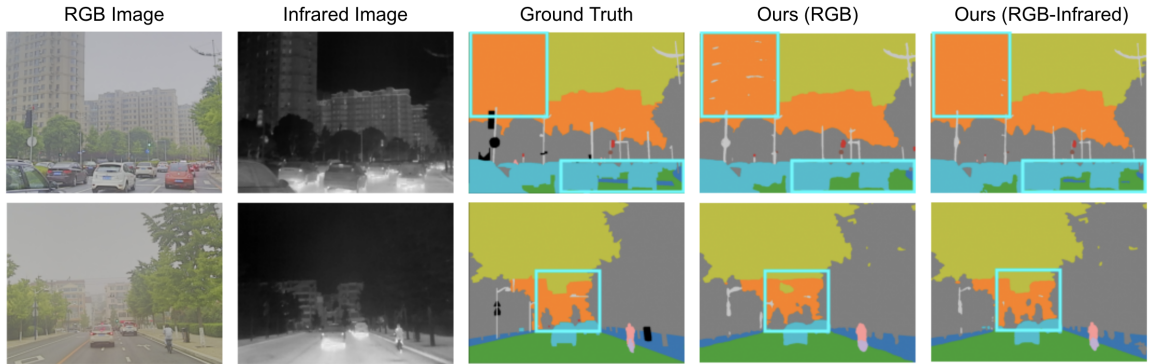
Furthermore, our model consistently outperforms the current state-of-the-art benchmark across all modality combinations. This consistent superiority underscores the robustness and versatility of our fusion block, demonstrating its ability to adapt and excel regardless of the specific modality combination provided.

2.4.5 Qualitative Analysis of the Predictions

Apart from quantitative analysis, we also perform qualitative analysis of the predicted segmentation maps. We show material segmentation results predicted by CMNeXt [215] model



(a) Visualization of predictions on MCubeS dataset for different modality combinations



(b) Visualization of predictions on FMB dataset for different modality combinations

Figure 2.3: Visualization of predicted segmentation maps for different modality combinations on MCubeS and FMB datasets. Both figures show that prediction accuracy increases as we incrementally add new modalities. They also illustrate the fusion block’s ability to effectively combine information from different modality combinations.

and our proposed MMSFormer model in Figure 2.2a. For brevity, we only show RGB images and ground truth material segmentation maps in the illustrations. We show RGB only predictions and all modalities (RGB-A-D-N) predictions for both of the models. As highlighted in the rectangular bounding boxes, our proposed MMSFormer model identifies asphalt, sand and water with greater accuracy than CMNeXt [215] for both RGB only and all modalities (RGB-A-D-N) predictions.

Table 2.7: Ablation study of the Fusion Block on FMB dataset with RGB and infrared modalities. The table shows the contribution of different modules in fusion block in overall model performance.

Structure	Parameter Count (M)	% mIoU (Change)
MMSFormer	61.26	61.68
- without channel attention	61.21	58.32 (-3.36)
- without parallel convolutions	61.17	57.17 (-4.51)
- with 3x3, 7x7 and 11x11 convolutions	61.36	56.32 (-5.36)
- only linear fusion	59.57	52.43 (-9.25)

We also compare our prediction on PST900 dataset with RTFNet [164] and FDCNet [224] on Figure 2.2b. We show the input RGB image, thermal images, ground truth segmentation maps and predictions from the models. As highlighted by the rectangular bounding boxes, MMSFormer shows better accuracy in detecting objects with more precise contours compared to other methods.

2.4.6 Ablation Study on the Fusion Block

We conducted a number of ablation studies to investigate the contributions of individual components within the fusion block to the overall model performance. The findings, as detailed in Table 2.7, shed light on the critical importance of these components. We used both RGB and infrared modalities of the FMB dataset during training and testing in these experiments. First, we observed that the absence of channel attention in the residual connection had a negative impact, resulting in a reduction in performance by 3.36%. This indicates that feature calibration along channel dimension plays an important role in capturing and leveraging crucial information effectively. Additionally, while comparing larger convolution kernel sizes (3×3 , 7×7 , and 11×11) to the originally employed (3×3 , 5×5 , and 7×7), we noted a decrease in performance by 5.36%. This result underscores

Table 2.8: Per class % IoU comparison on MCubeS dataset for different modality combinations. As we add modalities incrementally, overall performance increases gradually. This table also shows that specific modality combinations assist in identifying specific types of materials better.

Modalities	Asphalt	Concrete	Metal	Road marking	Fabric	Glass	Plaster	Plastic	Rubber	Sand	Gravel	Ceramic	Cobblestone	Brick	Grass	Wood	Leaf	Water	Human	Sky	Mean
RGB	83.2	44.2	52.1	70.4	31.0	51.6	1.3	26.2	21.8	65.0	61.8	31.3	72.5	45.0	55.4	46.0	74.7	56.0	22.7	96.4	50.4
RGB-A	86.5	46.5	55.9	73.0	35.3	56.0	0.8	27.3	27.8	66.2	67.0	28.6	69.6	43.0	57.6	49.6	76.4	53.8	8.4	96.6	51.3
RGB-A-D	86.0	44.0	55.5	68.1	31.9	54.8	2.3	30.0	29.7	69.4	73.7	32.2	69.4	41.4	59.2	48.3	76.6	50.6	20.9	96.7	52.0
RGB-A-D-N	88.0	48.3	56.2	72.2	35.4	54.9	0.5	34.6	29.4	67.2	69.0	29.9	73.4	44.7	59.5	47.8	77.1	50.5	26.9	96.6	53.1

the significance of the carefully chosen convolution kernel sizes within the fusion block.

Furthermore, completely removing the parallel convolutions from the block led to a performance decline of 4.51%, emphasizing their substantial contribution in capturing multi-scale features and overall model performance. Finally, if we only use the linear fusion layer to fuse the features and remove the parallel convolutions and channel attention from the fusion block, performance drop significantly by 9.25%. These studies demonstrate that multi-scale feature capturing via parallel convolutions and channel-wise feature calibration using channel attention is extremely important in learning better feature representation and thus crucial to overall model performance. These comprehensive ablation studies collectively underscore the significance of every component within the fusion block, revealing that each module plays a distinct and vital role in achieving the overall performance of our model.

2.4.7 Ablation Study on Different Modality Combinations

To analyze the contributions of different modalities in the identification of distinct materials, we conducted a series of ablation studies, focusing on per-class IoU for different modality com-

binations. The insights are summarized in Table 2.8. As we progressively integrate new modalities, performance gradually increases for specific classes, which include grass, leaf, asphalt, cobblestone and plastic classes. Particularly noteworthy is the assistance provided by NIR data in classifying asphalt, concrete, plastic, cobblestone, and human categories, leading to significant performance gains in these classes as NIR was added as an additional modality.

Conversely, certain classes such as water and brick exhibited a gradual performance decline as we introduced additional modalities. This suggests that RGB data alone suffices for accurately identifying these classes, and the inclusion of more modalities potentially introduces noise or redundancy that negatively impacts performance. Moreover, AoLP appears to be helpful in enhancing the recognition of materials like road markings, glass and wood. Similarly, DoLP improved performance for classes like plaster, rubber, sand, gravel and ceramic. These findings underscore the relationship between different imaging modalities and the unique characteristics of different types of materials, demonstrating that specific modalities excel in detecting particular classes based on their distinctive traits.

In Figure 2.3a, we present some examples to show how adding different modalities help improve performance. We show predictions for RGB, RGB-A, RGB-A-D and RGB-A-D-N from our proposed MMSFormer model. As we add new modalities, the predictions become more accurate as shown in the bounding boxes. The illustrations show that the identification of concrete and gravel becomes more accurate with additional modalities. Figure 2.3b shows predictions for RGB and RGB-Infrared from FMB dataset. As highlighted by the bounding boxes, adding new modality helps improve performance in detecting building, road and sidewalks. This also illustrates the capability of the fusion block to effectively fuse information from different modality combinations.

Table 2.9: Comparison of number of parameters in the fusion block and GFLOPs for MCubeS dataset having 4 input modalities with an input shape of $(3 \times 512 \times 512)$ for each modality. Our fusion block shows significantly lower complexity compared to existing methods.

Methods	Fusion Block Parameters (M)	GFLOPs
CMNeXt [215]	16.63	6.47
MCubeSNet [106]	7.41	12.10
HRFuser [11]	1.72	17.50
CMX [214]	16.59	6.41
DDF (Resnet-101) [232]	28.10	4.10
MMSFormer (Ours)	3.23	2.47

2.4.8 Computational Complexity of the Fusion Block

In addition to better performance, our fusion block is also computationally efficient compared to most of the fusion blocks proposed for these datasets. We show a comparison in terms of the number of parameters in the fusion block and GFLOPs for some recent models on Table 2.9 for MCubeS dataset having 4 input modalities with an input shape of $(3 \times 512 \times 512)$ for each modality. As observed from the table, our proposed fusion strategy demonstrates significantly lower complexity in terms of both the number of parameters and GFLOPs compared to existing methods. HRFuser [11] has a lower parameter count than ours but it requires more than $7\times$ GFLOPs. Other methods require significantly more parameters ($2.3\times - 8.7\times$) and GFLOPs ($1.6\times - 7\times$) compared to our proposed fusion strategy. Our comparison only includes models for which these results are available in the published literature.

2.5 Conclusion

In this work, we introduce a novel fusion module designed to combine useful information from various modality combinations. We also propose a new model called MMSFormer that integrates the proposed fusion block to accomplish multimodal material and semantic segmentation tasks. Experimental results illustrate the model’s capability to efficiently fuse information from different combination of modalities, leading to new state-of-the-art performance on three different datasets. Experiments also show that the fusion block can extract useful information from different modality combinations that helps the model to consistently outperform current state-of-the-art models. Starting from only one input modality, performance increases gradually as we add new modalities. Several ablation studies further highlight how different components of the fusion block contribute to the overall model performance. Ablation studies also reveal that different modalities assist in identifying different types of material classes. However, one limitation of the proposed model is the use of modality specific encoders and the number of encoders grows with number of modalities. Future work will include exploring the possibility and effectiveness of using a shared encoder for all the modalities, investigating and extending the model’s capability with other modalities and multimodal tasks.

Chapter 3

Robust Multimodal Learning with Missing Modalities via Parameter-Efficient Adaptation

3.1 Introduction

Multimodal learning (MML) [8, 197] is a general framework for processing, combining, and understanding information from multiple, diverse data sources. Fusing knowledge from multiple modalities (e.g., text, images, audio, and sensor data) is expected to provide more accurate and reliable systems. In recent years, MML has achieved remarkable success in a wide range of applications, including image segmentation [18, 178, 148], captioning [220, 205], classification [51, 152], sentiment analysis [161, 84], and autonomous driving [194, 150]. In all these applications, one often encounters situations where some modalities are corrupted or missing due to hardware limita-

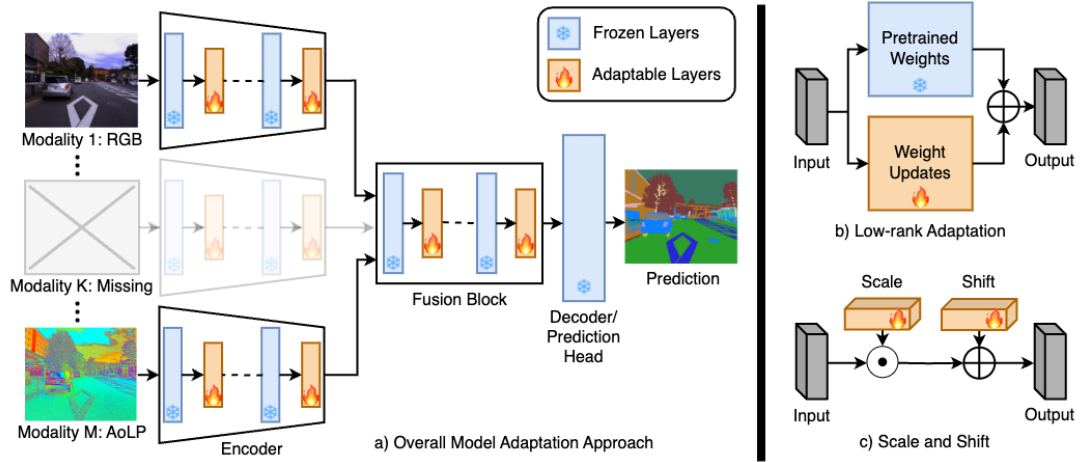


Figure 3.1: a) Overview of our model adaptation approach. A model pretrained on all the modalities is adapted using a small number of learnable parameters to handle different modality combinations. We insert adaptable layers after each layer of the encoders and the fusion block to learn the modulation as a function of the available input modalities to compensate for the missing modalities. The grayed-out branch (missing modality) is inactive and does not contribute to the output. b) Low-rank model adaption computes features using frozen weights and low-rank weight updates and combine them. c) Scale and shift adaptation transforms input by element-wise multiplication and addition.

tions/failures, privacy concerns or data acquisition cost/constraints. The ability to handle corrupt or missing modalities is thus crucial for the robustness and reliability of multimodal systems. However, most of the existing multimodal models are not designed to handle corrupt or missing modalities. The primary focus of this paper is to study and enhance robustness of existing multimodal models in different missing modality scenarios.

Recent studies [124, 61, 130] have shown that MML is not inherently robust to missing modalities and performance can drop significantly when modalities are missing at test time. Existing approaches for robust MML usually work for specific combinations of modalities they are trained for and tend to perform poorly when applied to untrained combinations. For instance, one approach is to adopt robust training strategies such as modality dropout during training [136, 77], partial or full modality masking [5, 157], and knowledge distillation [166, 127]. These approaches either

require specialized training strategies or utilize extra models/sub-networks to guide the underlying model. Another approach replaces uninformative tokens with aggregated informative tokens from different modalities or learns to predict tokens for the specific missing modalities [178, 186, 157]. Training such separate (independent) networks for every possible modality combination is not feasible specially when the number of input modalities is large. One recent approach for robust MML is to impute missing modalities from the available modalities [203, 155, 33]. Performance of these methods depend on the generation model that imputes the missing modalities.

In this work, we propose a parameter-efficient approach to adapt existing multimodal networks to perform well on different missing modality scenarios. **Our main objective is to modify the network in a controllable manner as a function of available modalities.** For instance, if a modality is missing, we seek to modify how the features from available modalities are extracted and fused for the inference. Instead of learning an independent network for each modality combination, our goal is to perform parameter-efficient adaptation. Figure 3.1 illustrates our proposed method, where a given multimodal network can be adapted to arbitrary modality combinations by transforming the intermediate features of the available input modalities at different layers. To achieve parameter-efficient adaptation, we propose to use simple linear transformations such as scaling and shifting to modulate the intermediate features or low-rank increments of features to compensate for the missing modalities. Our method does not require retraining the entire model or any specialized training strategy. The adapted networks provide significant performance improvement over the multimodal networks trained with all modalities and tested with missing modalities. Performance of the adapted models is also comparable or better than the models that are exclusively trained for each input modality combination as shown in Table 3.1. We present a series of experiments to evaluate our

method and compare with existing methods for robust MML on five multimodal tasks across seven datasets (Section 3.4.1). We tested different parameter-efficient adaptation strategies and found intermediate feature modulation with scaling and shifting provides overall best performance, which is discussed in Section A1.4. Our method shows significant performance improvement with less than 1% additional learnable parameters as discussed in Section 3.4.7. The main contributions of this work can be summarized as follows.

- We propose parameter-efficient adaptation procedure for multimodal learning that is robust to missing modalities. The adapted model can easily switch to different network states based on the available modalities with minimal latency, computational, or memory overhead.
- The adapted networks provide notably improved performance with missing modalities when compared to models trained with all modalities and is comparable to or better than the networks trained for specific modality combinations (Table 3.1).
- Our approach is versatile and adaptable to a wide range of multimodal tasks, datasets and models. Detailed evaluations on different datasets and tasks show that our approach outperforms existing baseline methods and robust models designed for specific tasks and datasets (Section 3.4.3 - Section 3.4.6).

3.2 Related Work

Multimodal learning with missing modalities has been studied for different applications in recent years. For instance, robustness in vision-language tasks with multimodal transformers in [124], multimodal sentiment analysis in [61], multimodal classification in [94], and multimodal ac-

tion recognition in [186]. These studies have shown that the task performance can drop significantly when modalities are missing during test time.

Robust training strategies have been proposed to make models robust to different missing modalities. Such approaches include modality dropout during training [136, 77], unified representation learning [93], and supervised contrastive learning [46]. Modality masking during training has become a popular choice for enhancing robustness. [157] utilized complementary random masking, [65] used masked auto encoder, and [124] applied masked cross-modal attention for enhancing missing modality robustness. [61] proposed noisy perturbation of modalities during training for robust multimodal sentiment analysis. Recently, [101] proposed uni-modal ensemble with modality drop and substitution augmentation during training to adapt to different missing modality scenarios.

Design of robust models and fusion strategies is another approach for robust MML. [39] proposed a recursive meshing technique called SpiderMesh and [157] designed complementary random masking (CRM) and knowledge distillation based framework for robust RGB-thermal semantic segmentation. [178] proposed TokenFusion to dynamically detect and replace uninformative tokens with projected tokens from other modalities for robust RGB-depth semantic segmentation, image-to-image translation, and 3D object detection. [171] proposed a model that learns modality-shared and modality-specific features for robust brain tumour segmentation. [24] proposed a robust fusion strategy for multimodal classification. The main limitation of these methods is that they are generally designed for a specific modality combination and do not perform well when applied to other multimodal tasks [107].

Knowledge distillation and generation methods have also become popular for robust MML. Studies by [155] and [203] used GAN based generative models while [33] used VAE based

generative models for imputing missing modalities from available input modalities for underlying multimodal tasks. Recently [147] introduced an approach to learn missing modality tokens from available modalities. Different knowledge distillation approaches have also been applied in several multimodal tasks. [166] proposed mean teacher and [127] introduced multimodal teacher for semi-supervised image segmentation. [157] and [61] applied self-distillation loss for robust RGB-thermal semantic segmentation. Apart from these approaches, weight space ensembling [189], policy learning [124], optimal transport based approach [59] and optimal fusion strategy designing [127] were also studied for robust MML for various tasks.

These approaches are either designed for specific tasks/modality combinations [178, 157, 39] or require training extra modules/sub-networks [166, 127] for guiding the model under different missing modality scenarios. Our goal is to design a generic framework that is parameter-efficient and applicable to any model and modality combinations.

Parameter-efficient network adaptation has become very popular in recent years [70, 109]. A number of parameter-efficient methods have been proposed for transfer learning [64, 30] and uni-modal domain/task adaptation [71, 104]. We can divide the approaches into the following two major categories:

Low-rank/additive adaptation has been applied for uni-modal model fine-tuning and domain adaptation. For instance, LoRA [71], QLoRA [28], KronA [37], and KAdaptation [67] learn low-rank factors for task/domain adaptation. Let W be the weight matrix of any dense layer of a given pretrained uni-modal model. These approaches learn a low-rank weight update matrix ΔW to transform the input x to that layer as $h = Wx + \Delta Wx$, where h is the updated feature. Since the

update matrix ΔW is low-rank, the number of learnable parameters remains a fraction of the total number of model parameters.

Feature modulation based approach is another parameter-efficient method to transform intermediate features of the pretrained model [80, 4, 193, 104]. As shown in Figure 3.1c, it applies a linear transformation to the given input token/feature with learnable scale (γ) and shift (β) parameters. Given an input token x , this approach generates the output token as $h = \gamma \odot x + \beta$, where γ, β, x, h are vectors of same dimension and $\odot$ represents element-wise multiplication along the embedding dimension. These scale (γ) and shift (β) parameters are input-independent and learned during the training process to help the model adjust and fine-tune its representations for better performance on the underlying task.

Though parameter-efficient adaptation approaches have shown great potential in transfer learning, model fine tuning and task/domain adaptation, their potential remains unexplored in the context of missing modality in MML. In this study, we focus on parameter-efficient approaches to build our generic framework to enhance missing modality robustness in MML.

3.3 Proposed Method

In this section we first present a general framework for network adaptation for missing modalities. Then we discuss why we focus on parameter-efficient adaptation, present details of our proposed approach for missing modality adaptation and highlight the key benefits of our approach.

3.3.1 Network Adaptation for Missing Modalities

Let us denote the set of input modalities for a given multimodal task as $\mathcal{M} = \{m_1, \dots, m_M\}$.

Given the full set $\mathcal{M}$, we can train a model f with parameters $\Theta_{\mathcal{M}}$ that maps inputs from all the modalities (denoted as $\mathcal{X}_{\mathcal{M}}$) to an output $y_{\mathcal{M}}$ as

$$y_{\mathcal{M}} = f(\mathcal{X}_{\mathcal{M}}; \Theta_{\mathcal{M}}). \quad (3.1)$$

While we can ensure the availability of all input modalities during training, it is possible that some modalities may be inaccessible at test time, especially after real-world deployment. Any subset of modalities $\mathcal{M}$ can get missing due to hardware failure, data acquisition cost or privacy concerns. If we use a model trained on all the input modalities as denoted by (3.1), significant performance drop is observed when a subset of modalities gets missing during test time as shown in Table 3.1.

Naïve approach

When a subset of the modalities $\mathcal{M}$ is missing, a simple and naïve approach is to train a new model for the available input modalities. Without loss of generality, suppose $\mathcal{K} \subset \mathcal{M}$ represents missing modalities. We can use the available input modalities $\mathcal{S} = \mathcal{M} \setminus \mathcal{K}$ to retrain the model f for a new set of parameters $\Theta_{\mathcal{S}}$ as

$$y_{\mathcal{S}} = f(\mathcal{X}_{\mathcal{S}}; \Theta_{\mathcal{S}}), \quad (3.2)$$

where $\mathcal{X}_{\mathcal{S}}$ represents input data for modalities in $\mathcal{S}$. In principle, we can train one model for every possible $\mathcal{S} \subset \mathcal{M}$ and use the corresponding model at the test time. Such an approach is infeasible because of computational and storage resources required to train models for a large number of possible modality combinations. Furthermore, deploying a large number of trained models and

selecting one of them at test time is not feasible in real-world scenarios. Another drawback of this method is that, even though we would like $y_S \approx y_M$, the training process mentioned earlier does not guarantee it.

Parameter-efficient approach

We propose an alternative approach to adapt a single model for all subsets of input modalities $S \subset \mathcal{M}$ in a parameter-efficient manner. First, we select a model f trained on the full set of modalities $\mathcal{M}$ as shown in (3.1) and freeze the parameters Θ_M . Then we learn a small number of parameters Δ_S , specific to the available input modality set S , and update the model as

$$\hat{y}_S = f(\mathcal{X}_S; \Theta_M, \Delta_S), \quad (3.3)$$

where $\hat{y}_S$ represents the prediction of the updated model. Our goal is to keep $\hat{y}_S$ close to all modality prediction y_M in the best case ($\hat{y}_S \approx y_M$) and close to the prediction y_S made by a model trained on the available input modalities in the worst case ($\hat{y}_S \approx y_S$).

The adaptation method shown in (3.3) is considered parameter-efficient if the number of parameters in Δ_S is significantly smaller compared to the total number of parameters in Θ_M . During adaptation, we keep Θ_M frozen and demonstrate that less than 1% of the total parameters for Δ_S are sufficient for network adaptation (Section 3.4.7).

Need for parameter-efficient adaptation

In recent years, a number of approaches have been proposed for MML with missing modalities. To the best of our knowledge, parameter-efficient adaptation is still unexplored in this field. The current methods for robust MML, as discussed in Section 3.2, require retraining the whole

model with specialized training strategy [157, 77] or utilize extra module/sub-network to guide the multi-modal model [127, 166]. Furthermore, these methods are not very generic and do not perform well on different missing modality scenarios as shown in Table 3.2 and 3.3. To solve these issues, we propose parameter-efficient adaptation for enhancing missing modality robustness of MML. Our approach requires learning a very small number of parameters for different missing modality scenarios without the need to retrain the whole network. Furthermore, it is also applicable to diverse model architectures, tasks and modality combinations as discussed in Section 3.4.

3.3.2 Parameter-Efficient Adaptation for Robust MML

This section outlines our approach for multimodal network adaptation for missing modalities. We explain the reasons behind selecting intermediate feature modulation and compare with other parameter-efficient methods, highlighting key benefits of our approach.

Adaptation for multimodal models. To the best of our knowledge, no parameter-efficient adaptation approach has been proposed or applied for multimodal model adaptation to handle missing modalities. We draw our motivation from low-rank adaptation [71, 9, 4, 80] and feature modulation based approach [104]. These approaches can enhance the representation capabilities of deep models. We extend these adaptation approaches to build a generic framework that can transform the intermediate features of the available modalities to find an optimal feature representation to compensate for the performance gap due to missing modalities.

Training: model adaptation for missing modalities

Our approach is illustrated in Figure 3.1. Without loss of generality, let us assume a generic multi-modal model in which each modality goes through a separate encoder for feature

extraction, followed by a fusion block to fuse the extracted features. The fused feature is passed to a decoder head for making prediction. This setup can be easily generalized to models with shared encoder, different encoder/model architecture and/or different (early or mid) fusion strategy.

We train this multimodal network f with all available modalities in $\mathcal{M}$ to learn the parameters $\Theta_{\mathcal{M}}$ as shown in (3.1). Then we adapt f for different subsets of available modalities $\mathcal{S} \subset \mathcal{M}$. Unlike existing methods, we do not try to generate [125, 5], approximate [178, 171] or distill knowledge [157, 127] from any other modality/sub-network. Our goal is to learn a modified function for the available input modalities to appropriately learn and fuse features to compensate for any missing modality. Instead of re-training the entire network on the available modalities as shown in (3.2), we adapt the base network f and focus on learning a minimal set of parameters following (3.3).

To adapt the base model f , as shown in Figure 3.1a, we freeze the parameters $\Theta_{\mathcal{M}}$ (marked as ❄️ in light blue rectangles), which freezes all the layers in the model. Then we insert adaptable layers with learnable parameters $\Delta_{\mathcal{S}}$ (marked as 🔥 in light orange rectangles) after each frozen linear, convolutional, and norm layers. We show the missing modality branches as grayed-out indicating that they are inactive and do not contribute to the model output. Then we adapt f following (3.3) to learn $\Delta_{\mathcal{S}}$. While learning $\Delta_{\mathcal{S}}$ for a given modality combination, $\mathcal{S}$, we set the missing modalities to zero following standard practice [77, 127, 157, 107]. We minimize the cross-entropy loss with respect to $\Delta_{\mathcal{S}}$ for different modality combinations.

Below we discuss how to use low-rank and intermediate feature modulation-based adaptation to accommodate missing modalities. Our framework is generic and can also incorporate other parameter-efficient adaptation approaches.

Low-rank/additive adaptation. We extend low-rank/additive approaches to adapt multimodal model for missing modalities. Let us assume that W_m be one of the weight matrices from any layer for the m^{th} input modality where $m \in \mathcal{S}$. As shown in Figure 3.1b, we learn a low-rank weight update matrix ΔW_m for that layer to transform the input $h_{m,i}$ to the layer as

$$h_{m,o} = W_m h_{m,i} + \Delta W_m h_{m,i}, \text{ for all } m \in \mathcal{S}, \quad (3.4)$$

where $h_{m,o}$ is the transformed output feature that is passed to the next layer in the model. Since ΔW_m is low-rank, the total number of learnable parameters remains a fraction of the total number of model parameters. We can represent the learnable parameters $\Delta_{\mathcal{S}} = \{\Delta W_m\}_{m \in \mathcal{S}}$ as the collection of all low-rank update matrices.

Intermediate feature modulation. We extend SSF [104] method to work with multimodal models with missing modalities. The adaptable SSF layers modulate the intermediate tokens/features from each available modality at every layer as shown in Figure 3.1c. For the m^{th} input modality where $m \in \mathcal{S}$, we denote the learnable scale and shift parameters as $\gamma_m \in \mathbb{R}^d$ and $\beta_m \in \mathbb{R}^d$ respectively where d is the embedding dimension of the model. The output $h_{m,o} \in \mathbb{R}^{N \times d}$ from any frozen layer for the m^{th} input modality goes through the SSF layer that follows it. The SSF layer applies a linear transformation on $h_{m,o}$ as follows:

$$h_{m,i} = \gamma_m \odot h_{m,o} + \beta_m, \text{ for all } m \in \mathcal{S}, \quad (3.5)$$

where $h_{m,i} \in \mathbb{R}^{N \times d}$ is the transformed feature which is fed to the next frozen layer in the model and N is the number of tokens. Note that if the output of any layer is of shape $(H \times W \times d)$ (for convolutional layers), we reshape it to $(N \times d)$, where $N = H \times W$, before applying (3.5). We reshape the transformed feature back to the original shape (if required) before passing it to the

next layer. We can represent the learnable parameters as $\Delta_{\mathcal{S}} = \{\gamma_{\mathcal{S}}, \beta_{\mathcal{S}}\} = \{\gamma_m, \beta_m\}_{m \in \mathcal{S}}$. BitFit [9] method can also be used for adaptation as we only need to learn the bias/shift terms β_m for all $m \in \mathcal{S}$. We modify (3.5) as

$$h_{m,i} = h_{m,o} + \beta_m, \text{ for all } m \in \mathcal{S}, \quad (3.6)$$

and the learnable parameters can be represented as $\Delta_{\mathcal{S}} = \{\beta_m\}_{m \in \mathcal{S}}$. Thus the intermediate features from each available modality are modulated to find a better representation to compensate for the missing modalities.

Inference: model adaptation for missing modalities

At the test time, we load the base multimodal model f with the pretrained weights $\Theta_{\mathcal{M}}$. If all the modalities are available, then we can use $\Theta_{\mathcal{M}}$ to make predictions. When a subset of the modalities are missing, we can select the learned parameters $\Delta_{\mathcal{S}}$ corresponding to the available input modalities $\mathcal{S}$, insert them into the model and use them to make prediction as follows:

$$\hat{y}_{\mathcal{S}} = \begin{cases} f(\mathcal{X}_{\mathcal{S}}; \Theta_{\mathcal{M}}) & \text{if } \mathcal{S} = \mathcal{M}, \\ f(\mathcal{X}_{\mathcal{S}}; \Theta_{\mathcal{M}}, \Delta_{\mathcal{S}}) & \text{if } \mathcal{S} \subset \mathcal{M}. \end{cases} \quad (3.7)$$

Since we are inserting the adaptable layers after each layer, it does not require any major change to the model architecture and can be done easily without reloading all the model parameters $\Theta_{\mathcal{M}}$. We just need to load the parameters in $\Delta_{\mathcal{S}}$ and insert them into the model. Since we only insert a very small number of additional parameters, it adds very limited computational overhead. Furthermore, if a different subset of modalities becomes available, the adjustment is straightforward. We only need to replace the existing learned parameters $\Delta_{\mathcal{S}}$ with the corresponding parameters for

the available modality set, ensuring adaptability and flexibility in handling diverse combinations of available modalities during the testing phase.

We only insert adaptable layers in the encoders and fusion blocks, while keeping the decoder/prediction head unchanged. We observed that using pretrained decoder/prediction head provided a good overall performance with several missing modalities.

Feature modulation vs low-rank adaptation

While we present three adaptation approaches in (3.4), (3.5), and (3.6), we select intermediate feature modulation with SSF (3.5) as the main approach for our experiments. We primarily selected this technique because of its simplicity and effectiveness. Our experiments show that feature transformation via simple linear transformation with SSF works well for most of the scenarios compared to other parameter-efficient adaptation approaches as summarized in Table A1.3. We provided a detailed comparison in terms of mean accuracy, F1 score and % mIoU in Table A1.4, A1.5 and A1.6 in the supplementary section. SSF shows great promise in enhancing representation power [4], faster convergence [80], prevents loss of information in the representation learning process [193] and mitigates distribution mismatch between the upstream and downstream tasks [104]. These characteristics motivated us to extend this method for MML with missing modalities and build a generic framework that is very effective in learning the proper modulation of available input modalities to bridge the performance gap in the face of missing modalities.

Some key benefits of this approach are as follows. First, The parameters $\{\gamma, \beta\}$ are independent of the input features/modalities, which makes it applicable to diverse tasks and input modality combinations. Second, we can easily insert these learnable layers in existing models without changing the model architecture. We can easily switch/select the corresponding SSF parameters

for a given input modality combination. Finally, it introduces extremely small number of additional learnable parameters. The resulting adaptation offers significant savings compared to training a separate model for each input combination or retraining the model using some specialized training strategy like modality dropout [77, 136] or knowledge distillation [157, 127].

3.4 Experiments and Results

We performed detailed experiments to evaluate the effectiveness and generalizability of our proposed method on five multimodal tasks across seven datasets. In this section, we present comparison with existing baseline methods that are robust to missing modalities.

3.4.1 Datasets and Tasks

In this section, we provide a brief description of each dataset. Please refer to Section A1.1 in the supplementary materials for comprehensive details on each dataset.

Multimodal semantic segmentation.

MFNet dataset [55] contains 1569 registered RGB-Thermal image pairs and divided into train and test sets. Each image is 640×480 pixels, contains annotation for 9 classes.

NYUDv2 dataset [159] has 1449 pairs of aligned RGB-Depth image pairs. It is divided into train and test sets having 795 and 654 image pairs respectively. Each image is 640×480 pixels and contains annotation for 40 classes. We used HHA encoded images [54] instead of raw depth maps for our experiments.

Multimodal material segmentation

MCubeS dataset [106] has 4 input modalities: RGB, Angle of Linear Polarization (AoLP), Degree of Linear Polarization (DoLP) and Near-Infrared (NIR). The dataset is divided into train, validation and test sets containing 302, 96 and 102 sets of images respectively along with ground truth per-pixel annotation for 20 material classes.

Multimodal action recognition

NTU RGB+D (NTU60) dataset [154] contains 56,880 video samples across 60 action classes. It contains RGB videos (1920×1080), depth map sequences (512×424), infrared (IR) videos (512×424) and 3D skeletal data. We use RGB and depth data for our experiments and evaluate performance using cross subject protocol.

Multimodal sentiment analysis

CMU-MOSI dataset [210] contains audio, visual and text modality for multimodal sentiment analysis. The dataset has 2,199 samples divided into train, validation and test containing 1,284, 229 and 686 samples respectively.

CMU-MOSEI dataset [6] is another large scale dataset. It contains 23,453 samples having audio, visual and text. The dataset is divided into train, validation and test sets.

Multimodal classification

UPMC Food-101 dataset [177] is a popular multimodal classification dataset containing image and text as input modalities. The dataset contains 90,704 image-text pairs and 101 food categories.

3.4.2 Implementation Details

We use CMNeXt [215] as the base model for multimodal segmentation tasks, multimodal transformer [168] for multimodal sentiment analysis, UMDR [230] for multimodal action recognition and ViLT [90] for multimodal classification. We train the corresponding base model with all the input modalities for each dataset. To evaluate performance with missing modalities, we provide the available modalities and set the missing modalities to zero for images and empty string for texts. To perform model adaptation for any modality subset $\mathcal{S} \subset \mathcal{M}$, we fine tune the learnable parameters until convergence for all the tasks.

For multimodal segmentation tasks, we set the learning rate to 6×10^{-5} and apply polynomial learning rate scheduler with power = 0.9. The first 10 epochs are warm-up epochs and the learning rate is set to 0.1 times the original rate. The scale (γ) and shift (β) parameters were initialized with all 1s and 0s respectively. We use cross-entropy loss and AdamW optimizer [120], with $\epsilon = 10^{-8}$ and weight decay = 0.01. We set batch size to 4 and report single scale performance. All other hyper-parameters are the same as [215]. For multimodal sentiment analysis, action recognition and classification tasks, we used the default settings from [207], [230] and [94] respectively. Please refer to Section A1.2 in the supplementary materials for additional details.

For every task/dataset, we show the reported results from prior works where possible. It is important to note that, because of this criteria, some of the baseline methods may only be present in specific experiments depending on the availability of their reported numbers. We also perform detailed comparison of SSF with other parameter-efficient adaptation techniques which we discuss in Section A1.4 in the supplementary materials.

Table 3.1: Performance comparison with different baseline methods for multimodal semantic segmentation on MFNet and NYUDv2 datasets and multimodal material segmentation on MCubeS dataset. We use CMNeXt as the base model. **Bold** letters represent best results.

Dataset	Input Modalities	Missing Modalities	Pretrained Models	Modality Duplication	Dedicated Training	Adapted (Ours)
MFNet	RGB-Thermal	-	60.10	60.10	60.10	60.10
	RGB	Thermal	53.71	52.33	55.86	55.22
	Thermal	RGB	35.48	44.43	53.34	50.89
NYUDv2	RGB-Depth	-	56.30	56.30	56.30	56.30
	RGB	Depth	51.19	46.19	52.18	52.82
	Depth	RGB	5.26	13.94	33.49	36.72
MCubeS	RGB-AoLP-DoLP-NIR	-	51.54	51.54	51.54	51.54
	RGB-AoLP-DoLP	NIR	49.06	49.93	49.48	51.11
	RGB-AoLP	DoLP-NIR	48.81	49.23	48.39	50.62
	RGB	AoLP-DoLP-NIR	42.32	48.96	48.11	50.43

3.4.3 Experiments on Multimodal Segmentation

In this section, we present experimental results for multimodal semantic and material segmentation. First, we show an overall comparison of our approach with baselines methods and then we compare with existing robust methods.

Overall performance comparison

We report experimental results for different baseline methods in Table 3.1. **Pretrained** model refers to the base CMNeXt model trained with all the available modalities. **Modality Duplication** means that one of the available modalities is used as a substitution for the missing modality. **Dedicated** training indicates that we train one CMNeXt model for each input modality combination

and use the model corresponding to the available modalities when modalities get missing. **Adapted** model refers to the model that is adapted using our approach for each modality combination.

Pretrained model show significant performance drop with missing modalities. We see a 6.39% and 5.11% drop when Thermal is missing on MFNet and Depth is missing on NYUDv2, respectively, compared to the case when all modalities are available. The effect is amplified when RGB gets missing as we observe 24.62% and 51.04% drop on MFNet and NYUDv2 dataset respectively. On MCubeS dataset, we observe 2.48–9.22% drop in pretrained model when different modality combinations are missing. Similar trend of performance drop is observed for modality duplication approach though it performs better than pretrained models for most of the cases.

The overall performance of the Adapted model is significantly better than Pretrained model and Modality Duplication approach. For MFNet, an improvement of 1.51% and 15.41% is observed compared to the Pretrained model when RGB and Thermal are available respectively. The performance of the Adapted models is also close to the Dedicated models. For NYUDv2 dataset, we see 1.63% and 31.46% performance improvement compared to Pretrained model when depth and RGB are missing, respectively. For all input combinations on MCubeS dataset, we see 1.82–8.11% performance improvement compared to the Pretrained model. The Adapted model performs better than Dedicated models on NYUDv2 and MCubeS datasets. Per-class IoU analysis shows that adapted models perform better than pretrained models for most of the classes which provides an overall performance improvement as discussed in Section [A1.5](#).

Feature modulation during adaptation helps the model learn better feature representation and thus it performs better when modalities are missing (Section [3.4.7](#)). Results also indicates that we do not need to train a dedicated network for each modality combination which requires more

Table 3.2: Performance comparison with existing robust methods for MFNet dataset. RGB and Thermal columns report performance when only RGB and only Thermal are available. Average column reports average performance when one of the two modalities gets missing. ‘-’ indicates that results for those cells are not published. * indicates that available code and pretrained models from the authors were used to generate the results.

Methods	Backbone	Parameters (M)	RGB		Thermal		Average	
			mAcc	% mIoU	mAcc	% mIoU	mAcc	% mIoU
FuseNet [63]	VGG-16 [160]	-	11.11	10.31	41.33	36.85	26.22	23.58
MFNet [55]	DCNN [204]	0.73	26.62	24.78	19.65	16.64	23.14	20.71
RTFNet [164]	ResNet-152 [66]	254.51	44.89	37.30	26.41	24.57	35.65	30.94
SAGate [18]	ResNet-50 [66]	110.85	32.01	30.57	13.34	12.51	22.68	21.54
FEANet [26]	ResNet [66]	-	15.96	8.69	58.35	48.72	37.16	28.71
MDRNet [223]	ResNet-50 [66]	64.60	57.11	45.89	41.98	30.19	49.55	38.04
VPFNet [107]	ResNet-50 [66]	-	48.14	41.08	42.20	35.80	45.17	38.44
SpiderMesh [39]	ResNet-152 [66]	151.81	-	39.60	-	50.50	-	45.05
CRM [157]	Swin-S [117]	117.68	-	52.70	-	53.10	-	52.90
CMNeXt [215]*	MiT-B4 [195]	116.56	60.74	53.71	38.18	35.48	49.46	44.60
Adapted (Ours)	MiT-B4 [195]	117.35	67.18	55.22	66.70	50.89	66.94	53.06

time and computation resources. Rather adapting one base model is sufficient to have comparable or even better performance in missing modality scenarios with less time and computational overhead.

Comparison with robust methods on MFNet dataset

We compare the performance of the Adapted model with existing robust models for RGB-thermal semantic segmentation on MFNet dataset in Table 3.2. Results show that the Adapted model offers the best average performance compared to existing baseline methods. Among the robust models, complementary random masking and knowledge distillation based model CRM [157] shows

Table 3.3: Performance comparison with existing robust methods for NYUDv2 dataset. RGB and Depth columns report performance when only RGB and only Depth are available. Average column indicates average performance when one of the two modalities gets missing. * indicates that available code and pretrained models from the authors were used to generate the results.

Methods	Backbone	Parameters (M)	RGB		Depth		Average	
			mAcc	% mIoU	mAcc	% mIoU	mAcc	% mIoU
FCN [119]	VGG-16 [160]	134.00	44.70	31.60	35.70	25.20	40.20	28.40
Dilated FCN-2s [83]	VGG-19 [160]	55.81	47.10	32.30	39.30	26.80	43.20	29.55
AsymFusion [180]	ResNet-101 [66]	118.20	59.00	46.50	45.60	34.30	52.30	40.40
CEN [179]*	ResNet-101 [66]	118.20	51.77	39.59	28.98	19.32	40.38	29.46
TokenFusion [178]*	MiT-B3 [195]	45.92	63.49	49.32	46.83	36.84	55.16	43.08
CMNeXt [215]*	MiT-B4 [195]	116.56	64.10	51.19	8.30	5.26	36.20	28.23
Adapted (Ours)	MiT-B4 [195]	117.35	67.96	52.82	52.42	36.72	60.19	44.77

competitive performance with the Adapted model. The Adapted model performs better when only RGB is available while CRM performs better when only Thermal is available. Notably CRM is designed specifically for RGB-Thermal pairs and requires specialized training approach. In contrast, our approach is generic, applicable to any input modalities and does not require any specialized training technique. Our approach performs significantly better compared to partial masking and recursive meshing based SpiderMesh [39], variational probabilistic fusion based VPFNet [107] and modality discrepancy reduction based MDRNet [223] models.

Comparison with robust methods on NYUDv2 dataset

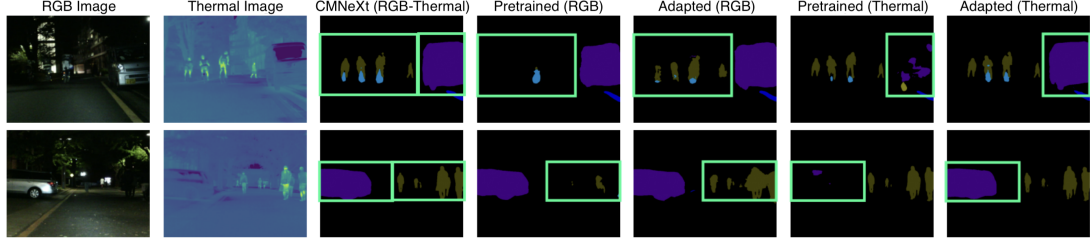
Table 3.3 shows the performance comparison with existing robust models for RGB-Depth semantic segmentation on NYUDv2 dataset. On an average, the Adapted model performs better than the existing robust models. Dynamic token selection and substitution based model TokenFu-

sion [178] performs slightly better (+0.12%) in mIoU when Depth is available and RGB is missing, but shows larger drop (-5.59%) in mean accuracy. On the other hand, the Adapted model performs significantly better (+3.5% mIoU and +4.47% mean accuracy) when RGB is available and Depth is missing. The average performance of the Adapted model is also better than the TokenFusion model despite the fact that TokenFusion was designed to work with RGB-Depth pair, whereas our approach is independent of input modalities. Our method also performs significantly better compared to dynamic channel exchange based CEN [179] and asymmetric fusion based AsymFusion [180] models.

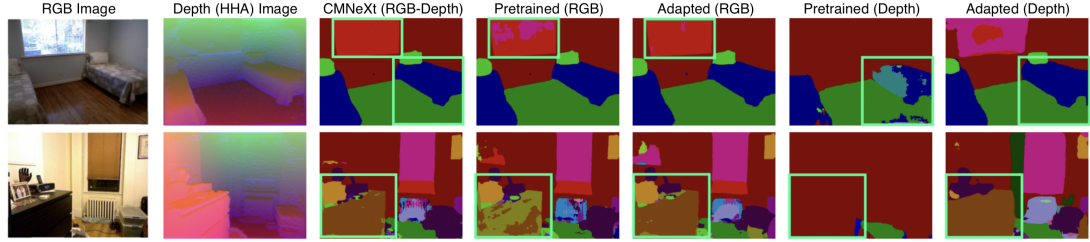
We observe that the CMNeXt model performs poorly when Depth is available and RGB is missing. This is due to its asymmetric architecture, which treats RGB as the primary modality and others as supplementary. As a result, performance drops significantly in the absence of RGB. However, the model overcomes this issue after adaptation and improves performance in all missing modality scenarios demonstrating the effectiveness of our adaptation approach.

Visualization of predictions

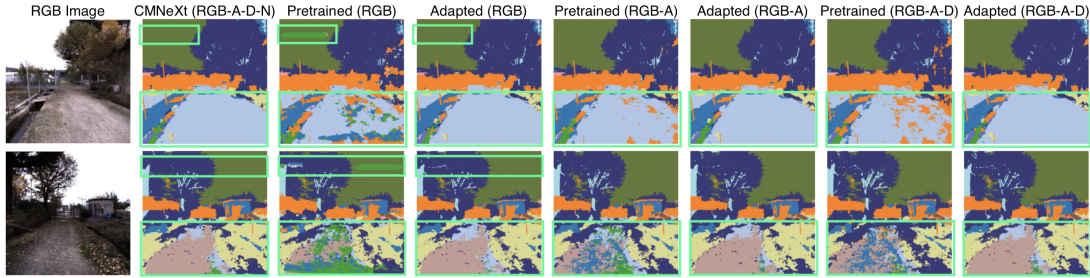
For qualitative analysis, we show some examples of the predicted segmentation maps from the Pretrained and Adapted models in Figure 3.2. For each dataset, we show the input images, predictions when all the modalities are available (CMNeXt column), predictions from the pretrained and adapted models for different available/missing modality scenarios (Available input modality names are shown in parentheses above each image). We see in Figure 3.2a, the Pretrained model fails to detect humans when only RGB images are available and cars when only Thermal images are available. The adapted model can detect both humans and cars with missing modalities.



(a) Visualization of predictions on MFNet dataset for multimodal semantic segmentation



(b) Visualization of predictions on NYUDv2 dataset for multimodal semantic segmentation



(c) Visualization of predictions on MCubeS dataset for multimodal material segmentation

Figure 3.2: Examples of predicted segmentation maps for the Pretrained and Adapted models. Title above each subimage shows method name (available modalities). Segmentation quality improves significantly after model adaptation for all input modality combinations. Green boxes highlight areas with salient differences in results. For MCubeS dataset, we only show RGB input images for brevity. A, D and N denote AoLP, DoLP, and NIR, respectively.

On NYUDv2 dataset, as shown in Figure 3.2b, the Adapted model can detect window, bed, and furniture with higher accuracy than the Pretrained model with missing modalities. On MCubeS dataset, the Adapted model can identify sand, sky, and gravel with higher accuracy than the pretrained model. In all cases, the predictions from the Adapted model with missing modalities

Table 3.4: Performance comparison with existing methods for multimodal sentiment analysis.

Datasets	Methods	Backbone	Parameters (M)	Audio		Visual		Audio-Visual		Average	
				ACC	F1	ACC	F1	ACC	F1	ACC	F1
CMU-MOSI	MuT [168]	Transformer [170]	2.58	48.31	40.98	52.44	51.77	48.93	41.95	49.89	44.90
	MFN [209]	LSTM [69]	2.17	56.86	44.81	55.95	42.94	56.86	51.07	56.56	46.27
	TFN [208]	LSTM [69]	5.04	42.23	25.07	42.38	25.40	42.23	25.07	42.28	25.18
	BERT_MAG [146]	BERT [86]	110.83	57.77	42.31	57.77	42.31	57.77	42.31	57.77	42.31
	LMF [118]	LSTM [69]	1.10	42.23	25.07	43.14	27.54	43.29	27.61	42.89	26.74
	Adapted (Ours)	Transformer [170]	2.60	50.00	46.71	54.88	54.39	55.49	53.96	53.46	51.69
CMU-MOSEI	MuT [168]	Transformer [170]	2.58	37.15	20.12	38.28	23.70	41.91	32.78	39.11	25.53
	MFN [209]	LSTM [69]	2.17	58.48	58.31	60.35	59.48	59.74	60.37	59.52	59.39
	TFN [208]	LSTM [69]	5.04	37.15	20.12	37.15	20.12	37.15	20.12	37.15	20.12
	BERT_MAG [146]	BERT [86]	110.83	62.83	48.50	61.39	49.70	62.83	48.51	62.35	48.90
	LMF [118]	LSTM [69]	1.10	42.38	34.48	57.15	57.85	55.94	56.63	51.82	49.65
	Adapted (Ours)	Transformer [170]	2.60	62.85	55.55	62.49	60.00	63.32	60.69	62.89	58.75

are closer to the predictions of the pretrained model with all the input modalities. We provide additional visualizations in Figure A1.3 in the supplementary materials.

3.4.4 Experiments on Multimodal Sentiment Analysis

We tested our adaptation method for multimodal sentiment analysis on CMU-MOSI [210] and CMU-MOSEI [6] datasets, and report the results in Table 3.4. We use multimodal transformer (MuT) [168] as the base model and adapt it using our approach. We observed that when text is available and either audio or video or both are missing, the test set performance does not drop significantly. Similar trend was reported in [61]. If text is missing at test time, then the performance

Table 3.5: Performance (top-1 accuracy) comparison with existing methods on NTU RGB+D action recognition dataset. RGB and Depth columns report performance when only RGB and only Depth are available. * indicates that available code and pretrained models were used to generate the results.

Method	Backbone	RGB	Depth	Average
Modality Distill. [41]	ResNet-50 [66]	73.42	70.44	71.93
Luo et al. [122]	ResNet-18 [66] + GRU [23]	89.50	87.50	88.50
DMCL [42]	ResNet-18 [66]	83.61	80.56	82.09
Motion-RGBD [231]	DSN + DTN [231]	90.30	92.70	91.50
ActionMAE [186]	ResNet-34 [66] + Transformer [170]	84.50	90.50	87.50
UMDR [230]*	DSN + DTN [231]	90.47	93.99	92.23
Adapted (Ours)	DSN + DTN [231]	91.53	94.29	92.91

of the base MulT model drops significantly. The Adapted models can partially compensate for missing modality and offer significantly better performance compared to the base MulT model.

For CMU-MOSI dataset, we observe 1.69% and 2.44% improvement in accuracy and larger improvement in F1 score over the base MulT model when only audio and only visual are available, respectively. The adapted model offers significant improvement when audio-visual modalities are available and text is missing. It shows 6.56% improvement in accuracy and 12.01% improvement in F1 score over the base MulT model. For CMU-MOSEI dataset, we see even greater improvement in all the metrics. Experiments show 25.7%, 24.21% and 21.41% improvement in accuracy for audio only, visual only and audio-visual scenarios compared to the MulT model. We also observe 27.91%-36.30% improvement in F1 score compared to the base MulT model.

We compare our adaptation method with existing methods for multimodal sentiment analysis. For CMU-MOSI dataset, BERT_MAG works better in terms of accuracy but our adaptation method works better in terms of F1 score. One thing to mention is that BERT_MAG uses a pre-

Table 3.6: Performance (accuracy) comparison with prompting based approach for multimodal classification on UPMC Food-101 dataset. Image and text columns indicate the amount of image and text modality available during both training and testing. † indicates that those values are approximated from the plots published in [94].

Available Modality		ViLT	Attention Level	Input Level	Adapted
Image	Text	[90]	Prompts [94]	Prompts [94]	(Ours)
100%	30%	66.29	72.57	74.53	75.38
30%	100%	76.66	86.05	86.18	88.31
65%	65%	69.25	78.09	79.08	81.77
100%	0%	63.60 [†]	67.70 [†]	68.10[†]	67.66
0%	100%	76.10 [†]	85.30 [†]	84.80 [†]	86.01
Average Accuracy		70.38	77.94	78.54	79.83
Total Params (M)		112.26	112.49	112.49	112.47
Learnable Params (M)		0.0	0.221	0.221	0.207
Change (%)		+0.0%	+0.20%	+0.20%	+0.18%

trained BERT model and finetunes it on the dataset but we are not using any pretraining on extra data. For CMU-MOSEI, our adaptation method works better for most of the cases.

3.4.5 Experiments on Multimodal Action Recognition

We evaluate our approach on NTU RGB+D [154] dataset for multimodal action recognition task. We use UMDR [230] as the base model and adapt it using our approach. As shown in Table 3.5, our adaptation performs better compared to recent modality masking and generation based approach ActionMAE [186] and modality de- and re-coupling based approaches Motion-RGBD [231] and UMDR [230]. Our adaptation shows 7.03% and 1.06% improvement over ActionMAE and UMDR respectively when RGB is available and depth is missing. We see 3.79% and 0.30% improvement over ActionMAE and UMDR respectively when depth is available and RGB is missing.

Moreover, our method outperforms all the existing baseline methods in all the scenarios. Which also indicates that our approach can learn better feature representation compared to modality masking, generation and distillation based approaches.

The base UMDR model has 75.82M parameters. Our adaptation method adds 0.24M additional learnable parameters, which is only 0.32% of the total model parameters. Other methods do not report their total parameter counts, so we omit the total parameters column for this table.

3.4.6 Experiments on Multimodal Classification

To further evaluate the effectiveness of our approach, we compare it with recent prompt based approach missing-aware prompts [94] on UPMC Food-101 [177] dataset. The results are summarized in Table 3.6. For fair evaluation, we use the same experimental setup and evaluation script as [94] to evaluate performance on different available and missing modality scenarios. Image and text columns indicate the amount of image and text modality available during both training and testing. Our adaptation method outperforms prompting based approach for most of the scenarios. On an average, our approach shows 1.29% improvement over the best prompting method and 9.45% improvement over the base ViLT model. These results corroborate the fact that adapting models by intermediate feature modulation helps the model learn optimal feature representation to perform better on different missing modality scenarios.

Efficiency on Parameters. We keep the pre-trained ViLT backbone frozen and compare the additional learnable parameters required for the learnable prompts [94] and our method. We require less additional parameters while performing better than both input level and attention level

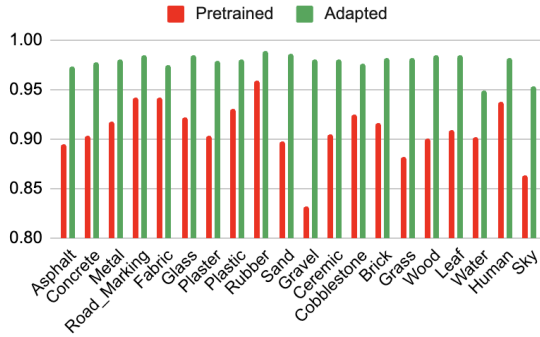


Figure 3.2: (a) Multimodal material segmentation on MCubeS. Available: RGB; Missing: A, D, N

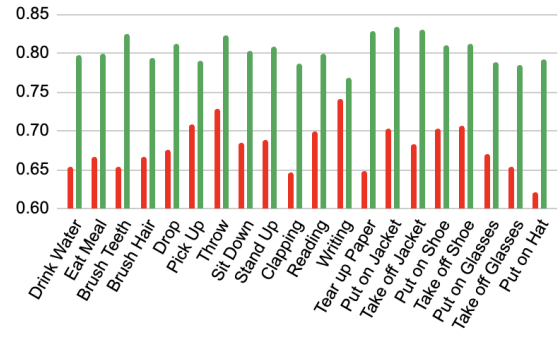


Figure 3.2: (b) Multimodal action recognition on NTU RGB+D. Available: RGB; Missing: Depth

Figure 3.3: Cosine similarity between complete and missing modality features of pretrained and adapted models on MCubeS and NTU RGB+D datasets.

prompts. Thus our adaptation method shows greater parameter efficiency and effectiveness compared to prompt based approach.

3.4.7 Feature and Parameter Analysis

We perform additional analysis to evaluate the effectiveness and generalizability of the adaptation approach. We discuss them in the section.

Why adapted model performs better?

To further analyze how the adaption is helping the model improve overall performance, we conducted cosine similarity analysis of the final fused features extracted from the last layer of the network. To be specific, we calculate the cosine similarity between the complete and missing modality features from the pretrained model (Pretrained), and the cosine similarity between the complete and missing modality features from the adapted model (Adapted). We show the cosine similarities for each class in Figure 3.3.

The adapted model demonstrates a higher cosine similarity to the complete modality features compared to the pretrained model on both MCubeS dataset for multimodal material segmentation and NTU RGB+D dataset for multimodal action recognition when RGB is available and other modalities are missing. This high similarity indicates that the adapted model better retains the essential information from the original complete modality input features, even when some modalities are missing. Consequently, this robustness in feature representation leads to a significant improvement in the model’s overall performance. These results demonstrate the effectiveness of the adapted model in handling scenarios with missing modalities and maintaining robust prediction quality.

We only show first 20 out of 60 classes for NTU RGB+D dataset here. We have included comparison for all the 60 classes including other missing scenarios in Section [A1.6](#).

Performance gain vs learnable parameters

Our method achieves significant performance gains with a small number of additional learnable parameters. As shown in Table [3.2](#) and [3.3](#), adapted models provide 8.46% and 16.54% improvement in mIoU on an average over the base CMNeXt model with only 0.79M additional parameters (i.e., 0.68% of the total model parameters). For multimodal sentiment analysis, as shown in Table [3.4](#), adapted models provide 3.57% and 23.78% improvement in accuracy and 6.79% and 33.22% improvement in F1 score for CMU-MOSI and CMU-MOSEI datasets, respectively over the base MulT model with only 0.02M additional parameters (i.e., 0.775% of the total model parameters). For multimodal classification on UPMC Food-101 dataset, as shown in Table [3.6](#), adapted models achieve an average performance improvement of 9.45% over the base ViLT model with only 0.207M additional learnable parameters (i.e., 0.18% of the total model parameters).

In summary, learning a small number of additional parameters in a base network provides significant performance improvement in the case of missing modalities across all tasks and architectures. The parameter complexity of our approach is comparable/better than existing robust methods like CRM [157] in Table 3.2 and prompts [94] in Table 3.6. However, existing works on missing modality robustness vary widely in terms of model architectures [39, 178], fusion methods [107, 230], training procedures [157, 39], and missing feature generation methods [186]. Due to this heterogeneity, a fair comparison based solely on model size/number of parameters is infeasible.

3.5 Limitations and Future Directions

In this work, our main focus was to enhance missing modality robustness of existing multimodal models. Though our method can make existing models robust to different missing modality scenarios, it has certain limitations. First, we only considered missing modality during test time. However in real life scenarios, modalities can be missing in both train and test times. Second, our method learns one set of adaptation parameters for every combination of missing modalities. While the number of adaptation parameters is small, the overall parameter complexity will scale with the number of modality combinations. For M modalities, we can have up to 2^M possible combinations, as each modality can either be available or missing. Our method will require $2^M - 2$ sets of adaptation parameters to accommodate every possible combination of missing modalities (excluding two cases when all or none of the modalities are available). If we expect one modality out of M to be missing at the test time, which is the case in most of the published work, our method will require M sets of adaptation parameters. Third, we insert the learnable layers after each layer of the encoders and the fusion block. We did not try to optimize the number of parameters or find

the optimal places to insert those learnable layers. Future study will explore these areas to further reduce the number of parameters, enhance the effectiveness and applicability of the approach in newer tasks and datasets.

3.6 Conclusion

Missing modalities at test time can cause significant degradation in the performance of multimodal systems. In this paper, we presented a simple and parameter-efficient adaptation method for robust multimodal learning with missing modalities. We demonstrated that simple linear operations can efficiently transform a single pretrained multimodal network and achieve performance comparable to multiple (independent) dedicated networks trained for different modality combinations. We evaluated the performance of our method and compared with existing robust methods for five different multimodal tasks. Our method requires an extremely small number of additional parameters (e.g., $< 1\%$ of the total parameters), while significantly improving performance compared to existing baseline models and methods for different missing modality scenarios. Our adaptation strategy is applicable to different network architectures, modalities and tasks, which can be a versatile solution to build robust multimodal systems.

Chapter 4

Robust Multimodal Learning via Cross-Modal Proxy Tokens

4.1 Introduction

Multimodal learning [138, 197] integrates information from multiple (diverse) input sources to improve the performance on downstream tasks. For example, examining a movie poster image along with a short text synopsis can provide a greater insight into the movie’s type and genre. In recent years, a number of methods have been proposed to effectively combine information from multiple modalities, including early-stage fusion [133], token-level fusion [178], channel exchange [179], and bottleneck mid-fusion [134]. Furthermore, specialized architectures have been proposed for specific multimodal applications, such as vision-language tasks [90, 99, 121], action recognition [187, 231], and segmentation [216, 148]. Most of these approaches assume that all modalities are available for every sample during both training and inference, yielding better performance than

unimodal baselines. However, their performance drops sharply when one or more modalities get missing [149, 123].

In practical applications, input modalities can be missing during training and/or inference for various reasons; such as lack of synchronous data capture (limited paired data during training) and sensor failure or privacy concerns (modalities missing during inference). [123] observed that multimodal transformers such as ViLT [90] exhibit a drastic performance drop when one or more modalities are missing during inference as shown in Figure 4.1b (*yellow curve*). They developed an optimal fusion strategy to recover the performance loss of multimodal transformers in the missing modality scenarios (*purple curve*). Subsequently, [94] developed a prompt tuning technique (*pink and blue curves*), and [89] combined the prompt-based approach with Variance Covariance Invariance regularization (VICReg) (*brown curve*) to further improve the robustness of multimodal models to missing modalities. Despite these efforts, achieving consistently strong performance under varying degrees of missing modalities remains an open challenge. In this work, we aim to improve robustness across a wide range of missing modality conditions while maintaining strong performance when all modalities are available.

In this work, we introduce a simple yet effective framework to *approximate* missing modalities from the available ones for robust multimodal learning. Specifically, we introduce **cross-modal proxy tokens (CMPTs)**, which learn to approximate the class tokens of missing modalities by attending to all tokens of the available modality, enabling effective handling of various missing modality scenarios. As illustrated in Figure 4.1a, our framework integrates pre-trained unimodal encoders into a multimodal learning setting while introducing CMPTs to partially bridge the performance gap in missing modality scenarios. To ensure efficient learning of CMPTs with

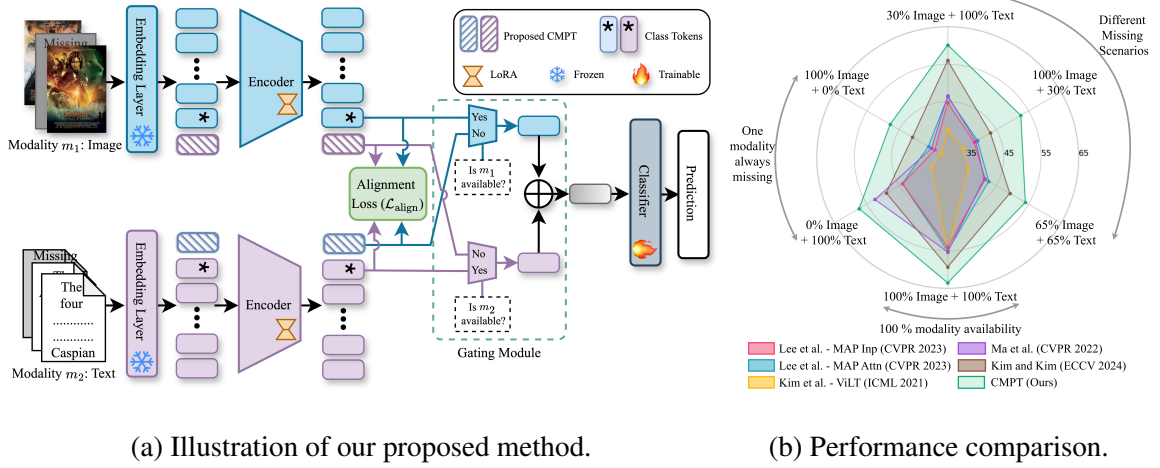


Figure 4.1: (a) We introduce Cross-Modal Proxy Tokens (CMPTs), a novel approach to address missing modality challenges. CMPTs effectively learn to approximate missing modality class tokens by adapting pretrained encoders through a joint optimization of alignment and task-specific objectives. Our approach accommodates both complete and missing modalities during training and inference, thereby enhancing robustness across varying missing modality scenarios. (b) CMPTs achieve state-of-the-art performance, consistently outperforming recent baseline methods in both complete and missing modality scenarios. Following the experimental setup of [94, 89], the radar plot illustrates F1-macro scores on the MM-IMDb dataset across varying modality availability.

minimal additional learnable parameters, we employ low-rank adapters within the frozen unimodal encoders and jointly optimize both an alignment loss and a task-specific loss. The radar plot in Figure 4.1b demonstrates the superior performance of our approach compared to five existing methods [90, 94, 89, 123] across six different complete and missing modality scenarios on the MM-IMDb dataset. Our CMPT-based model consistently outperforms all baselines across all the scenarios. We conduct extensive experiments on five multimodal datasets, comparing our approach against existing state-of-the-art baselines. Results show that our framework achieves state-of-the-art performance in both missing modality (Section 4.4.3) and complete modality (Section 4.4.4) scenarios, while maintaining a minimal number of learnable parameters. Ablation studies further confirm that our method significantly improves performance even in severe missing-modality scenarios and pro-

vides performance enhancement across most of the classes, leading to an overall performance boost (Section 4.4.6). Main contributions of our work are as follows.

- We propose CMPTs to effectively learn an *approximation* of the missing modality from the available one with minimal adaption of pre-trained unimodal encoders via an alignment loss.
- Our approach outperforms 12 recent baseline methods across four multimodal datasets in various missing-modality scenarios, demonstrating state-of-the-art performance (Section 4.4.3).
- Across five datasets, our method achieves on-par or superior performance compared to state-of-the-art techniques when all modalities are present during training and inference (Section 4.4.4).
- We demonstrate that CMPTs generalize effectively to unseen missing rates (Section 4.4.5) and boost performance in most of the classes (Section 4.4.6).

4.2 Related Work

Recent works [123, 61, 149] show that performance can degrade significantly when modalities are missing, as multimodal models are not inherently robust to such scenarios. Several approaches have been proposed to address this issue. In this section, we primarily discuss works where the main focus is on making multimodal models robust to missing modalities. More broader related work on multimodal learning is discussed in appendix Section A2.1.

Robust model design is one of the approaches for handling missing modalities. Recently, [157] developed a robust framework utilizing modality masking and knowledge distillation. [171] proposed ShaSpec to impute missing features through modality-specific and modality-shared fea-

ture learning. [178] proposed TokenFusion, which dynamically replaces uninformative tokens to enhance robustness. Additionally, different fusion strategies [24, 39, 107] have been developed to improve robustness across various missing-modality scenarios. However, many of these approaches are task-specific, limiting their generalization to other multimodal tasks. In our work, we focus on reusing pre-trained unimodal encoders, making it flexible for extension to different input modalities and multimodal tasks.

Robust training and model adaptation can enhance missing modality robustness. Recently, [149] proposed a parameter-efficient adaptation framework for different multimodal tasks. [156] introduced a modality-aware low-rank adaptation method (MoRA) that uses a shared down-projection layer to map inputs to a low-dimensional space, and then applies modality-specific up-projections for missing modality adaptation. [137] trained a single model to infer missing information by projecting available modalities. [125] introduced SMIL, a Bayesian meta-learning method for flexible handling of missing modalities during both training and testing. [123] proposed a multi-task optimization approach with a differentiable algorithm for optimal fusion strategy selection. [127] developed a semi-supervised framework that leverages unlabeled data to improve robustness, while [181] used a teacher network to transfer multimodal information for improved missing modality robustness. These methods, however, either require extensive adaptation or specialized training procedures. In our approach, we employ lightweight adaptation of unimodal encoders (using Rank-1 LoRA adapters) and a single extra loss term to explicitly learn cross-modal relationships. Hence, the overall training flow is mostly left unchanged.

Prompt tuning methods improve missing modality performance by incorporating learnable prompts across multiple layers of the model. [94] introduced missing-aware prompts that learns

a separate set of prompts for each modality combination. Later, [81] simplified the approach, showing that a single set of prompts per modality can achieve similar performance. More recently, [25] proposed a multi-step adaptive prompt learning method through modality alignment, while [89] introduced a read-only prompt tuning approach with Variance Covariance Invariance regularization. Although effective, these methods require learning and storing a large number of prompts, typically 16 prompts [94, 81] per modality per layer, for each missing modality scenario. In contrast, our method uses *a single cross-modal proxy token* per modality to estimate the missing modality feature from the available modality, thereby reducing the complexity.

We carry out extensive experimental comparisons with all the above prompt-based techniques to demonstrate superior performance across various missing modality scenarios. Furthermore, our approach is efficient, uses few learnable parameters, and is robust to missing modalities in both training and testing which we discuss in Section 4.4.

4.3 Proposed Approach

In this paper, we propose a new framework to train multimodal systems such that they retain strong performance under different scenarios of missing modalities during inference. Furthermore, we seek to develop a simple and efficient framework that can be used for different multimodal tasks and datasets without any specialized (task-specific) changes. To achieve these objectives, we propose **cross-modal proxy tokens (CMPTs)**, which effectively learn to approximate the class tokens of missing modalities by adapting pretrained, modality-specific (transformer-based) encoders through a joint optimization of alignment and task-specific objectives, as illustrated in Figure 3.1.

4.3.1 Multimodal Embedding Tokens

Let us consider a multimodal dataset containing two modalities $\mathcal{M} = \{m_1, m_2\}$. We denote the available dataset as $\mathcal{D} = \{\mathcal{D}_c, \mathcal{D}_{m_1}, \mathcal{D}_{m_2}\}$, which can be further divided into three subsets: $\mathcal{D}_c = \{x_{m_1}, x_{m_2}, y\}$ where both modalities are present with corresponding output labels y ; $\mathcal{D}_{m_1} = \{x_{m_1}, \emptyset, y\}$ where only modality m_1 (x_{m_1}) is available with corresponding labels y ; $\mathcal{D}_{m_2} = \{\emptyset, x_{m_2}, y\}$ where only modality m_2 (x_{m_2}) is available with corresponding labels y . Here, $\emptyset$ denotes a placeholder for the missing modality (e.g., empty string for text, zeros for image, audio, or video) following prior works [94, 149].

For each input modality, we assume that a transformer-based encoder, pre-trained on some large dataset, is available. This is a common and effective practice in modern multimodal learning, particularly given the wide variety of high-performing, pre-trained models now readily accessible [29, 34, 48]. Leveraging these modality-specific encoders trained on extensive data leads to improved performance and faster convergence [60, 197]. Specifically, each modality $m \in \mathcal{M}$ is processed through a pre-trained embedding layer, EmbeddingLayer_m , with parameters $\Theta_{e,m}$. The input x_m for each modality is passed through its corresponding embedding layer to generate modality-specific tokens as

$$\mathcal{T}_m = \text{EmbeddingLayer}_m(x_m; \Theta_{e,m}), \quad (4.1)$$

where $\mathcal{T}_m \in \mathbb{R}^{N \times d}$ denotes N tokens, each of dimension d , for modality m . During training, we keep the pretrained parameters $\Theta_{e,m}$ frozen.

4.3.2 Cross-Modal Proxy Tokens (CMPTs)

To compensate for the missing modalities, we introduce cross-modal proxy tokens (CMPTs) that *approximate* the *missing* modality by attending only to the tokens of the *available* modality. For each modality $m \in \mathcal{M}$, we concatenate one CMPT ($\mathcal{T}_{\text{CMPT},m} \in \mathbb{R}^{1 \times d}$) with the class token (CLS) ($\mathcal{T}_{\text{CLS},m} \in \mathbb{R}^{1 \times d}$) and all the modality specific tokens $\mathcal{T}_m$ to generate the modified embedding tokens $\tilde{\mathcal{T}}_m \in \mathbb{R}^{(N+2) \times d}$ as

$$\tilde{\mathcal{T}}_m = \text{Concat}(\{\mathcal{T}_{\text{CMPT},m}, \mathcal{T}_{\text{CLS},m}, \mathcal{T}_m\}). \quad (4.2)$$

Note that the proposed CMPT serves as an extra token, similar to the class token [34, 29, 48], but with a complementary purpose. Class token attends to all the tokens of a given modality to aggregate information and learn a rich feature representation of that modality. In contrast, CMPT attends to all the tokens of a given modality to learn an *approximation* of the class token of the other modality (e.g., $\mathcal{T}_{\text{CMPT},m_1}$ approximates $\mathcal{T}_{\text{CLS},m_2}$ while attending to $\mathcal{T}_{m_1}$).

4.3.3 Learn CMPTs by Alignment and Adaptation

For each modality, we assume that a pretrained transformer encoder f_m with parameters $\Theta_{f,m}$ is available. The encoder for each modality needs to be updated so that they can learn the corresponding CMPTs. To keep the number of trainable parameters minimal, we use low-rank adaptation (LoRA) [72] to learn a small number of parameters $\Delta_{f,m}$. The modified embedding tokens $\tilde{\mathcal{T}}_m$ are passed through the adapted encoder for modality m to get the final output tokens as

$$\hat{\mathcal{T}}_m = f_m(\tilde{\mathcal{T}}_m; \Theta_{f,m}, \Delta_{f,m}), \quad (4.3)$$

where $\hat{\mathcal{T}}_m \in \mathbb{R}^{(N+2) \times d}$ represents the CMPT, CLS, and remaining tokens from the last encoder layer. Note that the pretrained encoder parameters $\Theta_{f,m}$ remain frozen while we only learn $\Delta_{f,m}$.

We aim to train the CMPTs such that they act as a proxy of the class token of the missing modality at the inference time. In particular, we want the CMPT $\hat{\mathcal{T}}_{\text{CMPT},m_1}$ of modality m_1 to approximate the class token $\hat{\mathcal{T}}_{\text{CLS},m_2}$ of modality m_2 , and vice versa. To achieve this, we use an alignment loss $\mathcal{L}_{\text{align}}$, defined as

$$\mathcal{L}_{\text{align}} = \frac{1}{N_{\mathcal{D}_c}} \sum_{x \in \mathcal{D}_c} \mathcal{L}_{\text{MSE}}(\hat{\mathcal{T}}_{\text{CMPT},m_1}, \hat{\mathcal{T}}_{\text{CLS},m_2}) + \mathcal{L}_{\text{MSE}}(\hat{\mathcal{T}}_{\text{CMPT},m_2}, \hat{\mathcal{T}}_{\text{CLS},m_1}), \quad (4.4)$$

where $\mathcal{L}_{\text{MSE}}$ denotes the mean squared error loss and $N_{\mathcal{D}_c}$ denotes the number of samples in the training data where both modalities are present.

4.3.4 Gating Module and Fusion

After computing the final output tokens $\hat{\mathcal{T}}_m$, we extract the class token ($\hat{\mathcal{T}}_{\text{CLS},m}$) and the CMPT ($\hat{\mathcal{T}}_{\text{CMPT},m}$) for both modalities and pass them through the gating module. If both modalities are available, the gating module returns the class tokens from each modality. If one modality is missing, the gating module substitutes the missing class token with the CMPT from the available modality. The returned tokens are then fused using additive fusion to produce the final fused feature token $\mathcal{T} \in \mathbb{R}^{1 \times d}$ as

$$\mathcal{T} = \begin{cases} \hat{\mathcal{T}}_{\text{CLS},m_1} + \hat{\mathcal{T}}_{\text{CLS},m_2}, & m_1, m_2 \text{ are available,} \\ \hat{\mathcal{T}}_{\text{CLS},m_2} + \hat{\mathcal{T}}_{\text{CMPT},m_2}, & m_1 \text{ is missing,} \\ \hat{\mathcal{T}}_{\text{CLS},m_1} + \hat{\mathcal{T}}_{\text{CMPT},m_1}, & m_2 \text{ is missing.} \end{cases} \quad (4.5)$$

The fused feature token $\mathcal{T}$ is then passed through a linear classifier head to make a prediction $\hat{y}$ as

$$\hat{y} = h(\mathcal{T}; \Theta_h). \quad (4.6)$$

Θ_h represents the parameters of the classifier head that are learned during training.

Finally, we learn the low-rank factors for each encoder ($\{\Delta_{f,m}\}$) and classifier head (Θ_h) by minimizing the task-specific loss $\mathcal{L}_{\text{task}}$, which for our purposes is a standard cross-entropy loss for classification tasks, combined with the alignment loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda \mathcal{L}_{\text{align}}, \quad (4.7)$$

where λ is a hyperparameter that we set to $\lambda = 0.20$ in all our experiments. As discussed in Section 4.4.6, this value provides a balance in performance across different scenarios.

4.3.5 Handling Missing Modalities

Our model can handle missing modalities during both training and inference. During training on complete modality data, we apply random modality dropout, where we either use both modalities or randomly drop one in each iteration. This way, the model learns to make correct predictions in missing modality scenarios using the CMPTs from the available modality. When training data contains missing modalities, we learn the CMPTs by optimizing the alignment loss using only the samples with complete modalities, as shown in (4.4). For missing modalities, we employ placeholders (e.g., an empty string for text and zeros for images, audio, and video) following standard practices [94, 149], and use the CMPT from the available modality as an approximation to make predictions, as described in (4.5).

To enhance training efficiency, we leverage LoRA [72] with a rank of 1 to fine-tune the encoders in a parameter-efficient manner. LoRA layers are inserted after the query, key, value,

and output layers in every attention blocks within the encoders. Our experiments demonstrate that LoRA with rank 1 is sufficient to achieve better performance compared to existing state-of-the-art methods, as discussed in Section 4.4.4 and Section A2.4. Our method is also compatible with newer parameter-efficient adaptation techniques such as DoRA [115] and VeRA [92], as discussed in Section A2.10.

4.4 Experiments and Results

4.4.1 Datasets

We evaluate our approach on five popular multimodal datasets across different tasks. A brief description of each dataset is provided here, with further details in Section A2.2.

UPMC Food-101 [177] is a multimodal classification dataset consisting of 101 food classes. It contains 90,704 image-text pairs, divided into a training set of 67,988 and a test set of 22,716 pairs.

MM-IMDb [3] has 25,959 image-text pairs for multi-label movie genre classification across 23 classes. It is split into 15,552 training, 2,608 validation, and 7,799 test samples.

Kinetics-Sound (KS) [2] is a subset of Kinetics400 dataset [85] for human action recognition using audio and video as input modalities. It has 31 classes, with 14,739 samples in the training set and 2,594 samples in the test set.

Audio-Visual Event (AVE) dataset [167] contains 4,143 10-second videos for multimodal event localization divided into train/val/test sets containing 3,339/402/402 samples across 28 categories.

CREMA-D dataset [13] includes 7,442 short video clips from 91 actors expressing six emotion categories. It contains 6,698 training and 744 test samples for multimodal emotion recognition.

4.4.2 Implementation Details

Vision-Language datasets. We use pretrained ViT-B [34] from CLIP model [144] as image encoder and BERT-base-uncased model [121] as text encoder. The max token length is set to 40 for UPMC Food-101 and 256 for MM-IMDb dataset, respectively.

Audio-Video datasets. We use AST model [48] pretrained on AudioSet dataset [44] as audio encoder and ViT-B [34] from CLIP model [144] as video encoder. Raw audio is converted to a 128-bin mel spectrogram at 16 kHz, with a maximum length of 1024 frames. For video, we sample 3 frames randomly while training and uniformly while testing. Other pre-processing steps are similar to [218, 141, 35].

Hyperparameter settings. We use Python 3.8.19 and PyTorch 2.2.2 for training and evaluating our models. All the models are trained using two NVIDIA RTX 2080Ti GPUs. For vision-language datasets, we set the learning rate to 10^{-3} and train the models for 10 epochs with a batch size of 8. For audio-video datasets, the learning rate is set to 5×10^{-5} and models are trained for 100 epochs with a batch size of 4. We use AdamW [120] optimizer with $\epsilon = 10^{-8}$ and weight decay = 0.02. While training, we utilize cross entropy loss, polynomial learning rate scheduler with power=0.9 and set the first 5 epoch as warm-up. The learning rate is set to 0.1 times the original learning rate during warm-up epochs. We set LoRA rank-1 and insert them after query, key, value and output layers of each transformer block. Further details can be found in Section A2.3.

For each task and dataset, we compare against state-of-the-art published work. For this reason, the set of baselines varies across experiments based on the availability of published results. In each table, the best and second best results are bold and underlined, respectively.

Table 4.1: Performance comparison when modality gets missing during both training and inference. We train and evaluate our model with same modality ratio following existing baselines [94, 81, 89] and report the mean and standard deviation (SD) of three independent runs with different seeds.

Datasets	% availability		ViLT	MAP [94]		MSP	VisualBERT	Ma et al.	Kim & Kim	MuAP [25]		CMPT (Ours)
	Image	Text	[90]	Input	Attn.	[81]	[99]	[123]	[89]	Head	Cross	Mean $\pm$ SD
MM-IMDb (F1-Macro)	30%	100%	37.61	46.30	44.74	47.45	38.63	46.63	<u>56.03</u>	47.21	46.73	60.21 $\pm$ 0.40
	65%	65%	36.30	42.66	41.56	42.03	37.23	41.28	<u>49.24</u>	42.57	43.92	54.04 $\pm$ 0.23
	100%	30%	34.71	39.22	38.16	38.34	36.41	38.65	<u>43.21</u>	41.37	39.88	52.61 $\pm$ 0.12
UPMC Food-101 (Accuracy)	30%	100%	76.93	86.18	86.05	86.34	77.41	86.38	<u>87.11</u>	86.90	86.59	87.48 $\pm$ 0.13
	65%	65%	69.03	79.08	78.09	78.89	71.06	78.58	<u>82.67</u>	77.87	78.95	82.83 $\pm$ 0.25
	100%	30%	66.29	74.53	72.57	73.77	67.78	73.41	<u>78.81</u>	74.61	74.60	80.37 $\pm$ 0.18

4.4.3 Performance when Modalities are Missing

Existing literature considers missing modalities at both training and inference or only at inference. Our method achieves state-of-the-art performance in both scenarios, as detailed in the following two sections.

Missing modality during training and inference

In the first setup, a modality is missing at a fixed rate during both training and inference, as studied by [94, 81, 89]. We follow their experiment setup and compare our method against seven state-of-the-art approaches on MM-IMDb and UPMC Food-101 datasets. As shown in Table 4.1, our method consistently outperforms all baselines across various missing modality configurations. Specifically, compared to the most recent approach by [89], we achieve a 4.18%–9.40% improvement in F1-macro score on MM-IMDb dataset and a 0.16%–1.56% improvement in accuracy on UPMC Food-101 dataset across different missing modality scenarios. It is worth noting that while the performance of [89] on UPMC Food-101 dataset is comparable to ours in two scenarios, our

Table 4.2: Performance comparison when a complete modality gets missing during inference. We report % accuracy for all the datasets. Our method shows overall better performance compared to recent state-of-the-art methods.

Datasets	Available Modality	Missing Modality	FiLM	BiGated	OGM-GE	QMF	MLA	CMPT
			[142]	[88]	[141]	[217]	[218]	(Ours)
CREMA-D	Audio	Video	53.89	51.49	53.76	<u>59.41</u>	59.27	67.20
	Video	Audio	18.67	17.34	28.09	39.11	<u>64.91</u>	76.21
	Audio - Video	-	60.07	59.21	68.14	63.71	<u>79.70</u>	88.84
KS	Audio	Video	48.67	49.96	48.87	51.57	<u>54.67</u>	68.27
	Video	Audio	23.15	23.77	29.73	32.19	<u>51.03</u>	85.77
	Audio - Video	-	63.33	63.72	65.74	65.78	<u>71.35</u>	91.21
UPMC Food-101	Image	Text	4.68	14.20	22.35	45.74	<u>69.60</u>	75.66
	Text	Image	<u>85.84</u>	85.79	85.17	84.13	86.47	85.31
	Image - Text	-	87.21	88.87	87.54	92.87	<u>93.33</u>	94.47

method still performs best in every configuration. These results also suggest that the CMPTs can provide good proxy for the missing class tokens, which significantly improves the model robustness. To ensure a reliable evaluation, we run our method with three different random seeds and report the mean and standard deviation (SD) for various missing scenarios. The low SD values indicate the stability and reliability of our method. We also assessed the performance of the models with different distributions of modality ratios at training and testing which is discussed in Section A2.5.

Missing modality during inference only

In the second setup, following the work of [218], we evaluate our model when all modalities are available during training, but one modality is completely missing during inference. As shown in Table 4.2, our method outperforms existing baselines in most cases. For instance, when

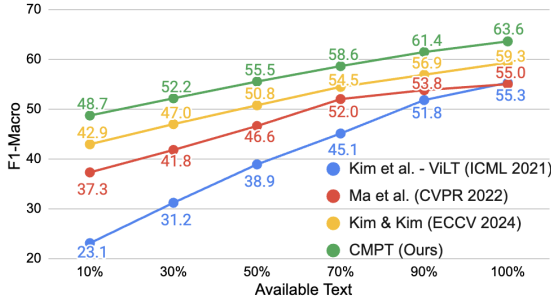
Table 4.3: Performance comparison with both modalities available during training and inference on Image-Text datasets. We report % accuracy for UPMC Food-101 and F1-Macro/F1-Micro score for MM-IMDb dataset. Here, ‘-’ indicate results that are not reported in corresponding papers.

Methods	Parameters (M)		UPMC Food-101	MM-IMDb
	Total	Learnable		
ViLT [90]	112.63	112.63	92.00	55.30 / 64.70
MBT [135]	196.00	196.00	93.56	59.60 / 64.81
MMBT [87]	170.00	170.00	<u>94.10</u>	60.80 / 66.10
MMLoRA [35]	196.10	196.10	93.70	<u>61.70 / 67.20</u>
PMF [102]	198.43	2.54	91.51	58.77 / 64.51
PMF-L [102]	643.95	4.44	91.68	61.66 / 66.72
MLA [218]	218.26	218.26	93.33	- / -
CMPT (Ours)	195.51	0.27	94.47	63.58 / 69.23

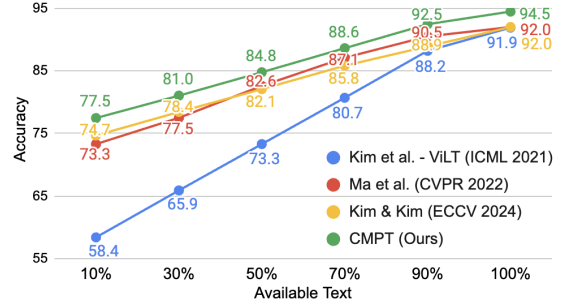
Table 4.4: Performance (% accuracy) comparison with both modalities available during training and inference on Audio-Video datasets. Here, ‘*’ denotes results generated using existing code and ‘-’ indicates that results for those cells are not published.

Method	Parameters (M)		KS	AVE	CREMA-D
	Total	Learnable			
G-Blend [175]	22.36	22.36	62.20	65.50	58.70
FiLM [142]	22.35	22.35	63.33	-	60.07
BiGated [88]	22.35	22.35	63.72	-	59.21
OGM-GE [141]	22.35	22.35	65.74	76.90	68.14
QMF [217]	22.36	22.36	65.78	68.16*	63.71
PMR [40]	22.36	22.36	-	66.40	61.80
MLA [218]	22.37	22.37	71.35	64.68*	79.70
MBT [87]	172.00	172.00	85.00	77.80	-
MMLoRA [35]	172.19	172.19	91.40	<u>96.20</u>	<u>88.60</u>
CMPT (Ours)	172.16	0.17	<u>91.21</u>	96.77	88.84

the audio modality is missing, all methods except MLA [218] experience a significant performance drop on the CREMA-D and Kinetics-Sound (KS) datasets. On the CREMA-D dataset, our method achieves 7.93% and 11.30% improvements over the most recent MLA approach, when video and audio are missing, respectively. Similarly, on the KS dataset, our method outperforms MLA with 13.60% and 34.74% improvements when video and audio are missing, respectively. While MLA performs slightly better when the image modality is missing on the UPMC Food-101 dataset, it suffers a larger performance drop when text is missing. In contrast, our method maintains a balanced performance across different missing modalities. Furthermore, across all datasets, our method achieves the best performance when all modalities are available and demonstrates superior robust-



(a) Dataset: MM-IMDb



(b) Dataset: UPMC Food-101

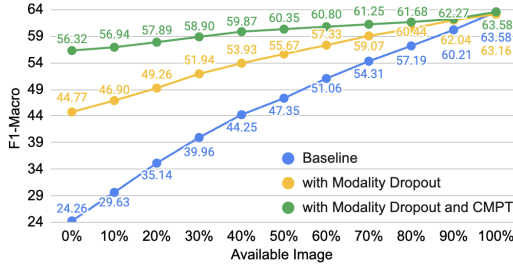
Figure 4.2: Generalization to varying missing rates during inference. Models are trained with 100% image + 100% text and evaluated with 100% image + $x\%$ text following [123, 89]. Our approach demonstrates better generalization, particularly under severe modality loss.

ness compared to other approaches. These results highlight the effectiveness of our method in handling scenarios where a modality is entirely absent during inference.

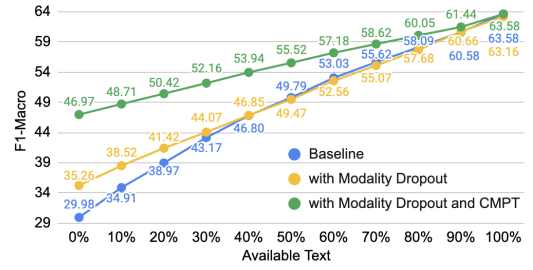
4.4.4 Performance when All Modalities are Available

A method designed to handle missing modalities should not compromise performance when all modalities are present. In this subsection, we demonstrate that our approach also achieves comparable or superior performance to state-of-the-art models across five datasets when all modalities are available.

To evaluate this, we train and test our method when all modalities are available during training and inference. Results in Table 4.3 demonstrate that our proposed method achieves state-of-the-art performance on both UPMC Food-101 and MM-IMDb datasets. Our method surpasses the most recent MLA approach [218] by 1.14% on UPMC Food-101. Our method achieves 1.88% and 2.03% improvement in F1-macro and F1-micro scores, respectively, over second-best performing MMLoRA [35] on MM-IMDb.



(a) Available: Text, Missing: Image.



(b) Available: Image, Missing: Text.

Figure 4.3: Effectiveness of CMPTs on the MM-IMDb dataset. All models are trained with 100% image and 100% text data, then evaluated under varying amount of missing modalities. CMPTs achieve significant performance improvement, especially under severe missing modality scenarios.

We present the results for audio-video datasets in Table 4.4, where we compare our method against nine state-of-the-art methods. Our approach consistently achieves comparable or superior performance across all three datasets, while using the fewest number of learnable parameters. These results highlights that incorporating CMPTs does not introduce any drawback when all modalities remain available.

4.4.5 Generalization to Different Missing Rates

In real-world scenarios, the missing rates for modalities during inference can be unpredictable. Any method designed to handle missing modalities needs to generalize effectively across varying levels of missing rates during inference. To evaluate this capability, we trained our models on complete multimodal data and assessed their performance under varying levels of missing modalities during inference. Specifically, we follow the experiment setup similar to [123, 89] where models are trained with 100% image + 100% text data and evaluated with 100% image + $x\%$ text.

As illustrated in Figure 4.2, our method consistently outperforms all baselines across different missing scenarios, demonstrating superior robustness. We observe a decline in performance

of all methods as the amount of available text decreases, but our approach maintains stronger resilience compared to other methods. For instance, on the MM-IMDb dataset, our method achieves a 5.8% higher F1-Macro score than the most recent method by [89] and a 25.6% improvement over ViLT [90] when only 10% of text is available. A similar trend is observed on the UPMC Food-101 dataset, where our method consistently yields superior performance, confirming its robustness to missing modalities at inference.

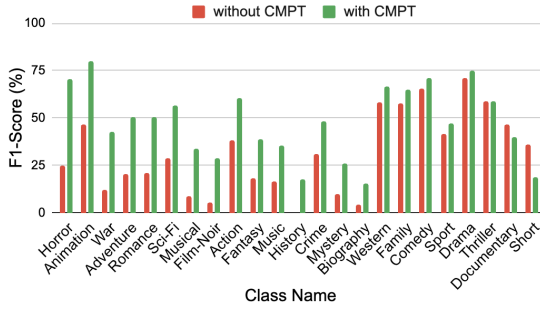
Note that published work by [123, 89, 90] did not report results for different missing rates of images, which prevents us from reporting a direct comparison in that setting. Nevertheless, our results highlight the effectiveness of CMPTs in handling different missing modality scenarios, making it a more generalizable solution for real-world applications.

4.4.6 Ablation Studies

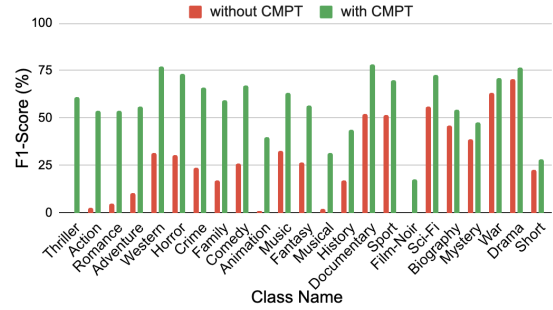
Effectiveness of CMPTs

To evaluate the effectiveness of CMPTs in handling missing modalities, we compare our method against two baselines: (i) a standard model trained without modality dropout or CMPTs (baseline) and (ii) a model trained with modality dropout, where modalities are randomly removed during training to improve robustness to missing data. For a fair comparison, all models were fine-tuned on the downstream task using LoRA with a rank of 1.

Figure 4.3 presents the main results for MM-IMDb dataset, where the baseline model experiences a sharp performance drop as the proportion of missing modalities increases. While modality dropout improves robustness, the performance gap remains large. Incorporating CMPTs further enhance performance, achieving a 32.06% improvement over the baseline and 11.55% over



(a) Available: Image, Missing: Text.



(b) Available: Text, Missing: Image.

Figure 4.4: CMPTs improve performance across most of the classes when modalities are missing. However, it fails to accurately predict certain classes (e.g., short/documentary) in the absence of text, likely due to the image modality’s insufficient information for those classes. Classes are sorted based on the amount of performance improvement.

modality dropout when the image modality is entirely missing as shown in Figure 4.3a. A similar trend is observed when text is missing as shown in Figure 4.3b. Notably, CMPTs become increasingly effective as the amount of missing data increases, demonstrating its strong ability to substantially mitigate performance degradation in the presence of severe modality loss. Similar trend appears in UPMC Food-101 dataset, which we discuss in Sec. A2.6 in the appendix.

Effects of CMPTs on different classes

When analyzing the effectiveness of CMPTs, a key question is whether the overall performance gains are driven by just a few classes or if CMPTs consistently improves performance across most classes. To address this, we conducted a per-class performance analysis with and without CMPTs on the MM-IMDb dataset.

As shown in Figure 4.4, CMPTs improve performance across nearly all classes in both text-missing and image-missing scenarios. Notably, it significantly mitigates performance degra-

Table 4.5: Ablation study on the alignment loss weight λ . We report performance on the UPMC Food-101 and AVE datasets across both complete and missing modality scenarios.

Dataset	Available Modality	Missing Modality	$\lambda = 0.0$	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.3$	$\lambda = 0.4$	$\lambda = 0.5$
UPMC Food-101	Image	Text	73.12	73.48	75.66	73.72	74.39	74.46
	Text	Image	84.31	84.79	85.31	84.51	84.81	84.88
	Image - Text	-	93.96	94.14	94.47	94.13	94.34	94.21
AVE	Audio	Video	84.82	84.33	85.21	84.82	84.33	84.82
	Video	Audio	82.83	83.08	84.58	85.32	81.84	84.83
	Audio - Video	-	95.77	96.02	96.77	96.52	96.52	96.27

dation in modality-sensitive classes. For example, when text is missing, classes like *Horror*, *Animation*, and *War* experience severe drops in performance, but CMPTs effectively recover much of the loss (Figure 4.4a). Similarly, when images are missing, CMPTs provide substantial gains in *Thriller*, *Action*, and *Romance* (Figure 4.4b). A slight drop is observed in *Short* and *Documentary* classes when text is missing, likely due to insufficient information in the image modality alone for those classes. These results demonstrate that CMPTs enhance performance across most classes, leading to a strong overall performance improvement. A similar trend is observed on the UPMC Food-101 dataset which is discussed in Section A2.7 of the appendix.

Determining the Optimal Alignment Loss Weight λ

To determine the optimal value for the alignment loss weight λ , we performed an ablation study on the UPMC Food-101 and AVE datasets, varying λ from 0.0 to 0.5. The results, shown in Table 4.5, indicate that the choice of λ is crucial for performance in different scenarios. For our experiments, $\lambda = 0.20$ provides an overall better performance compared to other values. On the UPMC Food-101 dataset, setting $\lambda = 0.20$ consistently achieves the best performance across

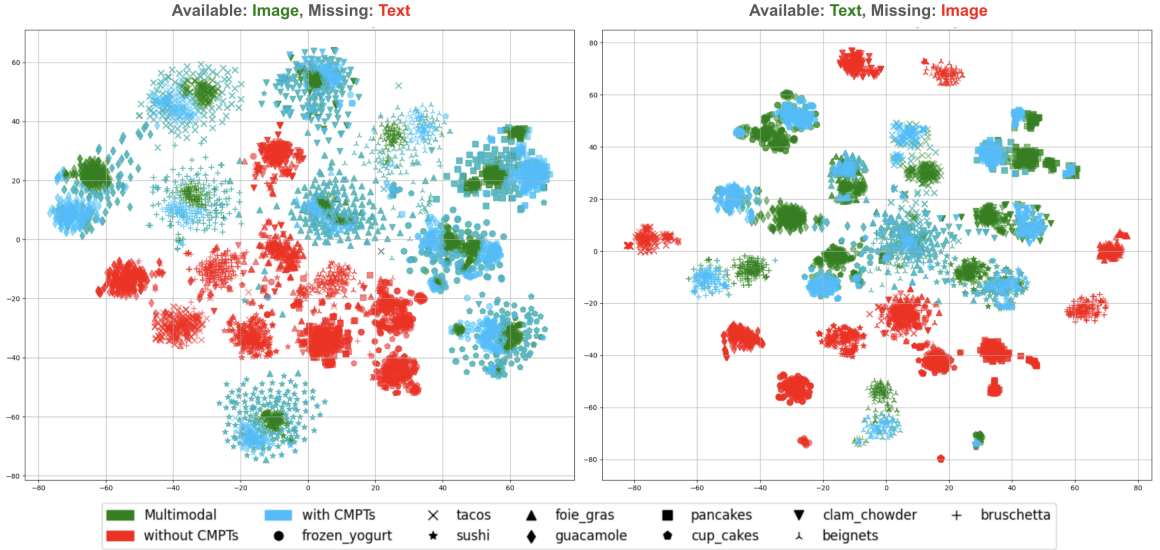


Figure 4.5: t-SNE plots of the fused feature tokens $\mathcal{T}$ for ten classes under different modality settings. The t-SNE plots show that fused features without CMPTs (red) deviate significantly from the complete multimodal features (green). Incorporating CMPTs (blue) provides embeddings that closely align with the complete multimodal features, indicating improved semantic alignment and robustness in different missing modality scenarios.

all conditions: with missing text, missing image, and with both modalities available. Using $\lambda = 0.0$, which is equivalent to removing the alignment loss, results in a significant performance drop, highlighting the importance of the alignment objective.

On the AVE dataset, the results show a similar trend. A weight of $\lambda = 0.20$ yields the best performance when video is missing and when both modalities are available. $\lambda = 0.30$ shows a slight performance improvement in the video-only scenario. Given that $\lambda = 0.20$ demonstrates superior or near-optimal performance consistently across both datasets, we set $\lambda = 0.20$ for all our experiments, as it provides a balance in performance across different modality conditions.

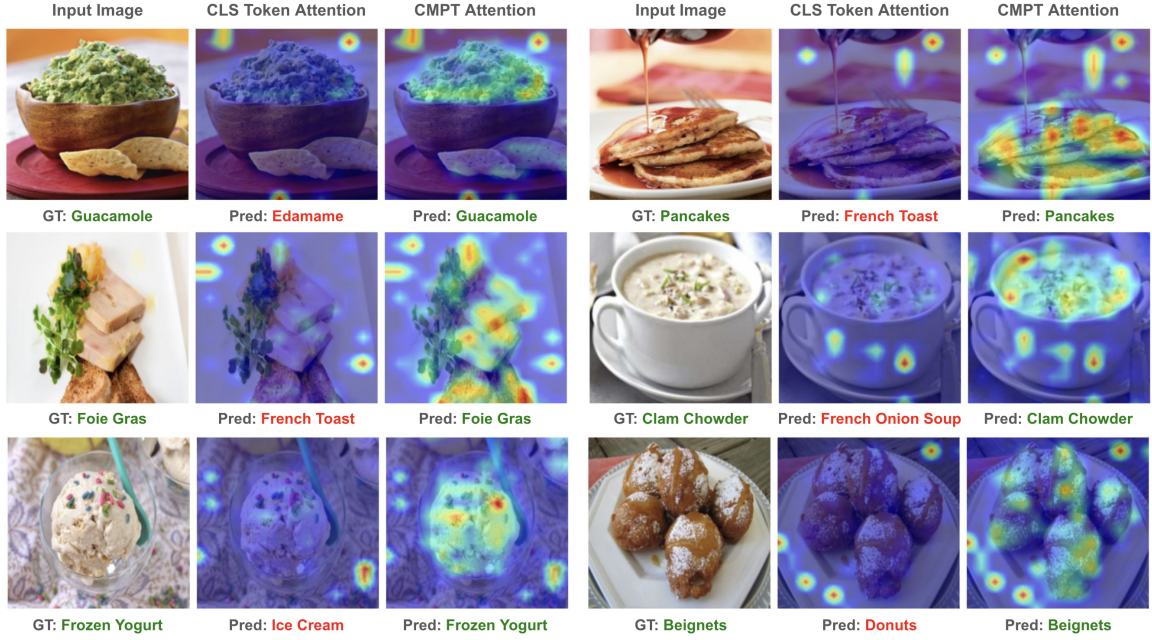


Figure 4.6: Attention map visualizations from the last layer of the vision encoder for the CLS token and CMPTs in the text-missing scenario on the UPMC Food-101 dataset. The attention maps show that CLS tokens attend to semantically-irrelevant regions, which likely leads to incorrect predictions. In contrast, CMPTs attend to semantically-relevant regions for the respective classes and lead to accurate predictions.

Qualitative Analysis of CMPT Effectiveness

To gain deeper insights into how CMPTs enhance robustness in missing modality scenarios, we conducted a qualitative analysis on the UPMC Food-101 dataset using both fused feature projections and attention visualizations. Figure 4.5 presents t-SNE plots of the fused feature $\mathcal{T}$ for ten representative food classes where CMPTs yield notable performance gains. We compare three settings: (i) complete multimodal input (green), (ii) missing modality without CMPTs (red), and (iii) missing modality with CMPTs (blue). The t-SNE plots show that fused features without CMPTs (red) deviate significantly from the complete multimodal features (green) in missing modality scenarios. Incorporating CMPTs (blue) provides embeddings that closely align with the complete multimodal features, indicating improved semantic alignment and robustness.

We also analyze the attention maps from the last layer of the vision encoder for the CLS token and CMPTs in the text-missing scenario, as shown in Figure 4.6. We selected random examples from six classes where fused features without CMPTs provide incorrect predictions, but including the CMPTs leads to correct predictions (approximately 15% samples in the dataset show this behavior). The attention maps show that CLS tokens attend to semantically-irrelevant regions, which likely leads to incorrect predictions. In contrast, CMPTs attend to semantically-relevant regions for the respective classes and lead to accurate predictions. These visualizations demonstrate that CMPTs not only act as proxies for the missing modality but also reshape the model’s attention to reflect modality-relevant cues. For the sake of completeness, we have also included attention maps for some cases when the fused features with and without CMPTs provide correct predictions (approximately 60% samples show this pattern) and when the fused features with CMPTs lead to incorrect predictions (approximately 3% samples show this pattern) in Section A2.8.

4.5 Limitations and Future Directions

While our approach demonstrates consistent performance improvements over existing methods in different missing modality scenarios, the remaining performance drop—especially in extreme cases—indicates room for further improvement. Additionally, we observe that certain classes are heavily dependent on a specific modality, which leads to significant performance drop in those classes when that specific modality is missing. A natural question arises as how can we design a better training strategy that encourages more uniform reliance on different modalities while maintaining overall performance in both complete and missing modality scenarios? Future work can also explore developing a general framework for explaining modality contribution and expected

performance drop under different missing modality scenarios. Finally, this paper primarily focused on two-modality settings due to the availability of standard benchmarks and baselines that enable direct and fair comparisons. Extending the CMPT framework is a natural next step, and we are actively investigating this direction as part of our ongoing work.

4.6 Conclusion

We presented a simple yet effective framework for utilizing pre-trained unimodal encoders for robust multimodal learning. Our proposed framework enables training of multimodal systems that offer robust performance under different scenarios of missing modalities during training and inference. To compensate for missing modalities, we propose to learn cross-modal proxy tokens (CMPTs) to approximate the missing modality class tokens from the available modality using an alignment loss during training. Instead of training the unimodal encoders from scratch, we only learn rank-1 adapters along with task-specific classifier heads. Our extensive evaluations across diverse tasks and datasets demonstrate the robustness and adaptability of our approach in both complete and missing modality scenarios. Our analysis also reveals that CMPTs can generalize well to different missing scenarios and enhance performance across most of the classes to provide an overall performance boost. Overall, our approach provides a robust and efficient solution for multimodal learning that maintains high performance in different real-world scenarios.

Chapter 5

SSAM: Singular Subspace Alignment for Merging Multimodal Large Language Models

5.1 Introduction

Multimodal Large Language Models (MLLMs) [202, 191] have demonstrated remarkable capabilities across diverse domains by integrating vision, audio, and language understanding within a unified reasoning framework. Building on the success of large language models (LLMs) [225, 132], early multimodal systems aligned different input modalities to a shared language representation space using modality-specific encoders paired with pretrained LLM decoders. This paradigm has proven effective for pairs of modalities such as vision-language [1, 110, 7], audio-language [145, 49], video-language [108, 183], and point cloud-language [198, 165]. By leveraging

large-scale paired multimodal datasets, these specialized MLLMs achieve strong alignment, generalization, and performance across a wide range of multimodal tasks and modality combinations.

Current efforts in this area are geared toward constructing MLLMs that can process multiple modalities simultaneously. One approach incorporates modality-specific [192, 226] or shared [57] encoders with a common language decoder and fine-tunes the model on large-scale paired multimodal datasets, as illustrated in Figure 5.1(a). Training such models or scaling existing MLLMs (e.g., vision–language or audio–language) to new modalities **requires massive paired multimodal pretraining and instruction-**

tuning datasets, along with significant computational resources. Alternative methods [58, 163] partially relax data requirements by aligning all modalities to an intermediate image embedding space before connecting to the LLM, as shown in Figure 5.1(b). While this reduces the need for all-modality paired data, it still depends on high-quality image and other modality pairs for alignment. These approaches face two major limitations: first, collecting and synchronizing heterogeneous sensory data (e.g., audio–video–text–point cloud) is costly and often impractical; and second, retraining or extending MLLMs for each new modality remains compute-intensive and laborious.

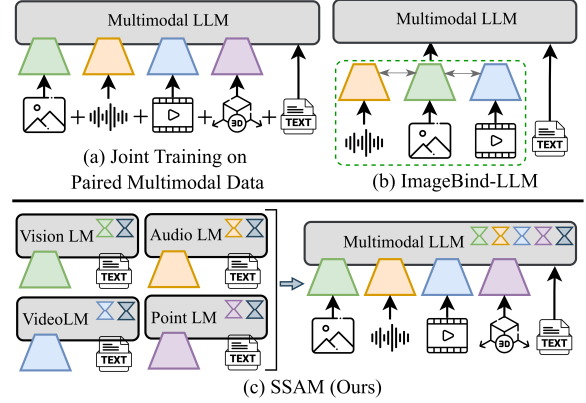


Figure 5.1: Comparison of multimodal learning paradigms. (a) Conventional methods train a single MLLM on paired multimodal datasets. (b) ImageBind-LLM [58] aligns all modalities to shared image embeddings. (c) SSAM offers a training-free method to merge independently trained specialist MLLMs into a single, unified model via subspace projection.

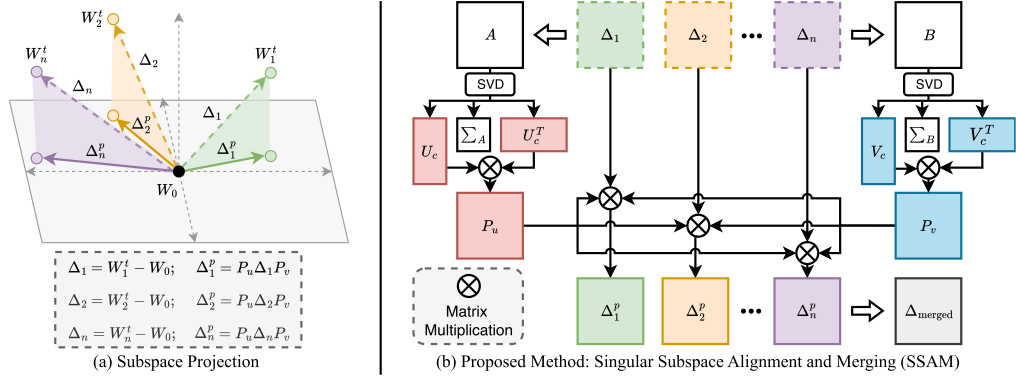


Figure 5.2: Overview of the proposed SSAM method. (a) Subspace projection: each language vector Δ_i is projected onto a shared low-rank subspace prior to merging. (b) SSAM constructs covariance matrices A and B from all Δ_i , extracts the shared left and right bases U_c and V_c , projects language vectors, and merge them into Δ_{merged} .

In this paper, we focus on the challenging problem of combining existing pretrained MLLMs with different input modalities (e.g., vision–language, audio–language) into a single model capable of processing all modalities simultaneously **without** training on any paired multimodal data, as illustrated in Figure 5.1(c). Our work is inspired by recent progress in training-free model merging, which demonstrate that knowledge from multiple expert models can be integrated without access to training data. Early studies [78, 43, 22, 199] revealed that fine-tuned models encode task-specific knowledge as directional “task vectors” in parameter space, enabling vector operations to combine model capabilities. *These methods primarily target homogeneous settings (e.g., multiple vision–language, vision only, or language only models).* Recent works attempt to bridge this gap via parameter decoupling [14] or structured optimization [182] to reduce task interference.

To address these challenges, we propose **Singular Subspace Alignment and Merging (SSAM)**, a training-free framework that merges independently trained specialist MLLMs into a single model *capable of processing arbitrary combinations of input modalities*. SSAM is motivated by the observation that task vectors of specialist models are inherently low rank [43] and often

span partially overlapping subspaces. Because we consider heterogeneous input modalities, SSAM maintains modality-specific parameter updates separately to reduce modality interference, while computing language-related parameter updates, referred to as *language vectors*, for each specialist model prior to alignment. Unlike prior methods that rely on direct weight interpolation [182, 14], SSAM explicitly identifies a *shared low-rank consensus subspace* that captures consistent parameter update directions across all the specialist MLLMs. Each language vector is projected onto this subspace before merging, ensuring alignment along dominant low-rank directions while filtering out noisy or conflicting updates as shown in Figure 5.2. This subspace-based alignment mitigates interference in shared language parameters and preserves complementary knowledge from the specialist MLLMs, enabling coherent multimodal reasoning without retraining or access to paired training data (Figure 5.3). **Our main contributions are summarized as follows:**

- We propose **Singular Subspace Alignment and Merging (SSAM)**, a training-free framework that aligns and merges specialist MLLMs by projecting their language vectors onto a shared low-rank subspace, effectively reducing parameter interference while preserving complementary knowledge.
- SSAM achieves state-of-the-art performance across four multimodal datasets, consistently outperforming both jointly trained multimodal models and existing training-free model merging approaches (Section 5.4.3, and Section 5.4.4).
- We further evaluate SSAM on three domain-specific benchmarks, showing that it preserves specialist capabilities and often improves performance. Extensive ablation studies analyze the impact of subspace rank and merging strategies, demonstrating that subspace alignment enhances stability and cross-modal generalization (Section 5.4.5).

- We present qualitative results demonstrating that the merged model effectively identifies shared semantic cues across image, audio, video, and point cloud modalities, validating SSAM’s coherent multimodal reasoning (Figure 5.3 and Section A3.5).

5.2 Related Work

Multimodal Large Language Models (MLLMs). Building on the success of large language models (LLMs) [225, 132], MLLMs [202, 191] extend pretrained LLMs to process diverse input modalities such as image, audio, and point cloud by aligning them to the language embedding space. Vision–language models such as Flamingo [1], LLaVA [110], and Qwen-VL [7] connect vision encoders to LLM decoders via learned projectors, while audio–language models like Whisper [145] and LTU [49] apply similar strategies for audio understanding tasks. This paradigm has since been extended to additional modalities including video [126, 212], sensor data [103, 79], and point clouds [198, 165]. Recent efforts aim to develop unified MLLMs capable of processing multiple modalities simultaneously. OneLLM [57] employs a shared encoder across all modalities, whereas NExT-GPT [192] uses modality-specific encoders with projector layers. These models are typically fine-tuned on large-scale paired multimodal datasets for downstream reasoning tasks. Alignment-based frameworks such as ImageBind-LLM [58] and PandaGPT [163] bind all modalities to a common image embedding space following ImageBind [45] approach before aligning to the language space, reducing but not eliminating the need for paired multimodal data. Despite their success, these methods face two key limitations: collecting large-scale aligned multimodal data is difficult, costly and time consuming due to modality heterogeneity, and training such models requires substantial compute resources.

Model Merging. Model merging has emerged as an efficient approach for combining complementary capabilities from multiple expert models. Task Arithmetic [78] introduced the concept of a *task vector*, representing a directional update in weight space, and showed that arithmetic operations on task vectors can steer model behavior. Subsequent works extend this idea through both training-free and unsupervised test-time adaptation strategies. Training-free approaches merge models without additional training data: Model Soup [188] averages model weights, while Fisher Merging [129] applies Fisher-weighted averaging. However, direct averaging often introduces task interference and degrades model performance [43]. To mitigate this, TIES-Merging [199] prunes and aligns weight signs before merging, DARE [206] randomly drops and rescales parameters, Consensus Merging [173] removes conflicting updates, WUDI [22] explicitly minimizes interference via gradient descent, and Core Space Merging [140] projects LoRA-based [72] parameter updates into a compact latent space before merging. Recent studies further explore subspace geometry and orthogonality to improve merging robustness: DO-Merging [229] decouples magnitude and direction components, OSRM [213] enforces orthogonal subspaces to reduce interference, KnOTS [162] aligns LoRA weights via joint SVD transformation, VisionFuse [19] combine multiple vision encoders, and Isotropic-Merging [128] flattens task spectra to enhance alignment. Test-time adaptation methods such as AdaMMS [36] and Representation Surgery [200] further refine merged models using unlabeled test data. Although these methods show strong results, they are primarily designed for unimodal (vision or language) or homogeneous multimodal settings (only vision–language models) and often require test data for further model finetuning, limiting their applicability to truly multimodal model merging.

Merging MLLMs with Heterogeneous Modalities. Recent studies have begun exploring model merging for MLLMs with heterogeneous input modalities (e.g., vision–language and audio–language). DAMC [14] mitigates task interference by processing each modality through separate modality-specific adapters. However, it merges the shared parameters via weighted averaging, which can still introduce interference due to conflicting updates from different source models. OptMerge [182] optimizes task vectors in low-rank space to reduce noise and interference, but direct weight interpolation still struggles to capture complementary information across distinct modalities, often leading to degraded performance.

In contrast, SSAM adopts a training-free approach that preserves modality-specific parameter updates while identifying a shared low-rank subspace for language-specific updates. By capturing the dominant and mutually consistent representational directions across the constituent specialist MLLMs, this subspace alignment strategy mitigates parameter interference during merging. SSAM consistently outperforms existing merging methods (Sections 5.4.3 and 5.4.4) and constituent specialist models (Section 5.4.5) across modality combinations, effectively capturing cross-modal semantics (Figure 5.3) without additional fine-tuning.

5.3 Proposed Method

We aim to construct a single unified MLLM, M_{merged} , capable of processing arbitrary combinations of input modalities by merging a collection of independently fine-tuned *specialist* models $\{M_1, M_2, \dots, M_n\}$, each trained on different modality–language pairs (e.g., vision–language, audio–language). Each M_i consists of two primary components: (1) a modality-specific encoder E_i that map raw modality inputs (e.g., image, audio, video, or point cloud) into the shared language

embedding space, and (2) a language decoder W_i that jointly processes the multimodal embeddings and textual queries to generate responses.

Our objective is to derive M_{merged} that inherits the multimodal reasoning capabilities of all $\{M_i\}_{i=1}^n$. *Moreover, we also seek to enable the model to understand arbitrary combinations of input modalities without requiring additional fine-tuning on large-scale paired multimodal data.* For instance, by merging a vision-language model and an audio-language model, the resulting merged model M_{merged} should not only retain the ability to process vision-language and audio-language inputs independently but also gain the capability to reason over combined vision-audio-language inputs. To achieve this, we propose a training-free merging framework: **Singular Subspace Alignment and Merging (SSAM)**. The key motivation behind SSAM is to preserve modality-specific parameter updates in isolation while identifying a shared, low-rank subspace for the language-specific parameter updates. This subspace is designed to capture the dominant and mutually consistent representational directions across all $\{M_i\}_{i=1}^n$, thereby mitigating interference during the merging process. A pseudocode for this method is shown in Algorithm 1 in Section A3.3.

5.3.1 Problem Formulation

We assume all $\{M_i\}_{i=1}^n$ share the same language decoder architecture that is fine-tuned from a common pretrained language decoder with weights W_0 . W_i denotes the language decoder weights of the i -th fine-tuned specialist ($i \in \{1, 2, \dots, n\}$) each fine-tuned to process a separate modality $m \in \{m_1, m_2, \dots, m_n\}$ paired with language t . We follow a parameter decoupling strategy similar to [14] and decouple modality-specific and language-specific parameters as W_i^m and W_i^t , respectively. An MLLM that is fine-tuned on modality m paired with language t , we can write the input to any layer l of the language decoder as $h_l = [h_l^m, h_l^t]$ where h_l^m and h_l^t denote the

modality-specific and language-specific hidden states, respectively. For simplicity of notation, we omit the layer subscript l in the subsequent derivations.

We calculate the query Q , key K and value V matrices for any layer as:

$$Q = [h^m W_{i,Q}^m, h^t W_{i,Q}^t]; \quad K = [h^m W_{i,K}^m, h^t W_{i,K}^t]; \quad V = [h^m W_{i,V}^m, h^t W_{i,V}^t]; \quad (5.1)$$

where $W_{i,(.)}^m$ and $W_{i,(.)}^t$ denote modality-specific and language-specific weights. Attention is applied to the whole sequence and the output is calculated as:

$$[h_a^m, h_a^t] = \text{Split}(\text{Attention}(Q, K, V) = \text{Split}(\text{softmax}(\frac{QK^\top}{\sqrt{d_k}})V) \quad (5.2)$$

$$[h_o^m, h_o^t] = [h_a^m W_{i,O}^m, h_a^t W_{i,O}^t]; \quad h_{l+1} = [\text{FFN}_m(h_o^m), \text{FFN}_t(h_o^t)] \quad (5.3)$$

where d_k is the dimension of K and FFN_m and FFN_t denote modality-specific and language-specific feed-forward layers. During attention, hidden states from all modalities attend to one another through a shared attention map, enabling cross-modal interactions in the joint sequence space. This design allows modality and language tokens to exchange information at every layer while preserving modality-specific parameters, thereby retaining modality-specific representations alongside shared multimodal reasoning. For our merging method, we keep all the modality-specific parameter updates $W_{i,(.)}^m$ separate and merge only language-specific weights $W_{i,(.)}^t$. For simplicity, we denote them as W_i^t .

5.3.2 Language Vectors

Following prior work [78, 22], we define the language-specific parameter updates, referred to as *language vectors*, as

$$\Delta_i = W_i^t - W_0, \quad (5.4)$$

where $\Delta_i \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ represents the parameter shift that adapts the general-purpose language decoder W_0 to a specific multimodal task (e.g., vision–language). We compute these language vectors independently for each linear layer l of the model. For models adapted using Low-Rank Adaptation (LoRA) [72], the language vector naturally takes the form $\Delta_i = A_i^t B_i^t$, where $A_i^t \in \mathbb{R}^{d_{\text{out}} \times r}$ and $B_i^t \in \mathbb{R}^{r \times d_{\text{in}}}$ are low-rank matrices.

Our goal is to estimate a merged language vector Δ_{merged} such that the resulting language decoder retains and integrates the capabilities of all $\{\Delta_i\}_{i=1}^n$. Since we only merge the language-specific parameter updates, we hypothesize that the subspaces spanned by the language vectors exhibit significant overlap, making it possible to identify a shared common subspace. Naïvely averaging language vectors in the full parameter space ($\frac{1}{n} \sum_{i=1}^n \Delta_i$) or in the LoRA space ($\frac{1}{n} \sum_{i=1}^n A_i^t$ and $\frac{1}{n} \sum_{i=1}^n B_i^t$) often leads to suboptimal performance due to parameter interference (Section 5.4.5) because conflicting parameter updates in Δ_i , A_i^t , or B_i^t can cancel out [199, 22]. SSAM addresses this by identifying a *shared consensus subspace* that captures consistent update directions across all $\{\Delta_i\}_{i=1}^n$ and merging the projected language vectors within that subspace.

5.3.3 Shared Consensus Subspace Construction

To identify a shared subspace that best captures the consistent parameter updates across all $\{\Delta_i\}_{i=1}^n$, we analyze their principal directions of variance. Each Δ_i represents a directional update in the model’s parameter space, reflecting the adaptation of W_0 toward a specific modality pair. Although these updates differ across $\{\Delta_i\}_{i=1}^n$, they often exhibit correlated directions because they all modify language-related parameters. Our objective is to extract this common structure while suppressing conflicting updates that may introduce noise or interference.

We first compute the row- and column-wise inner-product matrices, A and B , by aggregating the contributions of all $\{\Delta_i\}_{i=1}^n$:

$$A = \sum_{i=1}^n \Delta_i \Delta_i^\top, \quad B = \sum_{i=1}^n \Delta_i^\top \Delta_i. \quad (5.5)$$

These matrices capture the dominant column- and row-space structures shared among the language vectors. We then perform Singular Value Decomposition (SVD) as $A = U \Sigma_A U^\top$, $B = V \Sigma_B V^\top$, where U and V contain the orthonormal bases corresponding to the principal directions of variation across all $\{\Delta_i\}_{i=1}^n$. The largest singular values in Σ_A and Σ_B correspond to the most consistent and energy-preserving update directions shared among the specialists models. By selecting the top- k eigenvectors in U and V , we define the orthonormal bases of the **shared consensus subspace**:

$$U_c = U_{[:,1:k]}, \quad V_c = V_{[:,1:k]}. \quad (5.6)$$

The rank k controls the effective dimensionality of the shared subspace. In practice, k is chosen to retain the majority of cumulative spectral energy while removing task-specific noise [43, 182]. We evaluate $k \in \{64, 128, 256, 384, 512\}$ and find $k = 128$ empirically provides the best trade-off between shared variance preservation and parameter interference suppression as discussed in Section 5.4.5.

Justification for shared subspaces. One might intuitively consider computing the shared subspace by directly applying Truncated SVD to the naïvely summed language vectors $\bar{\Delta} = \sum_{i=1}^n \Delta_i$. However, the left singular vectors of this direct sum are derived from the covariance matrix:

$$\bar{\Delta} \bar{\Delta}^\top = \sum_{i=1}^n \Delta_i \Delta_i^\top + \sum_{i \neq j} \Delta_i \Delta_j^\top. \quad (5.7)$$

In the context of MLLMs independently fine-tuned on heterogeneous modalities, the cross-covariance terms ($\sum_{i \neq j} \Delta_i \Delta_j^\top$) often contain parameter interference and spurious cross-task noise. By explic-

itly defining $A = \sum_{i=1}^n \Delta_i \Delta_i^\top$ and $B = \sum_{i=1}^n \Delta_i^\top \Delta_i$ in Equation (5.5), SSAM deliberately drops these noisy cross-terms. Consequently, our bases U_c and V_c minimize the joint column- and row-space projection errors across all $\{\Delta_i\}_{i=1}^n$:

$$U_c = \arg \min_{U^\top U = I_k} \sum_{i=1}^n \|\Delta_i - U U^\top \Delta_i\|_F^2, \quad (5.8)$$

$$V_c = \arg \min_{V^\top V = I_k} \sum_{i=1}^n \|\Delta_i - \Delta_i V V^\top\|_F^2. \quad (5.9)$$

Next, we define the projection operators for the left and right subspaces and project each Δ_i onto the shared consensus subspace as:

$$P_u = U_c U_c^\top \in \mathbb{R}^{d_{\text{out}} \times d_{\text{out}}}, \quad P_v = V_c V_c^\top \in \mathbb{R}^{d_{\text{in}} \times d_{\text{in}}} \quad (5.10)$$

$$\Delta_i^p = P_u \Delta_i P_v \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}. \quad (5.11)$$

This operation aligns the language vectors onto a common coordinate system spanned by the shared bases U_c and V_c , ensuring that only the most salient update directions are preserved. By projecting $\{\Delta_i\}_{i=1}^n$ onto this subspace, SSAM captures the shared representational geometry across $\{M_i\}_{i=1}^n$ while suppressing conflicting updates and reducing parameter interference during merging.

5.3.4 Language Vector Merging

After projection, we compute the merged language vector as $\Delta_{\text{merged}} = \lambda \sum_{i=1}^n \Delta_i^p$ where λ is a scaling coefficient. For simplicity, we can set $\lambda = \frac{1}{n}$ when the norms of $\{\Delta_i\}_{i=1}^n$ are similar. However, since each M_i is fine-tuned independently on different modality pairs, scaling the merged language vector can improve performance on a given downstream task. In practice, λ can be determined using a small validation set for the target task or selected based on the performance of the merged model on individual modality-language pairs [14, 78].

For models fine-tuned using LoRA, we can represent the merged update as $\Delta_{\text{merged}} \equiv A_{\text{merged}}^t B_{\text{merged}}^t$, where $A_{\text{merged}}^t = U_c$ and $B_{\text{merged}}^t = U_c^\top \lambda \sum_{i=1}^n \Delta_i P_v$. This step preserves LoRA’s parameter efficiency and enables direct substitution into pre-trained model checkpoints. Bias and normalization parameters are merged via simple averaging.

In summary, SSAM constructs a shared consensus subspace that captures mutually consistent language-specific parameter updates across models, projects them onto this subspace, merges them, and re-factorizes the result into low-rank LoRA form when applicable. The resulting language decoder effectively integrates knowledge from all specialist MLLMs and enables robust multimodal reasoning without requiring any paired training data or additional fine-tuning.

5.4 Experiments and Results

5.4.1 Datasets and Benchmarks

We evaluate SSAM on four multimodal datasets from two representative benchmarks. In addition, we assess performance on three domain-specific benchmarks to verify that merging preserves the capabilities of the specialist models. A brief overview is provided below, with full dataset details in Section A3.1. We denote video, image, audio, point cloud, and language as V, I, A, P, and L, respectively.

Audio-Visual Question Answering (AVQA) Benchmarks. The AVQA task evaluates a model’s ability to answer questions on image, audio, and video inputs. We assess SSAM on two large-scale datasets: MUSIC-AVQA [96] and AVQA [201], each comprising three modalities (I, A, and V) paired with language. **MUSIC-AVQA** focuses on fine-grained multimodal understanding and

spatio-temporal reasoning in musical performance. **AVQA** targets complex interactions and relationships between real-world objects, sounds, and daily activities.

Multimodal Commonality Understanding Benchmark (MCUB). The MCUB benchmark [14] measures a model’s ability to identify shared semantic concepts across diverse input modalities. It includes two variants: **MCUB-3**, which includes four three-modality combinations (I+V+P, A+V+P, I+A+P, and I+A+V) paired with language, and **MCUB-4**, which integrates all four modalities (I+A+V+P) paired with language.

Domain-Specific Benchmarks. To assess whether the merged model preserves the capabilities of the original specialist models, we evaluate it on three domain-specific benchmarks: MMLU [68] for language understanding, OCRBench [116] for vision–language reasoning, and MMAU [153] for audio–language tasks. These evaluations demonstrate that merging preserves and, in some cases, improves the performance of the model.

5.4.2 Implementation Details

We consider four input modalities: Image (I), Audio (A), Video (V), and Point Cloud (P) paired with Language (L). For all experiments, we use the pretrained specialist model checkpoints provided by [14], which include vision–language, audio–language, video–language, and point–language models. Each specialist uses Vicuna-7B-v1.5 [228] as language decoder, CLIP-ViT-L-336px [144] as image encoder, BEATs-Iter3+ [17] as audio encoder, LanguageBind [238] video encoder, and PointLLM [198] point encoder. The audio encoder employs a Q-Former [98] with 32 query tokens as the projector, while other modalities use MLP projectors.

Table 5.1: Performance comparison on MUSIC-AVQA [96] and AVQA [201] datasets. SSAM achieves the best average performance across all modality combinations and consistently outperforms both training-free model merging baselines and models finetuned on paired multimodal data.

Dataset	Method	V+L	V+I+L	V+A+L	V+I+A+L	Average
MUSIC-AVQA	Models finetuned on paired multimodal data					
	ChatBridge-13B [226]	-	-	43.00	-	43.00
	OneLLM-7B [57]	-	-	47.60	-	47.60
	ImageBind-LLM [58]	37.24	38.76	39.72	38.16	38.47
	X-InstructBLIP [139]	45.83	41.23	48.34	47.39	45.70
	Proj-Only [14]	44.93	46.64	46.17	50.21	46.99
	Training-free model merging methods					
	Task Arithmetic [78]	53.53	53.22	53.62	55.98	54.09
	TSV [43]	53.81	54.26	53.63	54.73	54.11
	WUDI [22]	53.89	53.78	54.10	52.43	53.55
	OptMerge [182]	51.83	52.17	51.05	53.17	52.06
	NaiveMC [14]	49.00	52.52	50.66	53.63	51.45
	DAMC [14]	49.09	53.08	50.91	57.32	52.60
	SSAM (Ours)	54.34	54.55	54.68	56.30	54.97
AVQA	Models finetuned on paired multimodal data					
	ImageBind-LLM [58]	51.77	51.65	55.00	54.26	53.17
	X-InstructBLIP [139]	41.91	40.42	44.29	44.23	42.71
	Proj-Only [14]	67.99	66.65	67.65	66.85	67.29
	Training-free model merging methods					
	Task Arithmetic [78]	61.96	73.60	77.93	77.29	72.70
	TSV [43]	77.88	80.28	81.91	81.28	80.34
	WUDI [22]	64.30	76.09	79.06	78.69	74.54
	OptMerge [182]	78.37	80.26	82.10	81.52	80.56
	NaiveMC [14]	79.37	79.74	79.82	80.70	79.91
	DAMC [14]	79.15	80.30	80.40	81.31	80.29
	SSAM (Ours)	80.45	80.70	82.12	81.87	81.29

During merging, we experiment with ranks $k \in \{64, 128, 256, 384, 512\}$ and find that $k = 128$ yields overall better performance (Section 5.4.5). We report results from published works whenever available. For training-free model merging methods [78, 43, 22, 182], we reproduce results using the same checkpoints with the public codebase of [182]. Additional implementation details are provided in supplementary Section A3.2.

5.4.3 Results on AVQA Benchmarks

We summarize the performance of SSAM on the two AVQA datasets in Table 5.1. We compare SSAM with two categories of methods: (1) jointly trained multimodal models that rely on large paired multimodal datasets, including ChatBridge-13B [226], OneLLM [57], ImageBind-LLM [58], X-InstructBLIP [139], and Proj-Only [14]; and (2) training-free model merging approaches such as Task Arithmetic [78], TSV [43], WUDI [22], OptMerge [182], NaiveMC [14], and DAMC [14].

Performance on MUSIC-AVQA dataset

For MUSIC-AVQA, training-free merging methods generally outperform models trained on paired multimodal data, highlighting the effectiveness of model merging without retraining. SSAM achieves the highest average accuracy of 54.97%, surpassing the best finetuned baseline, Proj-Only (46.99%), by 7.98 percentage points. Within the training-free merging category, SSAM attains the best average accuracy and maintains consistent performance across all modality combinations. It achieves the top scores in three subsets: V+L (54.34%), V+I+L (54.55%), and V+A+L (54.68%). In V+I+A+L setting, DAMC slightly exceeds SSAM (57.32% vs. 56.30%), but SSAM

Table 5.2: Performance comparison on MCUB benchmark [14]. SSAM achieves the best average performance across all modality combinations and consistently outperforms both training-free merging baselines and models finetuned on paired multimodal data.

Method	MCUB-3					MCUB-4
	I+V+P+L	A+V+P+L	I+A+P+L	I+A+V+L	Average	I+A+V+P+L
Models finetuned on paired multimodal data						
ImageBind-LLM [58]	31.80	31.40	33.40	35.20	32.95	32.93
X-InstructBLIP [139]	29.40	25.20	21.20	41.40	29.30	27.94
Proj-Only [14]	42.60	43.80	42.80	47.40	44.15	43.00
Training-free model merging methods						
Task Arithmetic [78]	36.80	37.60	43.80	45.00	40.80	40.53
TSV [43]	48.00	48.00	58.00	51.60	51.40	52.29
WUDI [22]	38.20	40.40	44.40	44.40	41.85	40.95
OptMerge [182]	52.80	53.60	58.60	53.80	54.70	55.88
NaiveMC [14]	53.00	51.00	58.80	56.00	54.70	54.03
DAMC [14]	58.20	58.80	65.60	56.60	59.80	60.08
SSAM (Ours)	59.40	61.00	65.40	55.60	60.35	62.27

achieves the best overall mean accuracy across all combinations. These results indicate that SSAM effectively integrates complementary information from different specialist models while mitigating interference between modalities.

Performance on AVQA dataset

On the AVQA dataset, SSAM achieves an average accuracy of 81.29%, outperforming both finetuned and training-free baselines. Compared to the best finetuned model, Proj-Only (67.29%), SSAM improves by 14 percentage points. Among training-free merging methods, SSAM consistently achieves the highest accuracy across all modality combinations.

5.4.4 Results on MCUB Benchmarks

The results for MCUB-3 and MCUB-4 datasets are summarized in Table 5.2. We compare SSAM with (1) models finetuned on paired multimodal data [58, 139, 14], and (2) existing training-free model merging methods [78, 43, 22, 182, 14].

Performance on MCUB-3 dataset

Compared to models trained on paired multimodal data, training-free merging methods show substantial improvements, underscoring their potential for scalable and efficient merging of MLLMs. SSAM achieves the best overall performance with an average accuracy of 60.35%, surpassing the best finetuned baseline, Proj-Only (44.15%), by 16.20 percentage points. Within the training-free category, SSAM attains the highest average accuracy, outperforming the next-best method, DAMC (59.80%). A closer analysis reveals complementary strengths: SSAM leads in the I+V+P+L (59.40%) and A+V+P+L (61.00%) subsets, while DAMC performs slightly better in I+A+P+L (65.60%) and I+A+V+L (56.60%). Despite these minor variations, SSAM maintains strong overall performance across all subsets, demonstrating consistent generalization across diverse modality combinations.

Performance on MCUB-4 dataset

MCUB-4 presents a more challenging scenario, requiring reasoning over all four modalities (I, A, V and P) simultaneously. In this setting, SSAM achieves an accuracy of 62.03%, outperforming all training-free baselines, including DAMC (60.08%), and surpassing the best jointly trained model, Proj-Only (43.00%), by 19.03 percentage points. Figure 5.3 presents qualitative

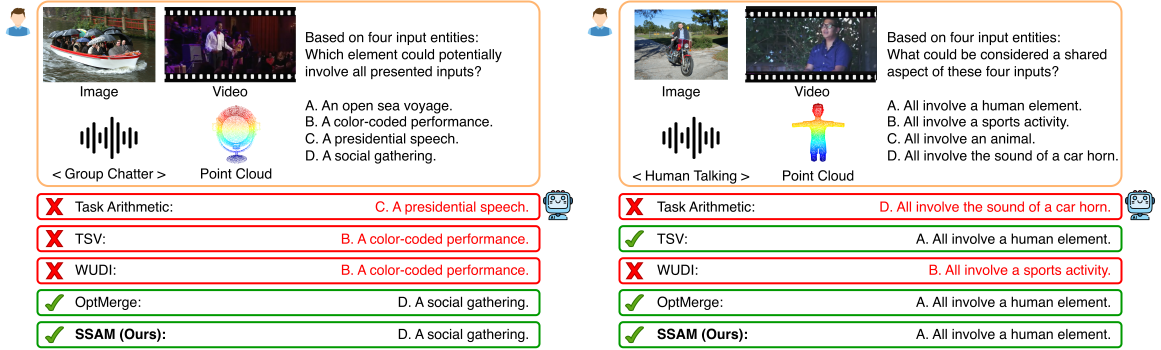


Figure 5.3: Qualitative examples on commonality understanding on the MCUB-4 dataset.

examples from the MCUB-4 dataset, showing that the merged model effectively identifies shared semantic cues across image, audio, video, and point cloud modalities, although none of the specialist models was fine-tuned on such combinations. Additional examples are provided in supplementary Section A3.5. These results highlight SSAM’s ability to integrate complementary modality information and maintain coherent multimodal reasoning across unseen modality combinations.

5.4.5 Ablation Studies

Model Merging Enhances Cross-Modal Performance. We compare the merged model with its constituent specialist vision–language, audio–language, and video–language models across different modality–language tasks and summarize the results in Table 5.3. The merged model consistently outperforms individual specialists on both MUSIC-AVQA and AVQA datasets, improving average performance by +8.36% and +9.46%, respectively. Notably, it generalizes effectively to the unseen four-modality (V+I+A+L) setting, despite none of the specialists being trained on this combination. We also provide qualitative examples in Figure 5.4, demonstrating that the merged model yields more accurate predictions than its specialist counterparts. Further examples are available in supple-

Table 5.3: Comparison of specialist vs. merged MLLMs on the MUSIC-AVQA and AVQA datasets. The merged model consistently outperforms individual specialists across modality pairs and generalizes well to the unseen four-modality (V+I+A+L) setting.

Dataset	Model	V+L	I+L	A+L	V+I+A+L	Avg.
MUSIC-	Specialist	48.88	51.22	33.25	–	44.45
AVQA	Merged	54.34	53.46	47.15	56.30	52.81
AVQA	Specialist	79.09	75.52	44.18	–	66.26
	Merged	80.45	76.33	64.21	81.87	75.72

Table 5.4: Comparison of specialist vs. merged MLLMs across three domain-specific benchmarks. VLM, Vid-LM, and ALM denote Vision-Language, Video-Language, and Audio-Language models, respectively. SSAM (ours) consistently outperforms individual specialists.

Benchmarks	Specialist MLLMs			Merged
	VLM	Vid-LM	ALM	SSAM
MMLU	50.52	50.25	42.62	51.82
OCRBench	319	–	–	329
MMAU (test-mini)	–	–	48.50	58.90

mentary Section A3.5. These gains arise from the subspace alignment mechanism, which projects language vectors onto a shared low-rank consensus subspace. By emphasizing consistent representational directions and suppressing noisy or conflicting updates, this alignment enables the model to integrate complementary strengths from all the specialist models, resulting in more robust and accurate multimodal reasoning.

Comparison with Simple Merging Strategies. We compare SSAM against two standard training-free merging baselines: Task Arithmetic [78], which directly adds the language vectors, and NaiveMC [14], which averages them. As shown in Figure 5.5, direct addition or averaging results in sub-optimal performance across all datasets (likely due to conflicting parameter updates and parameter interference) [43, 22, 182]. In contrast, SSAM projects language vectors into a shared low-rank subspace before merging, aligning consistent update directions while filtering modality-specific variations. This projection consistently improves performance across all four datasets, demonstrating that subspace alignment offers a more stable and coherent alternative to simple linear combination methods.

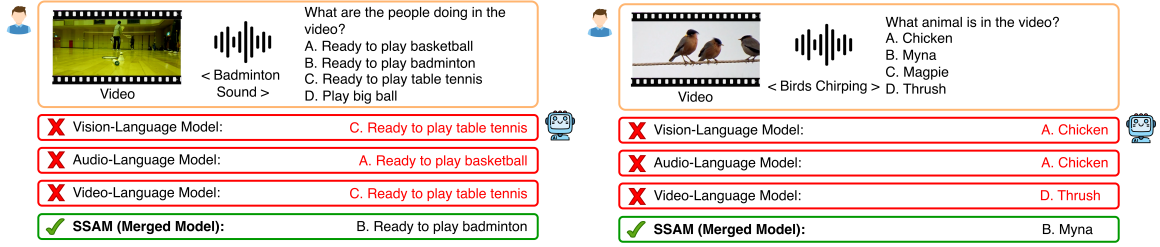


Figure 5.4: Qualitative examples on the AVQA dataset.

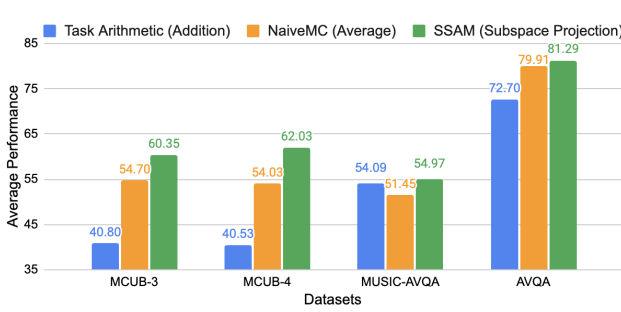


Figure 5.5: Comparison with simple model-merging baselines. SSAM consistently outperforms them across all four datasets.

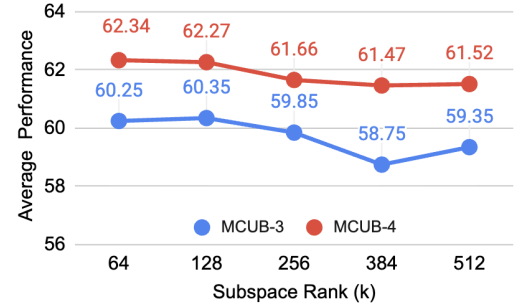


Figure 5.6: Effect of subspace rank (k) on average performance. Empirically $k = 128$ offers slightly better performance.

Merging Preserves Specialist Capabilities. To assess whether the merged model preserves the capabilities of the original specialist models, we evaluate it on three domain-specific benchmarks. The results are summarized Table 5.4. We evaluate SSAM on MMLU [68] benchmark to assess potential degradation of core language capabilities. SSAM slightly outperforms all specialist models indicating that merging does not degrade language capabilities of the original specialist models. To evaluate vision–language and audio–language reasoning, we evaluate SSAM on OCRBench [116] and MMAU (test-mini) [153], respectively. Results show that SSAM does not cause performance degradation on these benchmarks and slightly outperforms specialist models on both of them. These results indicate that SSAM does not degrade performance or suffer from catastrophic forgetting on domain-specific tasks and often provides performance improvements.

Effect of Subspace Rank (k). To determine the optimal subspace dimensionality, we evaluate SSAM with ranks $k \in \{64, 128, 256, 384, 512\}$ on the MCUB-3 and MCUB-4 datasets. As shown in Figure 5.6, ranks 64 and 128 exhibit similar performance, whereas increasing the rank gradually decreases performance. We hypothesize that higher ranks introduce redundant or conflicting updates within the language vectors, leading to parameter interference. Rank $k = 128$ empirically yields overall better performance across both datasets and is adopted in all the experiments. The reported averages are computed over all modality combinations, with detailed breakdowns provided in supplementary Section A3.4.

5.5 Discussion and Future Directions

Discussion. Model merging approaches such as SSAM offer a practical and scalable way to unify specialist MLLMs without retraining or access to paired multimodal data. As the number of fine-tuned models on open-source platforms like Hugging Face [184] continues to grow, model merging provides an efficient means to reuse, combine, and extend these models. It promotes model reuse, reduces data collection and training costs, preserves privacy by avoiding data sharing, and enables distributed development where independently trained specialist models can be integrated into a unified multimodal system.

Limitations and Future Directions. To maintain fairness with existing benchmarks and account for computational constraints, our experiments focused on models with approximately 7B parameters. We focus on merging models with the same architecture; extending to heterogeneous architectures remains an open challenge. We evaluated SSAM on four datasets that include diverse combinations of image, audio, video, point cloud, and language modalities. However, there is currently

no comprehensive benchmark specifically designed to evaluate *MLLM merging*. Developing such benchmarks would enable broader and more systematic evaluation across additional modalities, such as depth, thermal, IMU, or medical imaging, which are essential for assessing generalization to heterogeneous and complex modality interactions. Finally, the safety and alignment of merged MLLMs remain open challenges. Understanding how merging affects factual consistency, ethical behavior, and robustness is important for ensuring reliable and trustworthy multimodal systems.

5.6 Conclusion

We propose Singular Subspace Alignment and Merging (SSAM), a training-free framework for merging independently trained MLLMs. Unlike conventional joint multimodal training, which requires large paired datasets and substantial computational resources, SSAM integrates specialist MLLMs into a single model capable of processing arbitrary combinations of input modalities. SSAM preserves modality-specific parameter updates while identifying a shared low-rank subspace that captures dominant and mutually consistent directions across language-specific parameters. By aligning and merging updates within this subspace, SSAM constructs a unified MLLM that retains modality-specific representations while capturing shared knowledge in a common language space. This design mitigates cross-modal interference and enables effective integration without paired multimodal training data or additional fine-tuning. Extensive experiments on four datasets show that SSAM achieves state-of-the-art performance, outperforming prior training-free merging methods and models trained on large-scale paired multimodal data. Evaluation on three additional benchmarks shows that SSAM preserves and often improves performance on domain-specific tasks.

Overall, SSAM demonstrates that alignment in parameter space can serve as a practical alternative to costly joint multimodal training while maintaining strong performance.

Chapter 6

Conclusion

Multimodal learning has emerged as a fundamental paradigm for building intelligent systems capable of reasoning over heterogeneous data sources. This dissertation addressed three critical challenges that limit the practical deployment of multimodal systems: (i) fusing information from arbitrary combinations of input modalities, (ii) ensuring robustness under missing modality scenarios, and (iii) merging capabilities from multiple expert models to build a unified multimodal model without costly retraining or reliance on paired multimodal data.

6.1 Summary of Contributions

To address these challenges, this dissertation introduced a series of principled methods that progressively relax assumptions on modality availability, supervision, and data requirements.

6.1.1 Flexible Multimodal Fusion

In Chapter 2, we introduced MMSFormer, a transformer-based multimodal material and semantic segmentation architecture with a novel fusion block designed to handle arbitrary combinations of heterogeneous input modalities. Unlike prior methods restricted to fixed modality pairs, MMSFormer provides a scalable and computationally efficient mechanism for integrating heterogeneous modalities. Extensive experiments demonstrated that the proposed fusion strategy achieves state-of-the-art performance across multiple datasets and exhibits consistent performance improvements as additional modalities are incorporated. These results highlight the importance of designing modality-agnostic fusion mechanisms for flexible multimodal systems.

6.1.2 Robustness to Missing Modalities via Adaptation

In Chapter 3, we addressed the challenge of missing modality robustness through parameter-efficient adaptation. We proposed a lightweight framework that modulates intermediate feature representations of pretrained multimodal models using scaling and shifting. This approach enables dynamic adaptation of the model based on the available modalities without requiring model retraining or designing specialized architectures. Our results showed that the adapted models significantly reduce performance degradation in missing modality scenarios while maintaining strong performance when all modalities are present, using less than 1% additional learnable parameters.

6.1.3 Cross-Modal Approximation with Proxy Tokens

In Chapter 4, we further advanced robustness to missing modalities by introducing Cross-Modal Proxy Tokens (CMPTs). Unlike adaptation-based methods, CMPTs explicitly approximate

representations of missing modalities by leveraging information from available ones through attention mechanisms. This framework enables effective handling of diverse missing modality scenarios and achieves state-of-the-art performance across multiple datasets. Importantly, CMPTs maintain strong performance in both complete and missing modality settings, demonstrating their versatility and generalization capability.

6.1.4 Training-Free Multimodal Model Merging

In Chapter 5, we addressed the challenge of combining capabilities from multiple pre-trained multimodal models. We proposed Singular Subspace Alignment and Merging (SSAM), a training-free framework that aligns and merges independently trained specialist models within a shared low-rank subspace. By projecting model specific parameter updates onto a consensus subspace, SSAM reduces parameter interference while preserving complementary knowledge. This approach eliminates the need for paired multimodal training data and enables scalable integration of diverse modalities. Experimental results demonstrated that SSAM achieves state-of-the-art performance and, in some cases, surpasses jointly trained multimodal models.

6.1.5 Unified Perspective

Collectively, the contributions of this dissertation provide a unified framework for multimodal learning that emphasizes flexibility, robustness, and scalability. By addressing challenges at different levels—fusion, adaptation, representation, and model merging—this work moves toward practical multimodal systems capable of operating under real-world constraints.

6.2 Open Problems and Future Directions

Despite the progress made in this dissertation, several fundamental challenges are still open. Addressing these problems will be crucial for advancing multimodal learning toward more general, reliable, and efficient systems for real-world applications.

6.2.1 Generalization to New and Unseen Modality Combinations

Although the proposed methods improve flexibility across heterogeneous modality combinations, robust generalization to entirely new and unseen modality combinations remains an important open problem. One promising direction is to develop meta-learning or compositional learning frameworks that explicitly model higher-order interactions among modalities and enable adaptation to unseen combinations. Another direction is learning modality-invariant representations that capture shared semantic structure across modalities, allowing knowledge transfer beyond observed modality subsets. Progress in this area also requires benchmark datasets and evaluation methods that systematically assess generalization under novel modality configurations. In parallel, sparse or modular fusion strategies that dynamically select relevant modalities, together with efficient attention mechanisms that scale to large multimodal inputs, may offer practical pathways toward more generalizable multimodal systems.

6.2.2 Robustness Beyond Missing Modalities

This dissertation primarily focused on robustness under missing modality conditions. However, practical multimodal systems must also operate under broader forms of uncertainty and corruption. Real-world data may contain noisy or low-quality modalities with varying data reliabil-

ity, temporal or spatial misalignment across modalities, and even adversarial perturbations designed to exploit multimodal systems. Extending robustness beyond missing modalities therefore remains a critical research direction. In particular, uncertainty-aware or reliability-aware fusion mechanisms that dynamically assess the trustworthiness of each modality could provide a principled framework for robust multimodal reasoning under diverse failure modes.

6.2.3 Learning Without Paired Multimodal Training Data

While SSAM demonstrates the feasibility of training-free model merging without paired multimodal supervision, fully eliminating dependence on paired multimodal datasets remains an open challenge. Future research may explore self-supervised and contrastive learning approaches for cross-modal alignment that reduce or remove reliance on explicit paired data. Another promising direction is leveraging large-scale unimodal data together with weak or implicit supervision to bootstrap multimodal representations. Synthetic data generation and simulation environments may further provide scalable alternatives for learning multimodal representations when real paired datasets are scarce or expensive to collect.

6.2.4 Continual and Lifelong Multimodal Learning

Real-world multimodal systems must continually adapt to new modalities, tasks, and evolving data distributions, motivating the need for lifelong multimodal learning. Key challenges include incrementally incorporating new modalities without catastrophic forgetting, updating models efficiently without full retraining, and balancing adaptation with representational stability. Addressing these challenges may require integrating ideas from continual learning, parameter-efficient adaptation, and modular model composition. In particular, combining lightweight adaptation meth-

ods with model merging techniques presents a promising direction toward multimodal systems that can continuously expand their capabilities over time.

6.2.5 Interpretability and Explainability

As multimodal systems become increasingly complex, interpretability and explainability are becoming essential for both scientific understanding and trustworthy deployment. A key open question is how to identify the contributions of different modalities and their interactions in model predictions. Future work may focus on interpretable fusion mechanisms that expose cross-modal reasoning processes, attribution methods that quantify modality-specific contributions, and visualization tools for analyzing the structure of multimodal representations. Advancing interpretability in these directions could improve debugging, reliability, and trustworthiness in multimodal systems.

6.2.6 Toward General Multimodal Intelligence

A long-term goal of multimodal learning is the development of systems capable of general reasoning across diverse modalities and tasks. Achieving this objective will likely require unified architectures that integrate perception, reasoning, and generation within a common framework, together with scalable training paradigms that leverage both unimodal and multimodal data. Equally important is the development of robust evaluation frameworks that capture the complexity and variability of real-world environments. Progress along these directions may ultimately move multimodal learning beyond task-specific systems toward more general multimodal intelligence.

6.3 Final Remarks

This dissertation presented a comprehensive study of multimodal learning under realistic constraints, focusing on flexibility, robustness, and scalability. By addressing fundamental challenges in fusion, missing modality handling, and model merging, the proposed methods contribute toward the development of practical multimodal systems. As multimodal data continues to grow in scale and diversity, developing methods that can effectively leverage such data will be essential for the next generation of intelligent systems. The ideas and frameworks introduced in this work provide a foundation for future research in this rapidly evolving field.

Bibliography

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. [86](#), [90](#)
- [2] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*, pages 609–617, 2017. [72](#), [158](#)
- [3] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017. [72](#), [158](#)
- [4] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *Advances in Neural Information Processing Systems: Deep Learning Symposium*, 2016. [37](#), [40](#), [44](#)
- [5] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. MultiMAE: Multimodal multi-task masked autoencoders. In *European Conference on Computer Vision*, 2022. [4](#), [32](#), [41](#)
- [6] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 2236–2246, 2018. [46](#), [54](#), [139](#)
- [7] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. [86](#), [90](#)
- [8] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2018. [1](#), [31](#)
- [9] Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 1–9, 2022. [40](#), [43](#), [143](#)

- [10] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101—mining discriminative components with random forests. In *Computer vision—ECCV 2014: 13th European conference, zurich, Switzerland, September 6–12, 2014, proceedings, part VI* 13, pages 446–461. Springer, 2014. [139](#), [158](#)
- [11] Tim Broedermann, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Hrfuser: A multi-resolution sensor fusion architecture for 2d object detection. In *IEEE International Conference on Intelligent Transportation Systems (ITSC)*, 2023. [9](#), [12](#), [13](#), [29](#)
- [12] Yuqi Cai, Wujie Zhou, Liting Zhang, Lu Yu, and Ting Luo. Dhfnnet: dual-decoding hierarchical fusion network for rgb-thermal semantic segmentation. *The Visual Computer*, pages 1–11, 2023. [23](#)
- [13] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014. [72](#), [160](#)
- [14] Chi Chen, Yiyang Du, Zheng Fang, Ziyue Wang, Fuwen Luo, Peng Li, Ming Yan, Ji Zhang, Fei Huang, Maosong Sun, and Yang Liu. Model composition for multimodal large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11246–11262, Bangkok, Thailand, August 2024. Association for Computational Linguistics. [xv](#), [xviii](#), [88](#), [89](#), [92](#), [93](#), [97](#), [99](#), [100](#), [101](#), [102](#), [103](#), [105](#), [178](#), [179](#), [180](#), [182](#), [183](#)
- [15] Jin Chen, Xijun Wang, Zichao Guo, Xiangyu Zhang, and Jian Sun. Dynamic region-aware convolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8064–8073, 2021. [19](#)
- [16] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. [11](#), [19](#)
- [17] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. Beats: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058*, 2022. [99](#), [179](#)
- [18] Xiaokang Chen, Kwan-Yee Lin, Jingbo Wang, Wayne Wu, Chen Qian, Hongsheng Li, and Gang Zeng. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation. In *European Conference on Computer Vision (ECCV)*, pages 561–577, 2020. [1](#), [8](#), [12](#), [31](#), [50](#), [141](#)
- [19] Zhuokun Chen, Jinwu Hu, Zeshuai Deng, Yufeng Wang, Bohan Zhuang, and Mingkui Tan. Enhancing perception capabilities of multimodal llms with training-free fusion. *arXiv preprint arXiv:2412.01289*, 2024. [91](#)

- [20] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. [11](#)
- [21] Heng-Da Cheng, X. H. Jiang, Ying Sun, and Jingli Wang. Color image segmentation: advances and prospects. *Pattern recognition*, 34(12):2259–2281, 2001. [7](#)
- [22] Runxi Cheng, Feng Xiong, Yongxian Wei, Wanyun Zhu, and Chun Yuan. Whoever started the interference should end it: Guiding data-free model merging via task vectors. In *Forty-second International Conference on Machine Learning*, 2025. [88](#), [91](#), [94](#), [95](#), [100](#), [101](#), [102](#), [103](#), [105](#), [181](#), [183](#)
- [23] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, October 2014. [55](#)
- [24] Jun-Ho Choi and Jong-Seok Lee. EmbraceNet: A robust deep learning architecture for multimodal classification. *Information Fusion*, 51:259–270, 2019. [35](#), [66](#)
- [25] Ruiting Dai, Yuqiao Tan, Lisi Mo, Tao He, Ke Qin, and Shuang Liang. Muap: Multi-step adaptive prompt learning for vision-language model with missing modality. *arXiv preprint arXiv:2409.04693*, 2024. [67](#), [74](#)
- [26] Fuqin Deng, Hua Feng, Mingjian Liang, Hongmin Wang, Yong Yang, Yuan Gao, Junfeng Chen, Junjie Hu, Xiyue Guo, and Tin Lun Lam. Feanet: Feature-enhanced attention network for rgb-thermal real-time semantic segmentation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, page 4467–4473. IEEE Press, 2021. [19](#), [21](#), [50](#)
- [27] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. [18](#)
- [28] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*, 2023. [36](#)
- [29] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, June 2019. [68](#), [69](#), [157](#)
- [30] Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235, 2023. [36](#)

- [31] Shaohua Dong, Wujie Zhou, Xiaohong Qian, and Lu Yu. Gebnet: Graph-enhancement branch network for rgb-t scene parsing. *IEEE Signal Processing Letters*, 29:2273–2277, 2022. [23](#)
- [32] Shaohua Dong, Wujie Zhou, Caie Xu, and Weiqing Yan. Egfnets: Edge-aware guidance fusion network for rgb–thermal urban scene parsing. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–13, 2023. [19](#), [21](#), [23](#)
- [33] Reuben Dorent, Samuel Joutard, Marc Modat, Sébastien Ourselin, and Tom Vercauteren. Hetero-modal variational encoder-decoder for joint modality completion and segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pages 74–82. Springer, 2019. [33](#), [35](#)
- [34] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. [68](#), [69](#), [73](#), [174](#)
- [35] Chenzhuang Du, Yue Zhao, Chonghua Liao, Jiacheng You, Jie Fu, and Hang Zhao. Improving discriminative multi-modal learning with large-scale pre-trained models. *arXiv preprint arXiv:2310.05193*, 2023. [2](#), [73](#), [76](#), [77](#), [158](#)
- [36] Yiyang Du, Xiaochen Wang, Chi Chen, Jiabo Ye, Yiru Wang, Peng Li, Ming Yan, Ji Zhang, Fei Huang, Zhifang Sui, et al. Adamms: Model merging for heterogeneous multimodal large language models with unsupervised coefficient optimization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9413–9422, 2025. [91](#)
- [37] Ali Edalati, Marzieh Tahaei, Ivan Kobyzev, Vahid Partovi Nia, James J Clark, and Mehdi Rezagholizadeh. Krona: Parameter efficient tuning with kronecker adapter. *arXiv preprint arXiv:2212.10650*, 2022. [36](#)
- [38] Omar Elharrouss, Somaya Ali Al-Maadeed, Nandhini Subramanian, Najmath Ottakath, Noor Almaadeed, and Yassine Himeur. Panoptic segmentation: A review. *ArXiv*, abs/2111.10250, 2021. [7](#)
- [39] Siqi Fan, Zhe Wang, Yan Wang, and Jingjing Liu. SpiderMesh: Spatial-aware demand-guided recursive meshing for RGB-T semantic segmentation. *arXiv:2303.08692*, 2023. [35](#), [36](#), [50](#), [51](#), [60](#), [66](#)
- [40] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. Pmr: Prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20029–20038, 2023. [76](#), [158](#)
- [41] Nuno C Garcia, Pietro Morerio, and Vittorio Murino. Modality distillation with multiple stream networks for action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 103–118, 2018. [55](#)

- [42] Nuno Cruz Garcia, Sarah Adel Bargal, Vitaly Ablavsky, Pietro Morerio, Vittorio Murino, and Stan Sclaroff. Distillation multiple choice learning for multimodal action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2755–2764, 2021. [55](#)
- [43] Antonio Andrea Gargiulo, Donato Crisostomi, Maria Sofia Bucarelli, Simone Scardapane, Fabrizio Silvestri, and Emanuele Rodola. Task singular vectors: Reducing task interference in model merging. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18695–18705, 2025. [88](#), [91](#), [96](#), [100](#), [101](#), [102](#), [103](#), [105](#), [181](#), [183](#)
- [44] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017. [73](#)
- [45] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15180–15190, 2023. [90](#)
- [46] Ahmed Gomaa, Andreas Maier, and Ronak Kosti. Supervised contrastive learning for robust and efficient multi-modal emotion and sentiment analysis. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 2423–2429, 2022. [35](#)
- [47] Tingting Gong, Wujie Zhou, Xiaohong Qian, Jingsheng Lei, and Lu Yu. Globally contextually guided lightweight network for rgb-thermal urban scene understanding. *Engineering Applications of Artificial Intelligence*, 117:105510, 2023. [23](#)
- [48] Yuan Gong, Yu-An Chung, and James Glass. AST: Audio Spectrogram Transformer. In *Proc. Interspeech 2021*, pages 571–575, 2021. [68](#), [69](#), [73](#)
- [49] Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James R. Glass. Listen, think, and understand. In *The Twelfth International Conference on Learning Representations*, 2024. [86](#), [90](#)
- [50] Wenchao Gu, Shuang Bai, and Lingxing Kong. A review on 2d instance segmentation based on deep neural networks. *Image and Vision Computing*, 120:104401, 2022. [7](#)
- [51] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. Multimodal semi-supervised learning for image classification. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 902–909. IEEE, 2010. [1](#), [31](#)
- [52] Yanming Guo, Yu Liu, Theodoros Georgiou, and Michael S Lew. A review of semantic segmentation using deep neural networks. *International journal of multimedia information retrieval*, 7:87–93, 2018. [7](#)
- [53] Zhe Guo, Xiang Li, Heng Huang, Ning Guo, and Quanzheng Li. Deep learning-based image segmentation on multimodal medical imaging. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 3(2):162–169, 2019. [7](#)

- [54] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. Learning rich features from RGB-D images for object detection and segmentation. In *European Conference on Computer Vision*, pages 345–360, 2014. [45](#), [138](#)
- [55] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5108–5115, 2017. [45](#), [50](#), [138](#)
- [56] Abdul Mueed Hafiz and Ghulam Mohiuddin Bhat. A survey on instance segmentation: state of the art. *International journal of multimedia information retrieval*, 9(3):171–189, 2020. [7](#)
- [57] Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26584–26595, 2024. [87](#), [90](#), [100](#), [101](#)
- [58] Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, et al. Imagebind-llm: Multi-modality instruction tuning. *arXiv preprint arXiv:2309.03905*, 2023. [xiii](#), [87](#), [90](#), [100](#), [101](#), [102](#), [103](#)
- [59] Wei Han, Hui Chen, Min-Yen Kan, and Soujanya Poria. MM-align: Learning optimal transport-based alignment dynamics for fast and accurate inference on missing modality sequences. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10498–10511. Association for Computational Linguistics, December 2022. [36](#)
- [60] Xue Han, Yi-Tong Wang, Jun-Lan Feng, Chao Deng, Zhan-Heng Chen, Yu-An Huang, Hui Su, Lun Hu, and Peng-Wei Hu. A survey of transformer-based multimodal pre-trained modals. *Neurocomputing*, 515:89–106, 2023. [68](#)
- [61] Devamanyu Hazarika, Yingting Li, Bo Cheng, Shuai Zhao, Roger Zimmermann, and Soujanya Poria. Analyzing modality robustness in multimodal sentiment analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 685–696, July 2022. [32](#), [34](#), [35](#), [36](#), [54](#), [65](#)
- [62] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers. Fusetnet: incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Asian Conference on Computer Vision*, November 2016. [3](#), [8](#), [12](#), [19](#)
- [63] Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers. FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture. In *Asian Conference on Computer Vision*, 2016. [50](#)
- [64] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. In *International Conference on Learning Representations*, 2022. [36](#)

- [65] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. [35](#)
- [66] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. [xviii](#), [50](#), [51](#), [55](#), [151](#)
- [67] Xuehai He, Chunyuan Li, Pengchuan Zhang, Jianwei Yang, and Xin Eric Wang. Parameter-efficient model adaptation for vision transformers. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(1):817–825, Jun. 2023. [36](#)
- [68] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. [99](#), [106](#), [178](#)
- [69] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. [54](#)
- [70] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Larousilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019. [36](#)
- [71] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. [36](#), [40](#), [143](#)
- [72] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. [69](#), [71](#), [91](#), [95](#), [173](#), [179](#)
- [73] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. [16](#), [17](#)
- [74] Xinxin Hu, Kailun Yang, Lei Fei, and Kaiwei Wang. Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In *IEEE International Conference on Image Processing (ICIP)*, pages 1440–1444, 2019. [8](#), [12](#), [23](#)
- [75] Zhanbo Huang, Jinyuan Liu, Xin Fan, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Re-conet: Recurrent correction network for fast and efficient multi-modality image fusion. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 539–555, Cham, 2022. Springer Nature Switzerland. [19](#), [21](#)
- [76] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 603–612, 2019. [23](#)

- [77] Ahmed Hussen Abdelaziz, Barry-John Theobald, Paul Dixon, Reinhard Knothe, Nicholas Apostoloff, and Sachin Kajareker. Modality dropout for improved performance-driven talking faces. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 378–386, 2020. [32](#), [35](#), [40](#), [41](#), [45](#)
- [78] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*, 2023. [4](#), [88](#), [91](#), [94](#), [97](#), [100](#), [101](#), [102](#), [103](#), [105](#), [181](#), [183](#)
- [79] Sheikh Asif Imran, Mohammad Nur Hossain Khan, Subrata Biswas, and Bashima Islam. Llasa: A multimodal llm for human activity analysis through wearable and smartphone sensors. *arXiv preprint arXiv:2406.14498*, 2024. [90](#)
- [80] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015. [37](#), [40](#), [44](#)
- [81] Jaehyuk Jang, Yooseung Wang, and Changick Kim. Towards robust multimodal prompting with missing modalities. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8070–8074. IEEE, 2024. [xvii](#), [xix](#), [67](#), [74](#), [162](#), [163](#)
- [82] Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L Iuzzolino, and Kazuhito Koishida. Mmtm: Multimodal transfer module for cnn fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13289–13299, 2020. [19](#)
- [83] Sharif Amit Kamran and Ali Shihab Sabbir. Efficient yet deep convolutional neural networks for semantic segmentation. In *2018 International Symposium on Advanced Intelligent Informatics (SAIN)*, pages 123–130, 2018. [51](#)
- [84] Ramandeep Kaur and Sandeep Kautish. Multimodal sentiment analysis: A survey and comparison. *Research Anthology on Implementing Sentiment Analysis Across Multiple Disciplines*, pages 1846–1870, 2022. [31](#)
- [85] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. [72](#), [158](#)
- [86] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, volume 1, page 2, 2019. [54](#)
- [87] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*, 2019. [76](#), [157](#)

- [88] Douwe Kiela, Edouard Grave, Armand Joulin, and Tomas Mikolov. Efficient large-scale multi-modal classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. [75](#), [76](#)
- [89] Donggeun Kim and Taesup Kim. Missing modality prediction for unpaired multimodal learning via joint embedding of unimodal models. In *European Conference on Computer Vision*, pages 171–187. Springer, 2024. [xiii](#), [xvii](#), [xix](#), [63](#), [64](#), [67](#), [74](#), [77](#), [78](#), [79](#), [162](#), [163](#)
- [90] Wonjae Kim, Bokyoung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR, 2021. [2](#), [47](#), [56](#), [62](#), [63](#), [64](#), [74](#), [76](#), [79](#), [157](#), [163](#)
- [91] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9404–9413, 2019. [7](#)
- [92] Dawid Jan Kopiczko, Tijmen Blankevoort, and Yuki M Asano. VeRA: Vector-based random matrix adaptation. In *The Twelfth International Conference on Learning Representations*, 2024. [72](#), [173](#)
- [93] Kenneth Lau, Jonas Adler, and Jens Sjölund. A unified representation network for segmentation with missing modalities. *arXiv preprint arXiv:1908.06683*, 2019. [35](#)
- [94] Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal prompting with missing modalities for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14943–14952, 2023. [xiii](#), [xvii](#), [34](#), [47](#), [56](#), [57](#), [60](#), [63](#), [64](#), [66](#), [67](#), [68](#), [71](#), [74](#), [142](#), [163](#)
- [95] Gongyang Li, Yike Wang, Zhi Liu, Xinpeng Zhang, and Dan Zeng. Rgb-t semantic segmentation with location, activation, and sharpening. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(3):1223–1235, 2023. [19](#), [21](#)
- [96] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19108–19118, 2022. [xv](#), [xvii](#), [98](#), [100](#), [177](#), [183](#)
- [97] J. Li, H. Dai, H. Han, and Y. Ding. Mseg3d: Multi-modal 3d semantic segmentation for autonomous driving. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21694–21704, Los Alamitos, CA, USA, jun 2023. IEEE Computer Society. [8](#)
- [98] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. [99](#), [179](#)
- [99] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. [2](#), [62](#), [74](#), [157](#)

- [100] Ping Li, Junjie Chen, Binbin Lin, and Xianghua Xu. Residual spatial fusion network for rgb-thermal semantic segmentation. *arXiv:2306.10364*, 2023. 8, 12
- [101] Siting Li, Chenzhuang Du, Yue Zhao, Yu Huang, and Hang Zhao. What makes for robust multi-modal models in the face of missing modalities? *arXiv preprint arXiv:2310.06383*, 2023. 35
- [102] Yaowei Li, Ruijie Quan, Linchao Zhu, and Yi Yang. Efficient multimodal fusion via interactive prompting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2604–2613, 2023. 76
- [103] Zechen Li, Shohreh Deldari, Linyao Chen, Hao Xue, and Flora D. Salim. SensorLLM: Aligning large language models with motion sensors for human activity recognition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 354–379. Association for Computational Linguistics, 2025. 90
- [104] Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 36, 37, 40, 42, 44
- [105] Mingjian Liang, Junjie Hu, Chenyu Bao, Hua Feng, Fuqin Deng, and Tin Lun Lam. Explicit attention-enhanced fusion for rgb-thermal perception tasks. *IEEE Robotics Autom. Lett.*, 8(7):4060–4067, 2023. 8, 12
- [106] Yupeng Liang, Ryosuke Wakaki, Shohei Nobuhara, and Ko Nishino. Multimodal material segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19800–19808, June 2022. 7, 9, 12, 13, 18, 19, 20, 24, 29, 46, 138
- [107] Baihong Lin, Zengrong Lin, Yulan Guo, Yulan Zhang, Jianxiao Zou, and Shicai Fan. Variational probabilistic fusion network for RGB-T semantic segmentation. *arXiv preprint arXiv:2307.08536*, 2023. 35, 41, 50, 51, 60, 66
- [108] Bin Lin, Yang Ye, Bin Zhu, Jiayi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984, 2024. 86
- [109] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965, 2022. 36
- [110] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 3, 86, 90
- [111] Jianbo Liu, Junjun He, Jiawei Zhang, Jimmy S. Ren, and Hongsheng Li. Efficientfcn: Holistically-guided decoding for semantic segmentation. In *Computer Vision – ECCV 2020*, pages 1–17. Springer International Publishing, 2020. 23

- [112] Jinfu Liu, Wujie Zhou, Yueli Cui, Lu Yu, and Ting Luo. Gcnet: Grid-like context-aware network for rgb-thermal semantic segmentation. *Neurocomput.*, 506(C):60–67, sep 2022. [23](#)
- [113] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5802–5811, 2022. [19](#), [21](#)
- [114] Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *International Conference on Computer Vision*, 2023. [9](#), [18](#), [19](#), [21](#)
- [115] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*, 2024. [72](#), [173](#)
- [116] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12), December 2024. [99](#), [106](#), [178](#)
- [117] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. [xviii](#), [50](#), [151](#)
- [118] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *Annual Meeting of the Association for Computational Linguistics*, 2018. [54](#)
- [119] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. [11](#), [51](#)
- [120] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. [19](#), [47](#), [73](#)
- [121] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. [62](#), [73](#), [157](#)
- [122] Zelun Luo, Jun-Ting Hsieh, Lu Jiang, Juan Carlos Niebles, and Li Fei-Fei. Graph distillation for action detection with privileged modalities. In *Proceedings of the European conference on Computer Vision (ECCV)*, pages 166–183, 2018. [55](#), [141](#)
- [123] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on*

- Computer Vision and Pattern Recognition*, pages 18177–18186, 2022. [xiii](#), [4](#), [63](#), [64](#), [65](#), [66](#), [74](#), [77](#), [78](#), [79](#)
- [124] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18177–18186, 2022. [32](#), [34](#), [35](#), [36](#)
 - [125] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2302–2310, 2021. [41](#), [66](#)
 - [126] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602, 2024. [1](#), [90](#)
 - [127] Harsh Maheshwari, Yen-Cheng Liu, and Zsolt Kira. Missing modality robustness in semi-supervised multi-modal semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1009–1019, 2024. [32](#), [36](#), [40](#), [41](#), [45](#), [66](#)
 - [128] Daniel Marczak, Simone Magistri, Sebastian Cygert, Bartłomiej Twardowski, Andrew D. Bagdanov, and Joost van de Weijer. No task left behind: Isotropic model merging with common and task-specific subspaces. In *Forty-second International Conference on Machine Learning*, 2025. [91](#)
 - [129] Michael S Matena and Colin A Raffel. Merging models with fisher-weighted averaging. *Advances in Neural Information Processing Systems*, 35:17703–17716, 2022. [91](#)
 - [130] Brandon McKinzie, Joseph Cheng, Vaishaal Shankar, Yinfei Yang, Jonathon Shlens, and Alexander Toshev. On robustness in multimodal learning. *arXiv preprint arXiv:2304.04385*, 2023. [32](#)
 - [131] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3523–3542, 2021. [1](#), [7](#)
 - [132] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024. [86](#), [90](#)
 - [133] Shentong Mo and Pedro Morgado. Unveiling the power of audio-visual early fusion transformers with dense interactions through masked modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27186–27196, 2024. [2](#), [62](#)
 - [134] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in neural information processing systems*, 34:14200–14213, 2021. [2](#), [62](#)

- [135] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in neural information processing systems*, 34:14200–14213, 2021. [76](#), [157](#)
- [136] Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. ModDrop: Adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706, 2015. [4](#), [32](#), [35](#), [45](#)
- [137] Niki Nezakati, Md Kaykobad Reza, Ameya Patil, Mashhour Solh, and M. Salman Asif. Mmp: Towards robust multi-modal learning with masked modality projection. *arXiv preprint arXiv:2410.03010*, 2024. [66](#)
- [138] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, Andrew Y Ng, et al. Multimodal deep learning. In *ICML*, volume 11, pages 689–696, 2011. [2](#), [62](#)
- [139] Artemis Panagopoulou, Le Xue, Ning Yu, Junnan Li, Dongxu Li, Shafiq Joty, Ran Xu, Silvio Savarese, Caiming Xiong, and Juan Carlos Niebles. X-instructblip: A framework for aligning x-modal instruction-aware representations to llms and emergent cross-modal reasoning. *arXiv preprint arXiv:2311.18799*, 2023. [100](#), [101](#), [102](#), [103](#)
- [140] Aniello Panariello, Daniel Marczak, Simone Magistri, Angelo Porrello, Bartłomiej Twardowski, Andrew D. Bagdanov, Simone Calderara, and Joost van de Weijer. Accurate and efficient low-rank model merging in core space. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. [91](#)
- [141] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8238–8247, 2022. [73](#), [75](#), [76](#), [157](#)
- [142] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. [75](#), [76](#)
- [143] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7077–7087, 2021. [8](#), [12](#), [19](#)
- [144] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. [73](#), [99](#), [179](#)
- [145] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023. [86](#), [90](#)
- [146] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained

- transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, page 2359, 2020. [54](#)
- [147] Merrey Ramazanova, Alejandro Pardo, Humam Alwassel, and Bernard Ghanem. Exploring missing modality in multimodal egocentric datasets. *arXiv preprint arXiv:2401.11470*, 2024. [36](#)
 - [148] Md Kaykobad Reza, Ashley Prater-Bennette, and M. Salman Asif. Mmsformer: Multimodal transformer for material and semantic segmentation. *IEEE Open Journal of Signal Processing*, 5:599–610, 2024. [31](#), [62](#), [157](#)
 - [149] Md Kaykobad Reza, Ashley Prater-Bennette, and M. Salman Asif. Robust multimodal learning with missing modalities via parameter-efficient adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–13, 2024. [63](#), [65](#), [66](#), [68](#), [71](#)
 - [150] Giulia Rizzoli, Francesco Barbato, and Pietro Zanuttigh. Multimodal semantic segmentation in autonomous driving: A review of current approaches and future perspectives. *Technologies*, 10(4):90, 2022. [1](#), [31](#)
 - [151] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. [11](#)
 - [152] Swalpa Kumar Roy, Ankur Deria, Danfeng Hong, Behnood Rasti, Antonio Plaza, and Jocelyn Chanussot. Multimodal fusion transformer for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–20, 2023. [1](#), [31](#)
 - [153] Sakshi Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*, 2024. [99](#), [106](#), [178](#)
 - [154] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016. [46](#), [56](#), [139](#)
 - [155] Anmol Sharma and Ghassan Hamarneh. Missing MRI pulse sequence synthesis using multimodal generative adversarial network. *IEEE Transactions on Medical Imaging*, 39(4):1170–1183, 2019. [33](#), [35](#)
 - [156] Zhiyi Shi, Junsik Kim, Wanhua Li, Yicong Li, and Hanspeter Pfister. Mora: Lora guided multi-modal disease diagnosis with missing modality. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 273–282. Springer, 2024. [66](#)
 - [157] Ukcheol Shin, Kyunghyun Lee, In So Kweon, and Jean Oh. Complementary random masking for rgb-thermal semantic segmentation. In *2024 IEEE International Conference on Robotics*

- and *Automation (ICRA)*, pages 11110–11117. IEEE, 2024. [32](#), [33](#), [35](#), [36](#), [40](#), [41](#), [45](#), [50](#), [60](#), [65](#)
- [158] Shreyas S. Shivakumar, Neil Rodrigues, Alex Zhou, Ian D. Miller, Vijay Kumar, and Camillo J. Taylor. Pst900: Rgb-thermal calibration, dataset and segmentation network. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9441–9447, 2020. [9](#), [18](#), [22](#), [23](#)
 - [159] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *European Conference on Computer Vision*, 2012. [45](#), [138](#)
 - [160] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR*, 2015. [50](#), [51](#)
 - [161] Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14, 2017. [31](#)
 - [162] George Stoica, Pratik Ramesh, Boglarka Ecsedi, Leshem Choshen, and Judy Hoffman. Model merging with svd to tie the knots. *arXiv preprint arXiv:2410.19735*, 2024. [91](#)
 - [163] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023. [87](#), [90](#)
 - [164] Yuxiang Sun, Weixun Zuo, and Ming Liu. RTFNet: RGB-Thermal Fusion Network for Semantic Segmentation of Urban Scenes. *IEEE Robotics and Automation Letters*, 4(3):2576–2583, July 2019. [3](#), [8](#), [12](#), [23](#), [26](#), [50](#)
 - [165] Yuan Tang, Xu Han, Xianzhi Li, Qiao Yu, Yixue Hao, Long Hu, and Min Chen. Minigpt-3d: Efficiently aligning 3d point clouds with large language models using 2d priors. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6617–6626, 2024. [86](#), [90](#)
 - [166] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. [32](#), [36](#), [40](#)
 - [167] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 247–263, 2018. [72](#), [160](#)
 - [168] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, July 2019. [47](#), [54](#), [141](#), [157](#)

- [169] Paul Upchurch and Ransen Niu. A dense material segmentation dataset for indoor and outdoor scene parsing. In *European Conference on Computer Vision*, pages 450–466. Springer, 2022. [7](#)
- [170] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. [54](#), [55](#)
- [171] Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15878–15887, 2023. [35](#), [41](#), [65](#)
- [172] Jie Wang, Kechen Song, Yanqi Bao, Liming Huang, and Yunhui Yan. Cgfnets: Cross-guided fusion network for rgb-t salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5):2949–2961, 2022. [23](#)
- [173] Ke Wang, Nikolaos Dimitriadis, Guillermo Ortiz-Jiménez, François Fleuret, and Pascal Frossard. Localizing task information for improved model merging and compression. In *International Conference on Machine Learning*, 2024. [91](#)
- [174] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. Understanding convolution for semantic segmentation. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 1451–1460. Ieee, 2018. [7](#)
- [175] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12695–12705, 2020. [76](#), [157](#)
- [176] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *IEEE/CVF international conference on computer vision*, pages 568–578, 2021. [11](#), [13](#), [14](#)
- [177] Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frederic Precioso. Recipe recognition with large multimodal food dataset. In *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2015. [46](#), [57](#), [72](#), [139](#), [158](#)
- [178] Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, and Yunhe Wang. Multimodal token fusion for vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12186–12195, 2022. [2](#), [31](#), [33](#), [35](#), [36](#), [41](#), [51](#), [52](#), [60](#), [62](#), [66](#), [157](#)
- [179] Yikai Wang, Wenbing Huang, Fuchun Sun, Tingyang Xu, Yu Rong, and Junzhou Huang. Deep multimodal fusion by channel exchanging. *Advances in neural information processing systems*, 33:4835–4845, 2020. [51](#), [52](#), [62](#)

- [180] Yikai Wang, Fuchun Sun, Ming Lu, and Anbang Yao. Learning deep multimodal feature representation with asymmetric multi-layer fusion. In *ACM International Conference on Multimedia (ACM MM)*, 2020. [51](#), [52](#)
- [181] Shicai Wei, Chunbo Luo, and Yang Luo. Mmanet: Margin-aware distillation and modality-aware regularization for incomplete multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20039–20049, 2023. [66](#)
- [182] Yongxian Wei, Runxi Cheng, Weike Jin, Enneng Yang, Li Shen, Lu Hou, Sinan Du, Chun Yuan, Xiaochun Cao, and Dacheng Tao. Unifying multimodal large language model capabilities and modalities via model merging. *arXiv preprint arXiv:2505.19892*, 2025. [88](#), [89](#), [92](#), [96](#), [100](#), [101](#), [102](#), [103](#), [105](#), [181](#), [183](#)
- [183] Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. Longvlm: Efficient long video understanding via large language models. In *European Conference on Computer Vision*, pages 453–470. Springer, 2024. [86](#)
- [184] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019. [107](#)
- [185] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. [19](#), [20](#), [21](#), [23](#)
- [186] Sangmin Woo, Sumin Lee, Yeonju Park, Muhammad Adi Nugroho, and Changick Kim. Towards good practices for missing modality robust action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2776–2784, 2023. [33](#), [35](#), [55](#), [56](#), [60](#), [141](#)
- [187] Sangmin Woo, Sumin Lee, Yeonju Park, Muhammad Adi Nugroho, and Changick Kim. Towards good practices for missing modality robust action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2776–2784, 2023. [62](#), [157](#)
- [188] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. PMLR, 2022. [91](#)
- [189] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, and Hongseok Namkoong. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7959–7971, 2022. [36](#)
- [190] Huikai Wu, Junge Zhang, Kaiqi Huang, Kongming Liang, and Yizhou Yu. Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation. *arXiv preprint arXiv:1903.11816*, 2019. [11](#)

- [191] Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S Yu. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256. IEEE, 2023. [86](#), [90](#)
- [192] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. In *Forty-first International Conference on Machine Learning*, 2024. [87](#), [90](#)
- [193] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. [37](#), [44](#)
- [194] Yi Xiao, Felipe Codevilla, Akhil Gurram, Onay Urfalioglu, and Antonio M López. Multimodal end-to-end autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 23(1):537–547, 2020. [1](#), [31](#)
- [195] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Neural Information Processing Systems (NeurIPS)*, 2021. [xii](#), [8](#), [9](#), [11](#), [13](#), [14](#), [15](#), [17](#), [18](#), [50](#), [51](#), [140](#), [141](#)
- [196] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):502–518, 2022. [19](#), [21](#)
- [197] Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12113–12132, 2023. [1](#), [2](#), [31](#), [62](#), [68](#)
- [198] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. In *European Conference on Computer Vision*, pages 131–147. Springer, 2024. [86](#), [90](#), [99](#), [179](#)
- [199] Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36:7093–7115, 2023. [4](#), [88](#), [91](#), [95](#)
- [200] Enneng Yang, Li Shen, Zhenyi Wang, Guibing Guo, Xiaojun Chen, Xingwei Wang, and Dacheng Tao. Representation surgery for multi-task model merging. In *Forty-first International Conference on Machine Learning*, 2024. [91](#)
- [201] Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM international conference on multimedia*, pages 3480–3491, 2022. [xv](#), [xvii](#), [1](#), [98](#), [100](#), [177](#), [183](#), [184](#)
- [202] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024. [4](#), [86](#), [90](#)

- [203] Biting Yu, Luping Zhou, Lei Wang, Jurgen Fripp, and Pierrick Bourgeat. 3D cGAN based cross-modality MR image synthesis for brain tumor segmentation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 626–630, 2018. [4](#), [33](#), [35](#)
- [204] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *4th International Conference on Learning Representations, ICLR*, 2016. [50](#)
- [205] Jun Yu, Jing Li, Zhou Yu, and Qingming Huang. Multimodal transformer with multi-view visual representation for image captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(12):4467–4480, 2019. [1](#), [31](#)
- [206] Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Forty-first International Conference on Machine Learning*, 2024. [91](#)
- [207] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10790–10797, 2021. [47](#)
- [208] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, September 2017. [54](#)
- [209] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. [54](#)
- [210] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88, 2016. [46](#), [54](#), [139](#)
- [211] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. [174](#)
- [212] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. [90](#)
- [213] Haobo Zhang and Jiayu Zhou. Unraveling LoRA interference: Orthogonal subspaces for robust model merging. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26459–26472, Vienna, Austria, July 2025. Association for Computational Linguistics. [91](#)
- [214] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on Intelligent Transportation Systems*, 2023. [8](#), [9](#), [12](#), [13](#), [29](#)

- [215] Jiaming Zhang, Ruiping Liu, Hao Shi, Kailun Yang, Simon Reiß, Kunyu Peng, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. Delivering arbitrary-modal semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1136–1147, 2023. [9](#), [12](#), [13](#), [18](#), [19](#), [20](#), [24](#), [25](#), [29](#), [47](#), [50](#), [51](#), [140](#), [141](#), [151](#)
- [216] Jiaming Zhang, Ruiping Liu, Hao Shi, Kailun Yang, Simon Reiß, Kunyu Peng, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. Delivering arbitrary-modal semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1136–1147, 2023. [62](#), [157](#)
- [217] Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In *International conference on machine learning*, pages 41753–41769. PMLR, 2023. [75](#), [76](#), [158](#)
- [218] Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, and Huaxiu Yao. Multimodal representation learning by alternating unimodal adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27456–27466, 2024. [73](#), [75](#), [76](#), [77](#), [158](#)
- [219] Yifei Zhang, Désiré Sidibé, Olivier Morel, and Fabrice Mériaudeau. Deep multimodal fusion for semantic image segmentation: A survey. *Image and Vision Computing*, 105:104042, 2021. [7](#), [11](#)
- [220] Dexin Zhao, Zhi Chang, and Shutao Guo. A multimodal fusion approach for image captioning. *Neurocomputing*, 329:476–485, 2019. [1](#), [31](#)
- [221] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. [11](#)
- [222] Lin Zhao, Hui Zhou, Xinge Zhu, Xiao Song, Hongsheng Li, and Wenbing Tao. Lif-seg: Lidar and camera image fusion for 3d lidar semantic segmentation. *IEEE Transactions on Multimedia*, pages 1–11, 2023. [8](#), [12](#)
- [223] Shenlu Zhao, Yichen Liu, Qiang Jiao, Qiang Zhang, and Jungong Han. Mitigating modality discrepancies for rgb-t semantic segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–15, 2023. [23](#), [50](#), [51](#)
- [224] Shenlu Zhao and Qiang Zhang. A feature divide-and-conquer network for rgb-t semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(6):2892–2905, 2023. [23](#), [26](#)
- [225] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2), 2023. [4](#), [86](#), [90](#)
- [226] Zijia Zhao, Longteng Guo, Tongtian Yue, Sihan Chen, Shuai Shao, Xinxin Zhu, Zehuan Yuan, and Jing Liu. Chatbridge: Bridging modalities with large language model as a language catalyst. *arXiv preprint arXiv:2305.16103*, 2023. [87](#), [100](#), [101](#)

- [227] Zixiang Zhao, Shuang Xu, Chunxia Zhang, Junmin Liu, Jianshe Zhang, and Pengfei Li. Didfuse: Deep image decomposition for infrared and visible image fusion. In *IJCAI*, pages 970–976. ijcai.org, 2020. [19](#), [21](#)
- [228] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023. [99](#), [179](#)
- [229] Shenghe Zheng, Hongzhi Wang, Chenyu Huang, Xiaohui Wang, Tao Chen, Jiayuan Fan, Shuyue Hu, and Peng Ye. Decouple and orthogonalize: A data-free framework for lora merging. *arXiv preprint arXiv:2505.15875*, 2025. [91](#)
- [230] Benjia Zhou, Pichao Wang, Jun Wan, Yanyan Liang, and Fan Wang. A unified multimodal de-and re-coupling framework for rgb-d motion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11428–11442, 2023. [47](#), [55](#), [56](#), [60](#), [141](#), [149](#)
- [231] Benjia Zhou, Pichao Wang, Jun Wan, Yanyan Liang, Fan Wang, Du Zhang, Zhen Lei, Hao Li, and Rong Jin. Decoupling and recoupling spatiotemporal representation for rgb-d-based motion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20154–20163, 2022. [55](#), [56](#), [62](#), [141](#), [157](#)
- [232] Jingkai Zhou, Varun Jampani, Zhixiong Pi, Qiong Liu, and Ming-Hsuan Yang. Decoupled dynamic filter networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6647–6656, 2021. [19](#), [29](#)
- [233] Wujie Zhou, Shaohua Dong, Meixin Fang, and Lu Yu. Cacfnets: Cross-modal attention cascaded fusion network for rgb-t urban scene parsing. *IEEE Transactions on Intelligent Vehicles*, pages 1–10, 2023. [22](#), [23](#)
- [234] Wujie Zhou, Shaohua Dong, Jingsheng Lei, and Lu Yu. Mtanet: Multitask-aware network with hierarchical multimodal fusion for rgb-t urban scene understanding. *IEEE Transactions on Intelligent Vehicles*, 8(1):48–58, 2023. [23](#)
- [235] Wujie Zhou, Shaohua Dong, Caie Xu, and Yaguan Qian. Edge-aware guidance fusion network for rgb–thermal scene parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3571–3579, 2022. [23](#)
- [236] Wujie Zhou, Xinyang Lin, Jingsheng Lei, Lu Yu, and Jenq-Neng Hwang. Mffnet: Multiscale feature fusion and enhancement network for rgb–thermal urban road scene parsing. *IEEE Transactions on Multimedia*, 24:2526–2538, 2022. [23](#)
- [237] Wujie Zhou, Jinfu Liu, Jingsheng Lei, Lu Yu, and Jenq-Neng Hwang. Gmnet: Graded-feature multilabel-learning network for rgb-thermal urban scene semantic segmentation. *IEEE Transactions on Image Processing*, 30:7790–7802, 2021. [19](#), [21](#), [23](#)
- [238] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiayi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*, 2023. [99](#), [179](#)

A1 Robust Multimodal Learning with Missing Modalities via Parameter-Efficient Adaptation

A1.1 Datasets

MFNet Dataset introduced by [55], is a popular dataset for RGB-thermal urban scene segmentation, particularly in the context of supporting autonomous driving applications. It comprises a total of 1569 aligned pairs of RGB-thermal images. Within this collection, 820 image pairs were captured during daytime, while 749 pairs were acquired during nighttime. The dataset is divided into distinct training and test sets, each accompanied by pixel-level annotations that define semantic labels for nine classes. Each image is 640×480 pixels.

NYU Depth v2 (NYUDv2) Dataset from [159] is a well-known dataset for RGB-D semantic segmentation. This dataset contains 1449 pairs of aligned RGB-depth images of indoor scenes. The images are divided into training and test sets containing 795 and 654 pairs of images respectively. The dataset also provides per pixel annotations for 13 classes, 40 classes and 894 classes ground truth semantic labels. For our experiments we used the standard 40 classes annotation. Each image is 640×480 pixels and the dataset contains both raw and processed depth maps. For our experiments we used HHA images as proposed by [54] instead of depth maps.

Multimodal Material Segmentation (MCubeS) Dataset was introduced by [106] for accurate multimodal material segmentation with the help of thermal and polarized images alongside RGB images. This dataset has four modalities: RGB, Angle of Linear Polarization, Degree of Linear Polarization and Near-Infrared. Alongside these modalities, the dataset also provides ground truth annotation for semantic and material segmentation. There are 500 image sets divided into train,

validation and test sets having 302, 96 and 102 image sets respectively. The images are 1224×1024 pixels each and have 20 classes in total.

NTU RGB+D (NTU60) dataset [154] is a popular multimodal action recognition dataset. The dataset contains 56,880 action samples divided into 60 classes. The actions can be broadly categorized into three different categories: daily actions, medical conditions and mutual actions. It has four different input modalities: RGB videos (1920×1080), depth map sequences (512×424), infrared (IR) videos (512×424) and 3D skeletal data (25 major body joints). Three Microsoft Kinect V2 cameras were used to capture the videos simultaneously. It has two evaluation protocols: cross subject and cross view. We used RGB and depth data for our experiments and evaluated on cross subject protocol.

CMU-MOSI dataset from [210] is a popularly used for multimodal sentiment analysis. The dataset has 2199 samples each having audio, visual and text as input modalities. It is divided into train, validation and test sets containing 1284, 229 and 686 samples respectively along with annotated sentiment for each sample.

CMU-MOSEI is a large scale sentiment analysis dataset from [6]. It is 10 times larger than CMU-MOSI and contains audio, visual and text modalities along with ground truth sentiment annotations. The dataset contains 23453 samples divided into train, validation and test sets for multimodal sentiment analysis and emotion recognition.

UPMC Food-101 dataset [177] is a popular challenging multimodal classification dataset. It has 90,704 image-text pairs divided into train, validation and test sets. The dataset is annotated for 101 classes. Classes are identical to the ETHZ Food-101 dataset [10]. The samples are noisy as they were collected in an uncontrolled environment and thus huge diversity among samples is observed.

Table A1.1: Hyperparameters for the experiments on CMU-MOSI and CMU-MOSEI datasets for multimodal sentiment analysis.

Hyperparameters	CMU-MOSI	CMU-MOSEI
Batch Size	16	4
Initial Learning Rate	0.002	0.0005
Optimizer	Adam	Adam
Attention Dropout	0.3	0.4
Embedding Dropout	0.2	0.0
Output Dropout	0.5	0.5
Gradient Clip	0.6	0.6
Weight Decay	0.005	0.001
Temporal Conv Kernel Size (T/A/V)	5/5/5	5/1/3
# of Crossmodal Blocks	4	4

A1.2 Implementation Details

We used Python¹ 3.8.12 and PyTorch² 1.9.0 to for our implementation. The experiments were done using two NVIDIA RTX 2080 Ti GPUs. We applied automatic mixed precision (AMP) training provided by PyTorch. For CMNeXt model, we use their publicly available code³ and models trained on all the available modalities for each dataset. We trained the multimodal transformer models on all the modalities using the available code and preprocessed data from the repository⁴ for CMU-MOSI and CMU-MOSEI datasets.

MFNet Dataset: We divided the 4 channel RGB-T images into three channel RGB and one channel thermal images. Then data pre-processing and augmentation was applied following CMNeXt from [215]. MiT-B4 from [195] was the backbone for the base CMNeXt model. One set of scale and shift

¹<https://www.python.org/>

²<https://pytorch.org/>

³<https://github.com/jamycheung/DELIVER>

⁴<https://github.com/thuiar/MMSA>

parameters was learnt for each input modality combination. Input images were sized at 640×480 for both training and testing and we report single scale performance for all the experiments. The scale and shift parameters were trained for 100 epochs with a batch size of 4.

NYUDv2 Dataset: For processing depth maps, we follow SA-Gate by [18] and CMNeXt by [215] and use HHA-encoded images instead of raw depth maps. The already preprocessed dataset can be downloaded from the SA-Gate repository⁵. RGB and HHA images were sized at 640×480 pixels each and we used this size for training and testing. The backbone was set to MiT-B4 as suggested in CMNeXt paper. One set of scale and shift parameters was learnt for each input modality combination by feeding available input modalities and setting the missing modality to zero. We train the scale and shift parameters for 100 epochs with a batch size of 4 and report single scale performance.

MCubeS Dataset: We follow the same data pre-processing and augmentations used by the base CMNeXt model from [215]. MiT-B2 from [195] was used as the backbone for this dataset. We set the input image resolution to 512×512 during training and 1024×1024 during testing and report single scale performance with predicted segmentation maps sized at 1024×1024 . Similar to other two datasets, we train the learnable parameters for 100 epochs with a batch size of 4.

NTU RGB+D (NTU60) Dataset: We followed the same pre-processing and experimental setup as [231, 230]. We used RGB and depth data for our experiments and evaluated our method using cross subject protocol for fair comparison with [122, 186, 231, 230]. We extract 16 frames per video following previous methods and utilize DSN and DTN for spatial and temporal information encoding following [231, 230]. Then we train the SSF layers for 20 epochs with a batch size of 6.

CMU-MOSI and CMU-MOSEI Datasets: We used Multimodal Transformer (MulT) from [168]

⁵https://github.com/charlesCXX/RGBD_Semantic_Segmentation_PyTorch

Table A1.2: Learnable parameter counts for different parameter efficient model adaptation methods. As seen from the table, scale and shift introduce less than 0.7% of the total model parameters.

Method	Total Parameters (M)	Learnable Parameters (M)	% of Total Parameters
Norm	116.560	0.126	0.108
BitFit	116.560	0.378	0.324
LoRA	116.957	0.397	0.340
Scale and Shift	117.349	0.789	0.673

as the base model. Preprocessed datasets and all the configurations are available on the repository⁶.

First we trained the multimodal transformer (MulT) model on all the available modalities and then adapted the pretrained model for different missing modality scenarios. The hyperparameters for the experiments are shown in Table A1.1.

UPMC Food-101 Dataset: For multimodal classification on this dataset, we use ViLT as the base model and follow the experimental setup used by [94]. We use the same hyper-parameters and script for generating and evaluating different missing modality combinations. The SSF layers were trained for 10 epochs with a learning rate of $1e^{-5}$.

A1.3 Number of Learnable Parameters

We report the number of learnable parameters for different parameter-efficient adaptation techniques (for multimodal segmentation) in Table A1.2. We insert scale and shift layers after each linear, convolutional and norm (both batch norm and layer norm) layers. The number of learnable parameter varies with the size of the backbone. We used MiT-B4 as the backbone while counting these learnable parameters. Scale and shift adds only 0.789M learnable parameters which is less than 0.7% of the total model parameters. Despite this very few parameters, it improves performance

⁶<https://github.com/thuiar/MMSA>

Table A1.3: Performance comparison (% mIoU) of different parameter-efficient adaptation techniques for MFNet, NYUDv2, and MCubeS datasets. Each column reports mIoU of the Adapted model with the corresponding modalities, and Avg indicates average performance. A and D denote Angle and Degree of Linear Polarization.

Datasets	MFNet			NYUDv2			MCubeS			
Methods	RGB	Thermal	Avg	RGB	Depth	Avg	RGB	RGB-A	RGB-A-D	Avg
Pretrained	53.71	35.48	44.60	51.19	5.26	28.23	42.32	48.81	49.06	46.73
Dedicated	55.86	53.34	54.60	52.18	33.49	42.84	48.16	48.42	49.48	48.69
Scale Only	54.77	49.23	52.00	53.04	36.12	44.58	50.16	50.55	51.13	50.61
Shift Only	54.57	48.96	51.77	53.04	36.25	44.65	50.13	50.40	50.86	50.46
BitFit	54.39	49.07	51.73	53.09	36.64	44.87	50.19	50.57	51.07	50.61
LoRA	54.19	47.45	50.82	52.87	34.97	43.92	49.59	50.07	50.80	50.15
Norm	54.65	47.49	51.07	53.05	34.73	43.49	49.95	50.51	51.07	50.51
Scale and Shift	55.22	50.89	53.06	52.82	36.72	44.77	50.43	50.62	51.11	50.72

significantly in different missing modality scenarios. For this study we mainly focused on improving missing modality robustness and did not try to optimize the number of learnable parameters. We will leave that part for future studies.

A1.4 Performance Comparison with Parameter Efficient Model Adaption Techniques

We performed a detailed performance comparison with other parameter efficient model adaptation methods for the three segmentation datasets. Comparison among different parameter-efficient adaptation methods show that SSF-based adaptation provides overall best performance. We summarize the results for scale only, shift only, BitFit [9], norm layer fine-tuning and LoRA [71] in Table A1.3. We also show detailed comparison in Table A1.4 for RGB-thermal segmentation on MFNet dataset, Table A1.5 for RGB-depth segmentation on NYUDv2 dataset and Table A1.6

Table A1.4: Performance comparison with parameter efficient model adaptation techniques on CM-NeXt model for MFNet dataset. Average column indicates average performance when one of the two modalities gets missing. Mean accuracy, F1 score and % mIoU are shown for all experiments.

Methods	RGB			Thermal			Average		
	mAcc	F1	% mIoU	mAcc	F1	% mIoU	mAcc	F1	% mIoU
Pretrained	60.74	66.91	53.71	38.18	45.11	35.48	49.46	56.01	44.60
Dedicated	66.28	68.22	55.86	68.35	65.29	53.34	67.32	66.76	54.60
Scale Only	67.09	68.03	54.77	64.00	60.92	49.23	65.55	64.48	52.00
Shift Only	65.82	67.42	54.57	59.77	60.54	48.96	62.80	63.98	51.77
BitFit	66.49	67.40	54.39	61.06	60.59	49.07	63.78	64.00	51.73
LoRA	66.44	67.32	54.19	57.10	59.04	47.45	61.77	63.18	50.82
Norm	66.43	67.07	54.65	57.55	59.22	47.49	61.99	63.15	51.07
Scale and Shift	67.18	68.04	55.22	66.70	62.64	50.89	66.94	65.34	53.06

for multimodal material segmentation on MCubeS dataset. For each method, we take a model trained on all the available modalities. Then we freeze the pretrained weights and tune the learnable parameters for the corresponding adaption method. We have shown mean accuracy (mAcc), F1 score (F1) and % mIoU for each of the experiments.

Overall Performance Comparison

We summarize the results in Table A1.3. For MFNet dataset, SSF outperforms all the methods and performance is significantly better than the Pretrained model and close to the Dedicated models. For NYUDv2 and MCubeS datasets, the Adapted model performs better than both Pretrained and Dedicated models. These experiments also show that SSF performs better than other methods for most of the input modality combinations for all the datasets. We show a detailed comparison for each dataset in terms of mean accuracy, F1 score and % mIoU in Table A1.4 - A1.6.

Table A1.5: Performance comparison with parameter efficient model adaptation techniques on CM-NeXt model for NYUDv2 dataset. Average column indicates average performance when one of the two modalities gets missing. Mean accuracy, F1 score and % mIoU are shown for all experiments.

Methods	RGB			Depth			Average		
	mAcc	F1	% mIoU	mAcc	F1	% mIoU	mAcc	F1	% mIoU
Pretrained	64.10	65.70	51.19	8.30	7.95	5.26	36.20	36.83	28.23
Dedicated	66.00	66.62	52.18	44.80	46.79	33.49	55.40	56.71	42.84
Scale Only	68.18	67.38	53.04	51.54	49.88	36.12	59.86	58.63	44.58
Shift Only	67.54	67.35	53.04	50.30	49.76	36.25	58.92	58.56	44.65
BitFit	67.31	67.33	53.09	50.68	50.27	36.64	59.00	58.80	44.87
LoRA	66.67	67.14	52.87	49.34	48.66	34.97	58.01	57.90	43.92
Norm	67.18	67.34	53.05	48.74	48.06	34.73	57.96	57.70	43.89
Scale and Shift	67.96	67.18	52.82	52.42	50.60	36.72	60.19	58.89	44.77

Performance Comparison for RGB-Thermal Semantic Segmentation on MFNet Dataset

Table A1.4 summarizes the results on MFNet dataset when the base CMNeXt model is adapted with other parameter efficient model adaptation techniques. Experiments show that scale and shift shows the best performance in all three matrices compared to all other methods. It shows a significant improvement of +8.46% in mIoU, +9.33% in F1 score and +17.48% in mean accuracy on an average over the pretrained model. The average performance is also close to dedicatedly trained models.

Performance Comparison for RGB-Depth Semantic Segmentation on NYUDv2 Dataset

Similar trend is observed for RGB-Depth semantic segmentation on NYUDv2 dataset as shown in Table A1.5. Scale only and BitFit adapted models show slightly better performance for some of the matrices. But in most of the cases scale and shift adapted model performs better. For

Table A1.6: Performance comparison with different parameter efficient model adaptation techniques on CMNeXt model for MCubeS dataset. Average column indicates the average performance. Mean accuracy, F1 score and % mIoU are shown for all experiments.

Methods	RGB			RGB-AoLP			RGB-AoLP-DoLP			Average		
	mAcc	F1	% mIoU	mAcc	F1	% mIoU	mAcc	F1	% mIoU	mAcc	F1	% mIoU
Pretrained	51.63	55.91	42.32	58.66	62.00	48.81	60.06	62.43	49.06	56.78	60.11	46.73
Dedicated	57.70	60.95	48.16	57.56	61.17	48.42	59.12	61.91	49.48	58.13	61.34	48.69
Scale Only	59.64	63.06	50.16	60.28	63.55	50.55	60.96	64.14	51.13	60.29	63.58	50.61
Shift Only	59.82	63.17	50.13	60.10	63.36	50.40	60.61	63.78	50.86	60.18	63.44	50.46
BitFit	59.98	63.24	50.19	60.12	63.52	50.57	60.84	64.03	51.07	60.31	63.60	50.61
LoRA	59.08	62.50	49.59	59.81	63.05	50.07	60.69	63.84	50.80	59.86	63.13	50.15
Norm	59.57	62.89	49.95	60.22	63.49	50.51	60.98	64.08	51.07	60.26	63.49	50.51
Scale and Shift	60.23	63.41	50.43	60.40	63.59	50.62	60.94	64.04	51.11	60.52	63.68	50.72

all the matrices, scale and shift shows a significant improvement of +16.54% in mIoU, +22.05% in F1 score and +23.99% in mean accuracy over the pretrained model on an average and consistently outperforms dedicated training.

Performance Comparison for Multimodal Material Segmentation on MCubeS Dataset

Table A1.6 shows comparison with parameter efficient model adaptation techniques on MCubeS dataset. Scale and shift outperforms all other methods in most of the matrices for all input combinations. It also shows an improvement of +3.99% in mIoU, +3.57% in F1 score and +3.74% in mean accuracy on an average over the pretrained model. Furthermore, Scale and shift also outperforms dedicated training for all input modality combinations. These experiments corroborate the fact that scale and shift provides better model adaption for different missing modality scenarios.

Table A1.7: Per class % IoU comparison between pretrained and adapted CMNeXt model on MFNet dataset. Adapted model show better performance for most of the classes leading to overall performance improvement.

Modalities	Methods	Unlabeled	Car	Person	Bike	Curve	Car_Stop	Guardrail	Color_Cone	Bump	Mean
RGB-Thermal	CMNeXt	98.31	90.27	74.52	64.52	46.64	39.19	15.09	52.56	59.79	60.10
RGB	Pretrained	97.79	87.62	51.13	61.94	30.05	39.36	21.04	45.55	48.95	53.71
	Adapted	97.79	88.06	55.55	61.20	34.19	40.52	15.78	48.67	55.21	55.22
Thermal	Pretrained	95.97	55.24	68.47	9.27	31.85	2.75	0.0	16.87	38.92	35.48
	Adapted	97.46	82.83	70.12	49.03	40.89	26.79	1.84	36.24	52.83	50.89

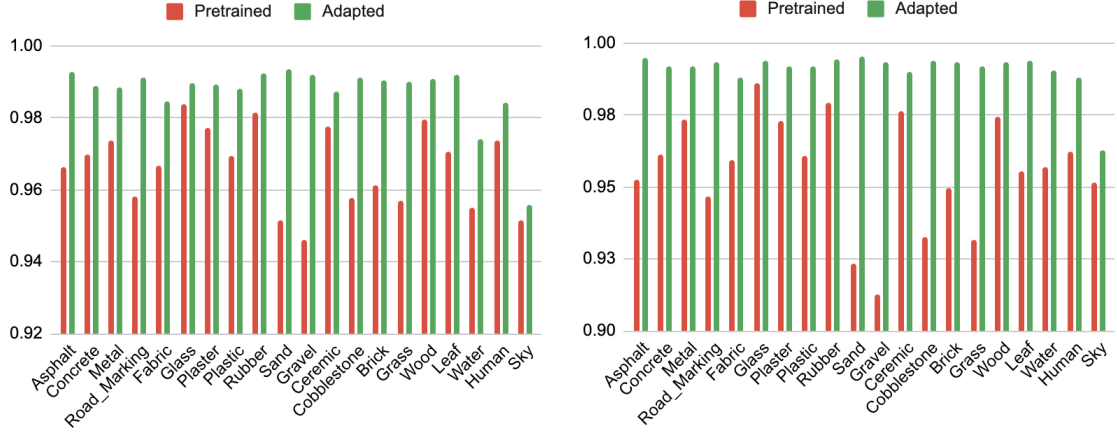
A1.5 Per Class IoU Comparison

To further analyze how the adaption is helping the model improve overall semantic and material segmentation performance, we conduct a per-class % intersection over union (IoU) analysis on the pretrained and adapted models. Table A1.7 and A1.8 summarize the results. We show the per class % IoU comparison for different missing modality situations on MFNet dataset on Table A1.7. From the table we can see that when RGB is available and thermal is missing, the adaptation helps improve performance for most of the classes. Though we see some performance drop for bike (-0.74%) and guardrail (-5.25%) classes, the rest of the classes have better % IoU than the pretrained model. Bump (+6.26%), person (+4.42%), and curve (+4.14%) classes show greater improvement after adaptation. When thermal is available and RGB is missing, adaptation improves performance for all the classes. Among the classes, bike (+39.76%), car (+27.59%), car stop (+24.04%), color cone (+19.37%) and bump (+13.91%) are showing impressive performance improvement.

Table A1.8: Per class % IoU comparison between pretrained and adapted CMNeXt model on MCubeS dataset. Adapted model show better performance for most of the classes leading to overall performance improvement. Here A, D, and N stand for Angle of Linear Polarization (AoLP), Degree of Linear Polarization (DoLP) and Near-Infrared (NIR) respectively.

Modalities	Methods	Asphalt	Concrete	Metal	Road_Marking	Fabric	Glass	Plaster	Plastic	Rubber	Sand	Gravel	Ceramic	Cobblestone	Brick	Grass	Wood	Leaf	Water	Human	Sky	Mean
RGB-A-D-N	CMNeXt	84.4	44.9	53.9	74.6	32.1	54.0	0.8	28.7	29.8	67.0	66.2	27.7	68.5	42.8	58.7	49.7	75.3	55.6	19.1	96.52	51.5
RGB	Pretrained	69.7	39.2	47.6	67.3	26.9	44.6	0.2	20.9	15.2	61.8	36.7	19.1	67.2	36.0	49.5	36.1	71.6	36.1	14.7	86.3	42.3
	Adapted	85.8	43.7	52.6	73.8	27.9	51.0	0.8	24.2	30.4	67.8	72.9	27.1	68.1	42.9	57.6	49.0	74.9	43.4	18.3	96.5	50.4
RGB-A	Pretrained	83.2	43.3	50.7	72.6	26.4	51.9	0.2	28.1	22.2	67.7	63.4	22.7	67.5	40.6	54.4	44.9	73.9	44.8	21.8	96.0	48.8
	Adapted	84.4	45.4	53.8	74.5	30.4	53.2	0.6	26.9	28.8	69.0	69.3	24.8	67.5	43.2	58.4	48.2	75.1	48.1	14.4	96.4	50.6
RGB-A-D	Pretrained	84.5	41.2	46.7	72.8	25.2	51.6	0.3	26.1	28.8	66.7	65.6	26.0	66.5	40.4	50.0	45.1	72.7	49.4	25.6	96.3	49.1
	Adapted	84.1	45.6	54.1	74.6	30.5	54.2	0.6	28.1	30.1	69.0	67.6	25.9	67.8	43.8	58.0	49.1	75.0	53.7	13.7	96.5	51.1

Results for MCubeS dataset is shown on Table A1.8. Here A, D, and N stand for angle of linear polarization (AoLP), degree of linear polarization (DoLP) and near-infrared (NIR) respectively. Experiments show that when only RGB is available and the rest of the modalities are missing, the adapted model performs better in detecting all the 20 classes present in the dataset. Gravel (36.2%), asphalt (16.1%), rubber (15.2%), wood (12.9%) and sky (10.2%) are some of the classes who show the most performance boost after adaptation. In other input combinations, most of the classes see performance improvement compared to the pretrained model. Though we see some performance drop in a few classes, most of the classes show improvement in % IoU which leads to the overall performance improvement after adaption.



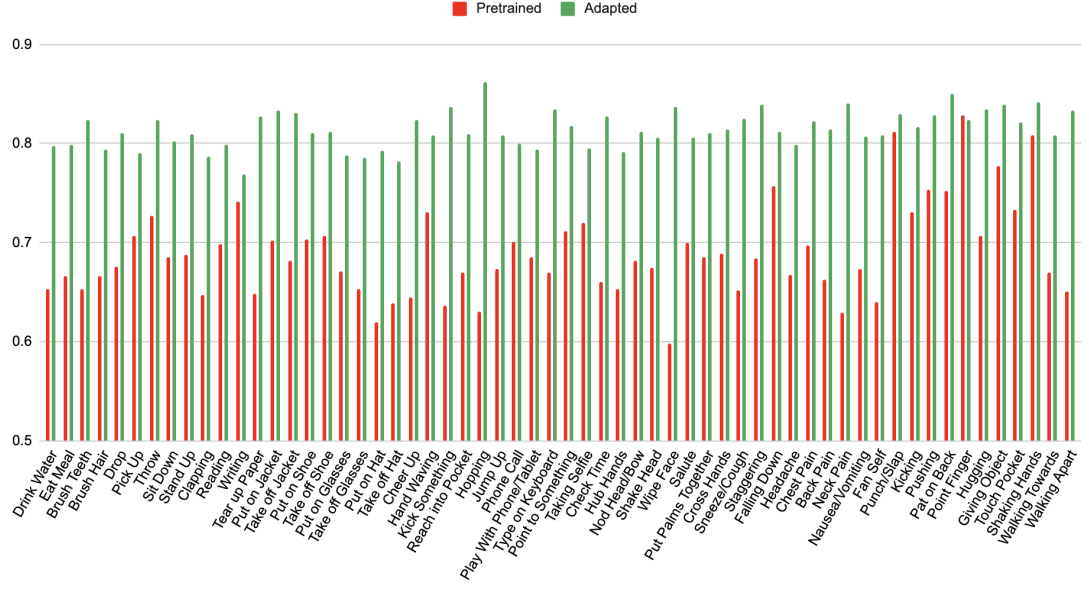
(a) Available: RGB, AoLP; Missing: DoLP, NIR (b) Available: RGB, AoLP, DoLP; Missing: NIR

Figure A1.1: Cosine similarity between complete and missing modality features of the pretrained model and complete and missing modality features of the adapted model under different missing modality scenarios on MCubeS dataset. The adapted model shows higher similarity to the complete modality features indicating less deviation and better handling of missing modalities.

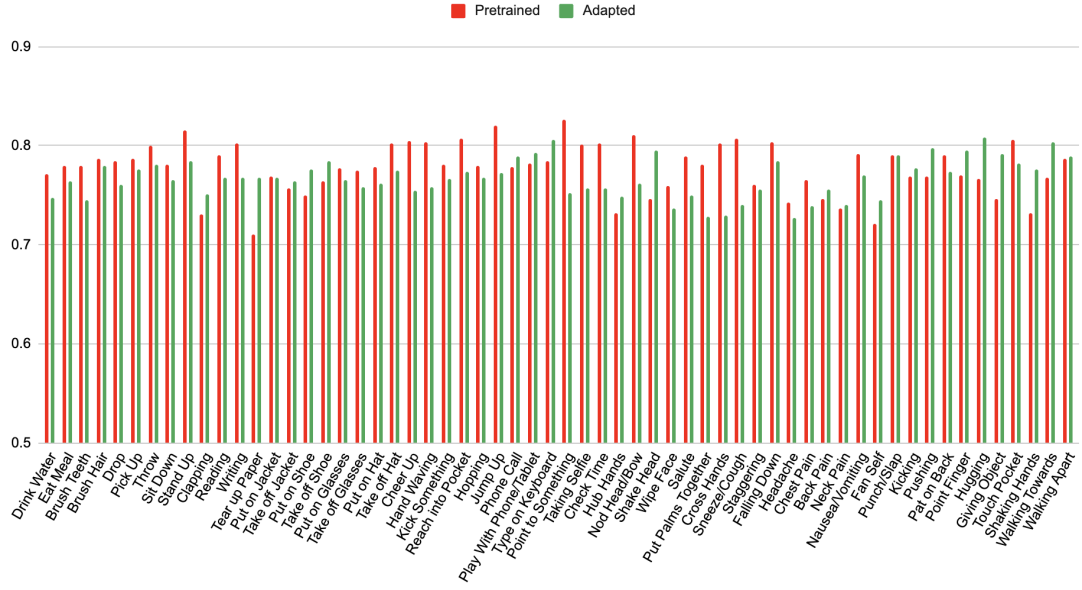
A1.6 Cosine Similarity Analysis

We show the cosine similarity for different missing modality scenarios on MCubeS dataset using CMNeXt as the base model in Figure A1.1. These results show similar trends as discussed in Section 3.4.7, demonstrating a consistent increase in cosine similarity for all of the classes. This leads to an overall increase in performance for the adapted model compared to the pretrained model under various missing modality scenarios.

For multimodal action recognition, we utilize the UMDR [230] as the base pretrained model. When RGB is available but depth is missing, the adapted model demonstrates a significant increase in cosine similarity compared to the pretrained model, as illustrated in Figure A1.2. This enhancement translates to a 1.06% improvement in overall performance, as shown in Table 3.5. When depth is available and RGB is missing, both the pretrained and adapted models show comparable cosine similarity. This is because the base UMDR model can handle depth-only data quite well



(a) Available: RGB - Missing: Depth



(b) Available: Depth - Missing: RGB

Figure A1.2: Cosine similarity between complete and missing modality features of the pretrained model and the adapted model under different missing modality scenarios on NTU RGB+D dataset.

Table A1.9: Missing modality performance comparison of base CMNeXt and Adapted CMNeXt model with Resnet-34 [66], Resnet-50 [66], and Swin-S [117] backbones on MFNet dataset.

Method	Backbone	Params (M)	RGB-Thermal		RGB		Thermal		Average	
			mAcc	%mIoU	mAcc	%mIoU	mAcc	%mIoU	mAcc	%mIoU
CMNeXt	ResNet-34	50.72	50.56	45.61	23.95	16.98	19.25	16.94	21.60	16.96
Adapted CMNeXt	ResNet-34	50.91	50.56	45.61	31.68	27.13	32.04	27.81	31.86	27.47
CMNeXt	ResNet-50	54.68	50.69	46.05	12.00	11.18	19.45	18.38	15.73	14.78
Adapted CMNeXt	ResNet-50	54.90	50.69	46.05	39.84	34.69	28.70	25.59	34.27	30.14
CMNeXt	Swin-S	123.73	45.24	41.02	15.95	11.98	14.77	13.91	15.36	12.95
Adapted CMNeXt	Swin-S	124.14	45.24	41.02	39.69	34.11	28.01	25.05	33.85	29.58

and maintain a higher performance, resulting in similar performance metrics for both the pretrained and adapted models. This consistency across datasets and tasks strengthens the generalizability and effectiveness of the adaptation process in promoting model robustness to missing modalities.

A1.7 Effectiveness of Our Adaptation Method on Different Backbones

To assess the effectiveness of our method across different backbones, we replaced the MiT-B4 backbone in the CMNeXt model with ResNet-34, ResNet-50, and Swin-S. We use the same experimental setup as the original CMNeXt model [215] while training the base models. In order to use the default window size and other configurations of Swin-S backbone, we resize the images to 448×448 for training and testing. For ResNets, we use the default image resolution of 640×480 . Other hyperparameters are selected in the same manner as the original CMNeXt model. Then we adapt the pretrained base models for different missing modality scenarios using our approach. For adaptation, we use the same hyper-parameters as described in Section 3.4.2.

Table A1.10: Per-class IoU comparison among Dedicated, Pretrained, and Adapted model on MFNet Dataset.

Class	RGB			Thermal		
	Dedicated	Pretrained	Adapted	Dedicated	Pretrained	Adapted
Unlabeled	97.95	97.79	97.79	97.85	95.97	97.45
Car	88.21	87.62	88.06	85.84	55.24	83.39
Person	62.97	51.13	55.55	71.00	68.47	69.96
Bike	63.03	61.94	61.20	56.50	9.27	49.19
Curve	36.02	30.05	34.19	40.03	31.85	40.53
Car_Stop	40.45	39.36	40.52	25.70	2.75	25.14
Guardrail	11.35	21.04	15.78	7.52	0.00	1.55
Color_Cone	52.99	45.55	48.67	42.24	16.87	33.85
Bump	49.78	48.95	55.21	53.35	38.92	53.22
Average	55.86	53.71	55.22	53.34	35.48	50.48

We summarize the test set performance of the pre-trained and adapted CMNeXt models on MFNet dataset with different backbones in Table A1.9. Our approach needs a small number of learnable parameters for each of the backbones. ResNet-34, ResNet-50, and Swin-S add only 0.37%, 0.40% and 0.33% additional learnable parameters, respectively. Adapted CMNeXt with ResNet-34 shows 10.26 and 10.51 points improvement in mean accuracy (mAcc) and %mIoU, respectively on average. Adapted CMNeXt with ResNet-50 provides an improvement of 18.54 and 15.36 points while the Adapted CMNeXt with Swin-S gains 18.49 and 16.63 points improvement in mAcc and %mIoU, respectively on average.

These results confirm that our adaptation approach generalizes across backbones, delivering significant performance gains with a small number of additional parameters.

Table A1.11: Per-class IoU comparison among Dedicated, Pretrained, and Adapted model on MCubeS Dataset.

Class	RGB			RGB+AoLP			RGB+AoLP+DoLP		
	Dedicated	Pretrained	Adapted	Dedicated	Pretrained	Adapted	Dedicated	Pretrained	Adapted
Asphalt	85.68	69.74	85.80	87.45	83.20	84.43	87.02	84.45	84.14
Concrete	43.41	39.18	43.72	45.29	43.28	45.36	43.82	41.20	45.57
Metal	51.36	47.57	52.64	53.11	50.72	53.75	50.65	46.68	54.08
Road_Marking	64.95	67.33	73.79	59.53	72.63	74.50	71.29	72.76	74.58
Fabric	30.08	26.88	27.86	30.52	26.37	30.42	29.86	25.20	30.50
Glass	51.20	44.56	50.95	53.25	51.91	53.22	50.88	51.59	54.23
Plaster	0.11	0.16	0.75	0.37	0.21	0.63	0.37	0.33	0.63
Plastic	21.81	20.86	24.18	23.22	28.05	26.90	22.05	26.05	28.10
Rubber	26.11	15.24	30.40	27.96	22.23	28.76	27.63	28.76	30.15
Sand	59.65	61.76	67.80	58.75	67.67	68.97	62.88	66.72	69.01
Gravel	64.24	36.73	72.86	66.24	63.40	69.25	71.08	65.63	67.63
Ceramic	27.11	19.12	27.10	30.11	22.70	24.83	30.30	25.97	25.86
Cobblestone	68.76	67.15	68.06	71.14	67.48	67.51	71.25	66.47	67.80
Brick	41.34	35.96	42.93	41.90	40.55	43.24	43.52	40.38	43.79
Grass	58.93	49.46	57.57	56.89	54.44	58.36	56.77	49.92	58.01
Wood	45.12	36.13	49.02	44.99	44.91	48.21	44.64	45.08	49.07
Leaf	76.62	71.55	74.91	75.70	73.90	75.05	75.60	72.67	75.03
Water	45.12	36.13	43.41	41.46	44.77	48.13	52.39	49.43	53.73
Human	4.27	14.70	18.27	3.32	21.79	14.43	1.33	25.59	13.71
Sky	96.39	86.25	96.47	96.53	96.04	96.42	96.34	96.33	96.49
Average	48.11	42.32	50.42	48.39	48.81	50.62	49.48	49.06	51.11

A1.8 Why Dedicated Model Performs Better on MFNet Dataset?

As shown in Table 3.1, the dedicated baseline performs better than adapted model on MFNet dataset. We further investigated the MFNet results and found that in the pretrained model a dominant modality (e.g., only RGB or Thermal input) contributes to the prediction of some specific classes as highlighted with gray background in Table A1.10. If the dominant modality is missing,

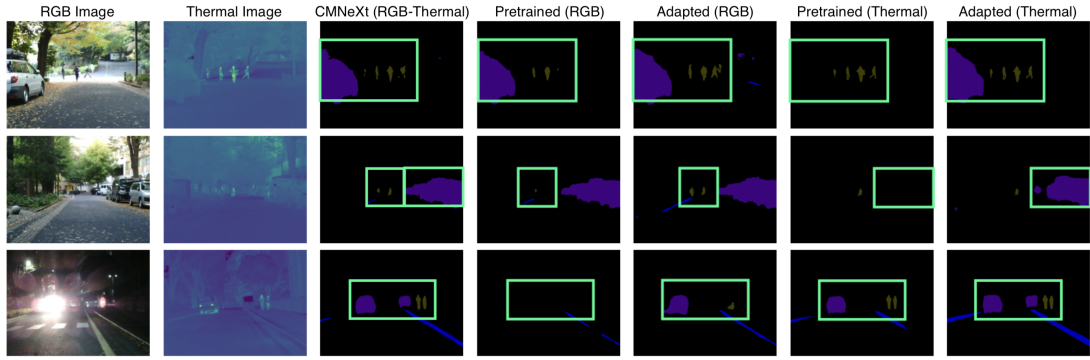
a significant performance drop is observed for that class. For instance, for Person class Thermal modality dominates as RGB only provides 51.13% IoU and Thermal only provides 68.47% IoU; for Bike class RGB dominates as RGB only provides 61.94% IoU whereas Thermal only provides 9.27% IoU. The adapted model exhibits a similar pattern since it builds on the pretrained model. While adaptation improves the IoU for some classes significantly, it falls short in some cases. On the other hand, dedicated networks are trained independently with RGB or Thermal and are forced to make correct predictions with the respective modality only. That can be one possible reason why dedicated networks perform better than adapted for MFNet dataset.

For other datasets like MCubeS , as shown in Table [A1.11](#), the adapted model can improve performance on most of the classes and outperforms the dedicated network. We also observe an overall performance boost compared to the pretrained and dedicated networks.

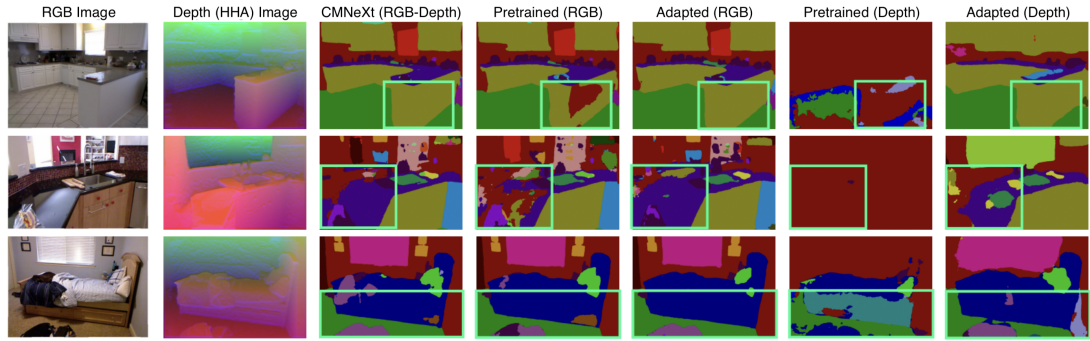
The main drawback of the dedicated models is that we have to train as many independent models as the number of modality combinations. In contrast, our proposed method adapts a single pretrained model for different missing modality scenarios using a small number of additional parameters. Thus, our approach offers compute and memory efficiency while performing on par or better than dedicated models.

A1.9 Visualization of Predicted Segmentation Maps

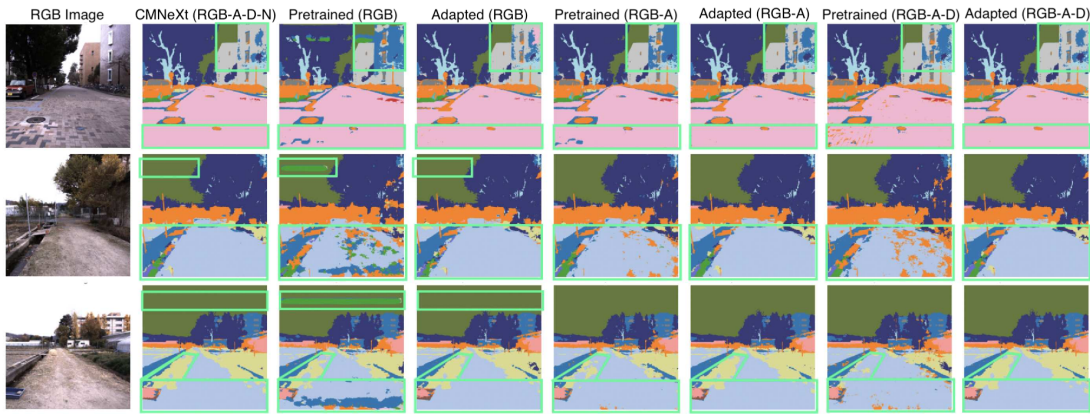
We show the predicted segmentation maps from the pretrained and adapted models in Figure [A1.3](#). For each dataset, we show the input images, predictions from the base CMNeXt model when all the modalities are available, predictions from the adapted and pretrained models for different missing modality scenarios. For brevity, we only show RGB input images for MCubeS



(a) Visualization of predictions on MFNet dataset.



(b) Visualization of predictions on NYUDv2 dataset.



(c) Visualization of predictions on MCubeS dataset.

Figure A1.3: Visualization of predicted segmentation maps for pretrained and adapted models on MFNet, NYUDv2, and MCubes datasets for multimodal semantic and material segmentation.

dataset. A, D and N stand for angle of linear polarization (AoLP), degree of linear polarization (DoLP) and near-infrared (NIR) respectively. Modalities that are available during testing are shown in parenthesis while other modalities are missing.

For MFNet dataset, Figure [A1.3a](#) shows that when only RGB is available, the pretrained model performs very poorly in detecting humans. On the other hand, if only thermal is available, the pretrained model can not detect cars very accurately. But the adapted model can detect both humans and cars more accurately in both of the scenarios. In all the cases, the predictions from the adapted model is closer to the predictions of the base CMNeXt model when all the modalities are available. Predictions from NYUDv2 dataset is shown on Figure [A1.3b](#). We can see that the adapted model can identify bed, furniture and other classes more accurately than the pretrained model for different missing modality scenarios. The pretrained model performs very poorly when only depth is available and RGB is missing. But detection accuracy improves significantly after model adaptation. For MCubeS dataset, as seen in Figure [A1.3c](#), predictions from the pretrained model shows artifacts when detecting different materials. On the other hand, the adapted model is showing more accuracy in detecting sky, cobblestone, sand and brick. For all the three datasets, the predictions from the adapted model is more accurate and closer to the all modality predictions of the base CMNeXt model.

A2 Robust Multimodal Learning via Cross-Modal Proxy Tokens

A2.1 Related Work on Multimodal Learning

In this section, we provide a general overview of multimodal learning, which involves integrating information from different input sources to improve performance on downstream tasks. We can broadly divide multimodal learning approaches into two main categories: designing specialized models and adapting existing models.

Designing specialized models to effectively fuse multiple input sources is a common strategy in multimodal learning. For vision-language tasks, models such as ViLT [90], ViLBERT [121], and Visual-BERT [99] process image patches and text tokens directly within a transformer architecture for tasks like image captioning and visual question answering. MMBT [87] integrates both text and images within a BERT-based framework [29] to model cross-modal dependencies. MBT [135] leverages fusion bottlenecks for efficient modality-specific interaction, particularly in audio-video tasks. Specialized multimodal models have also been proposed for other applications, including action recognition [187, 231], segmentation [216, 178, 148], and sentiment analysis [187, 168]. While these models excel in specific tasks, extending them to new modalities or domains often proves to be challenging and resource-intensive, as it requires retraining or significant modification of the model architecture.

Adapting existing models offers an alternative approach to multimodal learning, focusing on the use of pre-trained unimodal models and adapting them for multimodal tasks. One such strategy is gradient modulation, where techniques like G-Blend [175] modulate gradients to combine unimodal and multimodal knowledge. OGM-GE [141] improves upon this by emphasizing modality-specific information, enhancing the model’s ability to leverage unique features from each

input source. PMR [40] further refines this approach by modulating parameter groups based on the relevance of each modality, while QMF [217] introduces quadratic modulation for adaptive fusion. More recently, MLA [218] proposed alternating adaptation, which toggles between unimodal adaptation and gradient modulation to enable more efficient cross-modal learning. Another notable approach, MMLoRA [35], uses a two-phase strategy: unimodal fine-tuning followed by multimodal adaptation, which helps achieve optimal performance across a variety of tasks. While these methods allow for effective cross-modal learning, they often require significant computational resources and learning a large number of parameters to achieve optimal performance on downstream tasks.

A2.2 Detailed Dataset Description

UPMC Food-101 dataset [177] is a popular and challenging multimodal classification dataset. It has two input modalities: image and text. The dataset has 90,704 samples divided into training and test sets having 67,988 and 22,716 samples, respectively. It contains 101 classes, which are the same as those in the ETHZ Food-101 dataset [10]. The samples are noisy, each category contains around 5% irrelevant images as they were collected in an uncontrolled environment without any human human intervention during the data collection process.

MM-IMDb dataset [3] is a widely used multimodal dataset for multi-label movie genre classification task. It consists of 25,959 samples, each containing both image and text as input modality. The dataset is divided into three subsets: 15,552 samples for training, 2,608 for validation, and 7,799 for testing. It has 23 classes and each sample can have one or more genre. This dataset serves as a standard benchmark for evaluating multimodal models in the multi-label classification task.

Kinetics-Sound (KS) dataset [2] is a subset of the Kinetics-400 dataset [85]. It is used for human

Table A2.12: List of all the hyperparameters used in our experiments.

Hyperparameters	Image-Text Datasets	Audio-Video Datasets
Learning Rate	10^{-3}	5×10^{-5}
Scheduler	Polynomial	Polynomial
Power	0.9	0.9
Num Epochs/Warmup Epochs	10/5	100/5
Optimizer	AdamW	AdamW
ϵ	10^{-8}	10^{-8}
Weight Decay	0.02	0.02
λ	0.20	0.20
LoRA Rank/Alpha/Dropout	1/1/0.1	1/1/0.1
Image Encoder	ViT-B	ViT-B
Text Encoder	BERT-Base	-
Audio Encoder	-	AST
Image Resolution	224×224	224×224
Patch Size	16×16	16×16
Batch Size	8	4
Audio Sampling Rate	-	16 kHz
Frequency Bins	-	128
Maximum Length	-	1024 Frames
Frames Per Video	-	3
Python Version	3.8.19	3.8.19
PyTorch Version	2.2.2	2.2.2
GPU Model	RTX 2080Ti	RTX 2080Ti
Num GPUs	2	1

action recognition using audio and video as input modalities. This dataset includes 31 different action classes that span a variety of everyday human activities, making it suitable for evaluating multimodal models. The dataset is split into two subsets: a training set containing 14,739 samples and a test set containing 2,594 samples.

Audio-Visual Event (AVE) Localization dataset [167] is a benchmark dataset used for multimodal event localization task. It contains 4,143 10-second videos. The dataset is divided into training, validation, and test sets, with 3,339 videos for training, 402 for validation, and 402 for testing. The dataset encompasses 28 different event categories providing a broad range of scenarios where both audio and visual signals are crucial for accurate event detection. All videos in the AVE dataset are collected from YouTube.

CREMA-D dataset [13] is used for multimodal emotion recognition. It has audio and video as input modalities. The dataset contains 7,442 short video clips performed by 91 actors. The clips cover six emotions: angry, happy, sad, neutral, disgust, and fear. Emotion labels were obtained from 2,443 crowd-sourced raters to ensure diverse evaluations. The dataset is divided into 6,698 samples for training and 744 samples for testing.

A2.3 Implementation Details

In Section 4.4.2 of the main paper, we briefly discussed the hyperparameter settings used in our experiments. Here, we provide a complete list of all the hyperparameters and their values in Table A2.12. The “Image-Text Datasets” column includes the hyperparameters for UPMC Food-101 and MM-IMDb datasets, while the “Audio-Video Datasets” column lists the hyperparameters for Kinetics-Sound, AVE, and CREMA-D datasets.

A2.4 Determining the Optimal LoRA Rank

To determine the optimal LoRA rank, we conducted an ablation study evaluating ranks $r \in \{1, 2, 4\}$. For this analysis, we trained our models on modality-complete data and evaluated

Table A2.13: Performance comparison with different LoRA ranks (r). The models are trained on complete multimodal data and tested on both complete and missing modality scenarios. The best Accuracy (Acc.) for each scenario is highlighted in bold. Params (M) indicates the number of learnable parameters in millions.

Datasets	Available	Missing	$r = 1$		$r = 2$		$r = 4$	
			Params (M)	Acc. (%)	Params (M)	Acc. (%)	Params (M)	Acc. (%)
UPMC Food-101	Image	Text		75.66		75.62		70.38
	Text	Image	0.271	85.31	0.376	85.61	0.671	83.92
	Image - Text	-		94.47		94.64		93.30
AVE	Audio	Video		85.21		85.07		84.33
	Video	Audio	0.173	84.58	0.319	84.30	0.614	80.84
	Audio - Video	-		96.77		96.77		96.27
CREMA-D	Audio	Video		67.20		65.32		65.46
	Video	Audio	0.157	76.21	0.303	75.54	0.598	76.61
	Audio - Video	-		88.84		87.63		87.77
KS	Audio	Video		68.27		69.93		69.54
	Video	Audio	0.176	85.77	0.322	85.20	0.617	86.43
	Audio - Video	-		91.21		91.06		90.59

their performance in both complete and missing modality scenarios, with the results presented in Table A2.13.

Our analysis indicates that increasing the rank, and thereby the number of learnable parameters, does not guarantee a consistent improvement in performance. For instance, on the AVE dataset, a rank of $r = 1$ consistently achieves the highest accuracy. On the UPMC Food-101 dataset, performance between $r = 1$ and $r = 2$ is comparable. While $r = 2$ offers a marginal improvement in two scenarios, it requires a 38.7% increase in learnable parameters. In contrast, using $r = 4$ consistently resulted in degraded performance in both of these datasets compared to $r = 1$. A sim-

ilar trend is observed for the CREMA-D dataset, where $r = 1$ performs best in two of the three scenarios. Conversely, the KS dataset presented mixed results, with different ranks yielding optimal performance in different scenarios.

Overall, a rank of $r = 1$ demonstrates competitive, and often superior performance across a majority of the tested scenarios. The performance degradation at higher ranks, such as $r = 4$, suggest that a larger number of learnable parameters may not be necessary for these tasks and could introduce a risk of overfitting. This aligns with the common characteristic of LoRA, where a low rank is often sufficient to capture essential task-specific information. Based on these observations, we use LoRA with $r = 1$ for all our experiments as it provides an optimal balance, delivering robust performance with the greatest parameter efficiency.

A2.5 Different Modality Ratios During Training and Inference

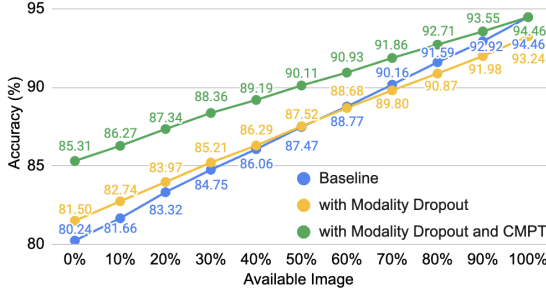
A robust multimodal learning method should maintain strong performance even when the missing modality ratio during inference differs from that during training. To evaluate the robustness of our approach, we follow the experimental setup of [81, 89] and train and test our model with varying modality ratios. Let η denote the missing rate. We consider three scenarios: missing-text, missing-image, and missing-both. In the first two cases, $\eta\%$ of the samples contain only images or only text, while $(1 - \eta)\%$ retain both modalities. For the missing-both case, $\frac{\eta}{2}\%$ of the samples contain only texts, $\frac{\eta}{2}\%$ contain only images, and $(1 - \eta)\%$ include both modalities. We set $\eta = 70$ throughout our experiments, following existing baselines [81, 89].

As summarized in Table A2.14, our method consistently outperforms existing state-of-the-art approaches in most scenarios. While the method proposed by [89] performs well in certain

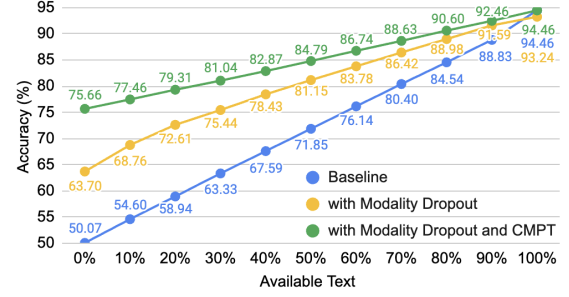
Table A2.14: Performance comparison on MM-IMDb and UPMC Food-101 datasets when modality gets missing during both training and inference following setup from existing baselines [81, 89].

Datasets	Missing	Train		Inference		ViLT	MAP	MSP	Kim & Kim	CMPT (Ours)
	Rate (η)	Image	Text	Image	Text	[90]	[94]	[81]	[89]	(Mean $\pm$ SD)
MM-IMDb (F1-Macro)		30%	100%	30%	100%	37.63	47.18	47.45	<u>56.03</u>	60.21 $\pm$ 0.40
				65%	65%	34.47	37.25	42.17	<u>49.81</u>	50.05 $\pm$ 0.53
				100%	30%	30.00	24.31	34.32	43.07	<u>36.10 $\pm$ 0.19</u>
	70%	65%	65%	30%	100%	36.48	42.72	44.81	<u>55.26</u>	57.45 $\pm$ 0.21
				65%	65%	36.54	40.84	42.03	<u>49.24</u>	54.04 $\pm$ 0.23
				100%	30%	33.76	37.80	37.30	<u>42.46</u>	50.05 $\pm$ 0.29
		100%	30%	30%	100%	27.25	20.17	34.11	54.67	<u>44.72 $\pm$ 0.46</u>
				65%	65%	32.51	29.79	36.97	<u>49.09</u>	49.26 $\pm$ 0.42
				100%	30%	34.26	36.89	38.34	<u>43.21</u>	52.61 $\pm$ 0.12
UPMC Food-101 (Accuracy)		30%	100%	30%	100%	76.40	86.34	86.34	<u>87.11</u>	87.48 $\pm$ 0.13
				65%	65%	59.03	57.92	71.88	81.22	<u>77.52 $\pm$ 0.29</u>
				100%	30%	41.94	29.71	56.87	75.41	<u>67.60 $\pm$ 0.26</u>
	70%	65%	65%	30%	100%	73.62	85.55	85.91	<u>87.35</u>	87.36 $\pm$ 0.20
				65%	65%	68.60	78.49	78.89	<u>82.67</u>	82.83 $\pm$ 0.25
				100%	30%	64.25	71.17	71.58	<u>78.13</u>	78.40 $\pm$ 0.09
		100%	30%	30%	100%	43.67	27.46	57.22	86.90	<u>85.34 $\pm$ 0.15</u>
				65%	65%	54.75	50.79	65.13	<u>82.35</u>	82.78 $\pm$ 0.22
				100%	30%	66.01	73.71	73.77	<u>78.81</u>	80.37 $\pm$ 0.18

cases, our approach demonstrates superior performance across both the MM-IMDb and UPMC Food-101 datasets. To ensure a rigorous evaluation, we report the mean and standard deviation (SD) over three independent runs with different seeds. The low SD values further indicate the stability



(a) Available: Text, Missing: Image.



(b) Available: Image, Missing: Text.

Figure A2.4: Effectiveness of CMPTs in handling missing modalities on UPMC Food-101 dataset. CMPTs significantly improves performance when modalities are severely missing. Models are trained on 100% image + 100% text and tested with varying levels of missing modalities.

and reliability of our method. These results substantiate our claim that CMPTs can effectively estimate missing modality features, enhancing robustness to varying missing modality conditions. Other methods reported in Table 4.1 do not report results for this specific experimental setup, and thus we are unable to compare with them.

A2.6 Effectiveness of CMPTs

We extend the study from Section 4.4.6 to evaluate the effectiveness of CMPTs in handling missing modalities on the UPMC Food-101 dataset. We compare it against two baselines: (i) a standard model trained without modality dropout or CMPTs (baseline) and (ii) a model trained with modality dropout, where modalities are randomly removed during training to improve robustness to missing data. For a fair comparison, all models were fine-tuned on the downstream task using LoRA with a rank of 1.

As illustrated in Figure A2.4, the baseline model experiences a significant performance drop as the proportion of missing modalities increases. While modality dropout improves robust-

ness, it fails to fully recover the lost performance. Incorporating CMPTs further enhances performance, achieving a 5.07% improvement over the baseline and 3.81% over modality dropout when the image modality is entirely missing as shown in Figure A2.4a. A similar trend is observed when text is missing as shown in Figure A2.4b, where CMPTs achieves 25.59% and 11.96% higher accuracy compared to the baseline and modality dropout, respectively. Consistent with our previous findings, CMPTs becomes increasingly effective as the proportion of missing data increases, demonstrating its strong capability to substantially mitigate performance degradation in the presence of severe modality loss.

A2.7 Do CMPTs Benefit All Classes?

To further evaluate CMPT’s effectiveness, we conduct a per-class performance analysis on the UPMC Food-101 dataset, following the approach from Section 4.4.6. As shown in Figure A2.5 and A2.6, CMPTs improve or maintain performance across most of the classes in both text-missing and image-missing scenarios. Notably, it mitigates performance degradation in modality-sensitive classes. For instance, when text is missing, classes such as *guacamole*, *pancakes*, and *foie_gras* experience severe drops in performance, but CMPTs effectively recovers much of the loss as shown in Figure A2.4b. Similarly, in image-missing scenarios, CMPTs provide substantial gains for *sushi*, *cup_cakes*, and *pizza* as shown in Figure 4.3a. However, a slight performance drop is observed in certain classes when either text or image is absent. For instance, CMPT fails to recover performance for classes such as *bibimbap* and *sashimi* in the absence of text, and for classes like *lobster_roll_sandwich* and *cheese_plate* when the image modality is missing. This is likely due to insufficient

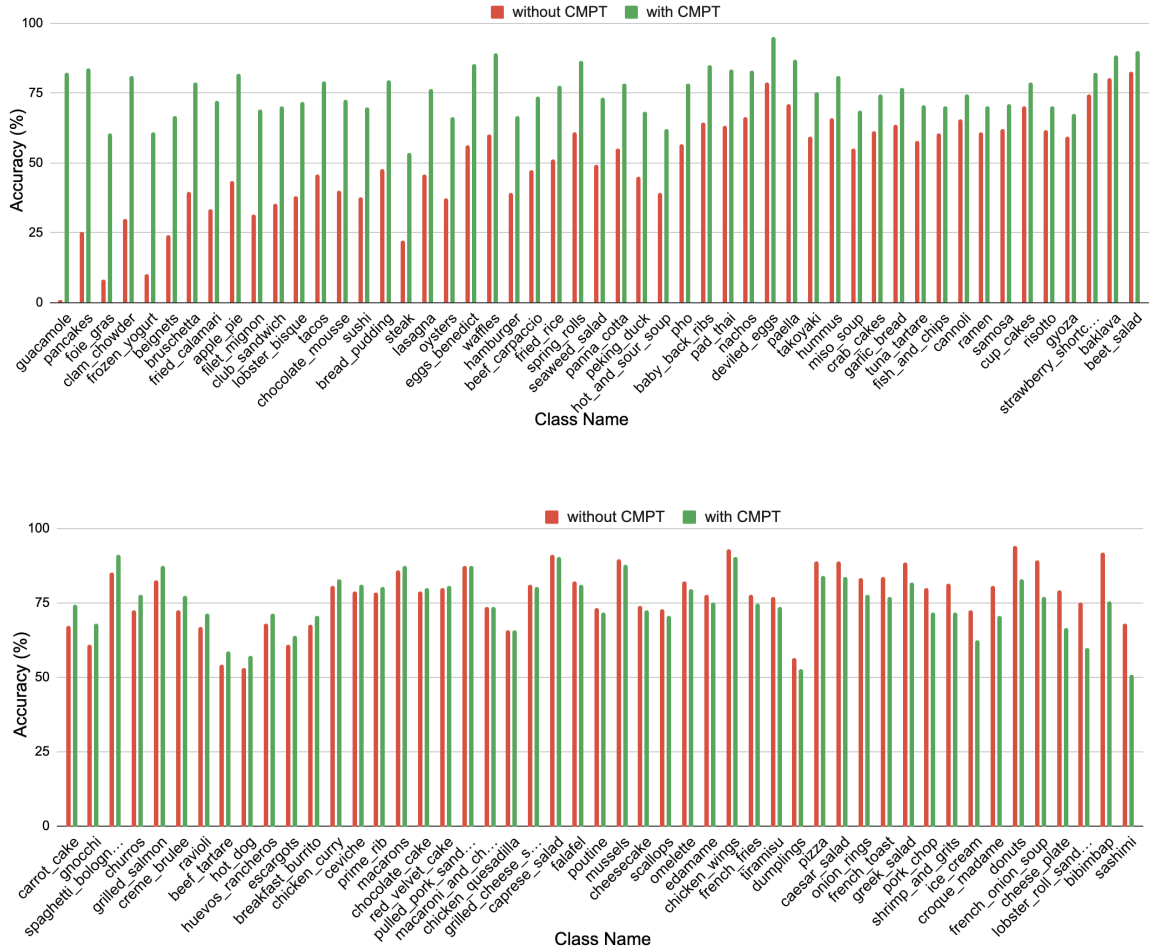


Figure A2.5: CMPTs improve performance across most of the classes when *text* modality is missing on UPMC Food-101 dataset. However, it fails to accurately predict certain classes (e.g., bibimbap/sashimi), likely due to the image modality’s insufficient information for those classes. Classes are sorted based on the amount of performance improvement.

information in the available modality alone for those classes. Overall, these results reinforce our previous findings, demonstrating that CMPTs enhance performance across most classes, leading to a strong overall performance improvement.

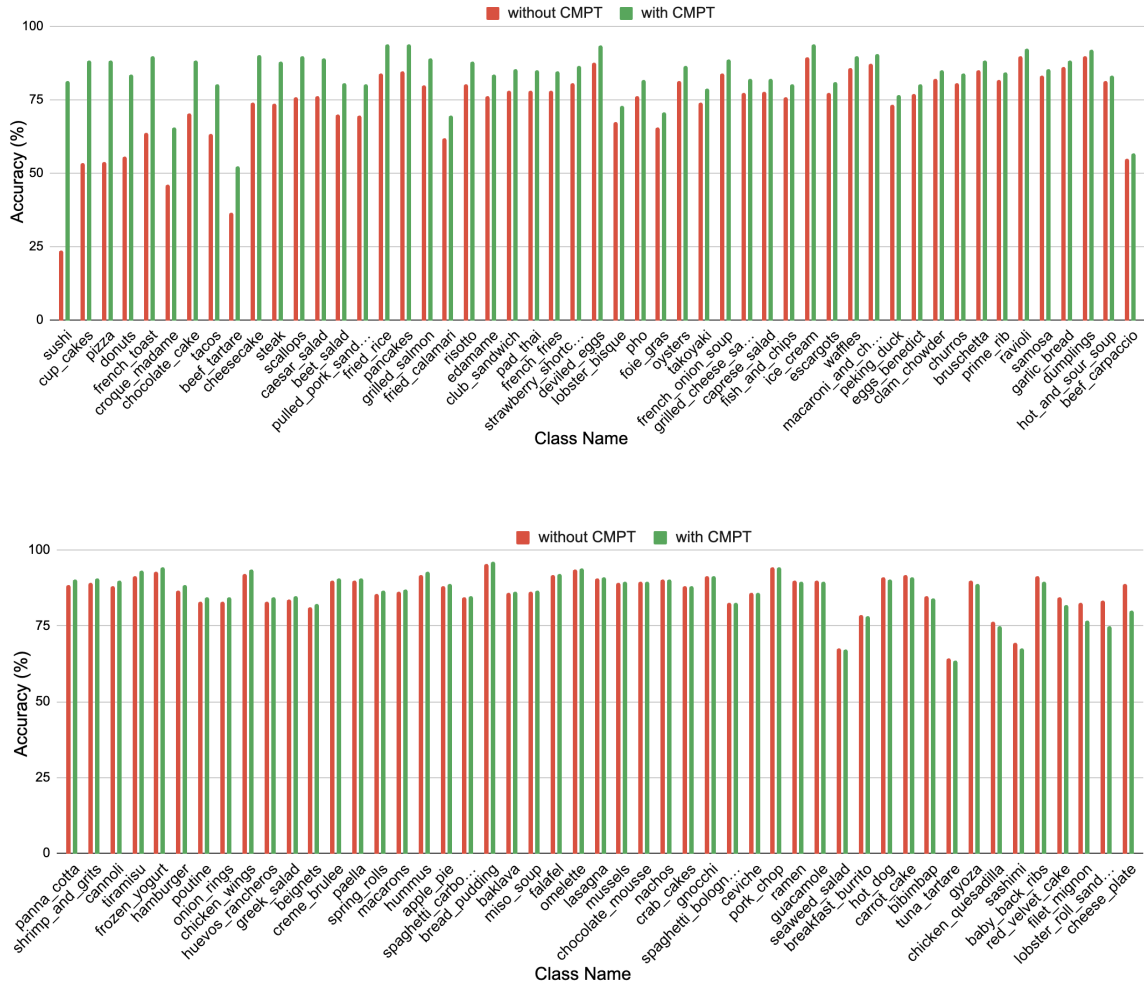


Figure A2.6: CMPTs improve performance across most of the classes when *image* modality is missing on UPMC Food-101 dataset. However, it fails to accurately predict certain classes (e.g., lobster_roll_sandwich/cheese_plate), likely due to the text modality’s insufficient information for those classes. Classes are sorted based on the amount of performance improvement.

A2.8 Limitations and Failure Case Analysis

While CMPTs demonstrate notable improvements in robustness, it is essential to understand their limitations. Figure A2.7 presents t-SNE plots of the fused feature tokens $\mathcal{T}$ for ten classes on the UPMC Food-101 dataset where the performance gains from CMPTs are limited. In these cases, the fused feature embeddings with CMPTs (blue) remain diffuse and overlapping, and

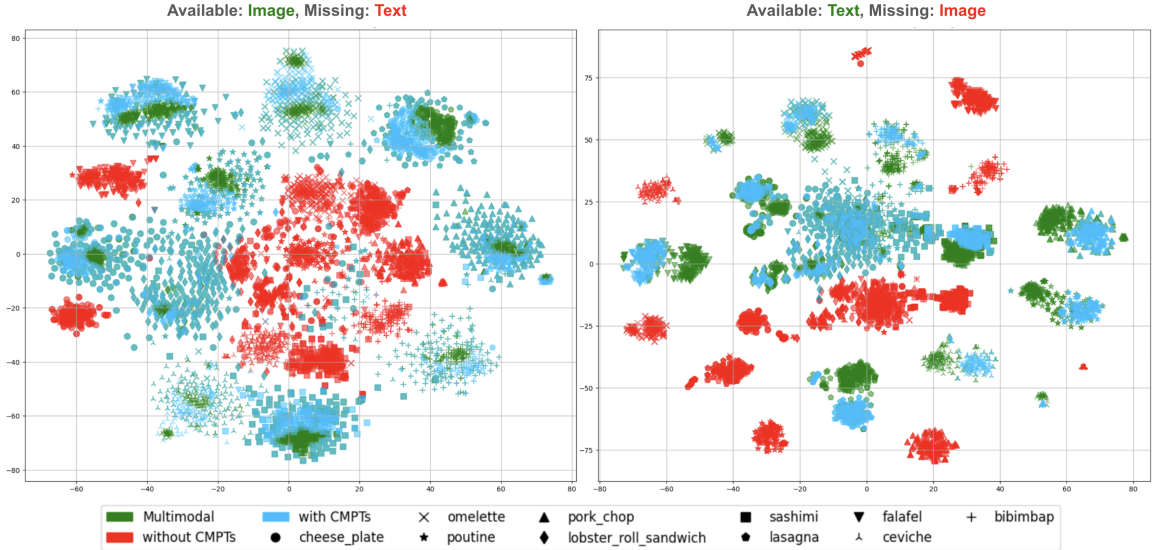


Figure A2.7: t-SNE plots of the fused feature tokens $\mathcal{T}$ for ten classes where the performance gains from CMPTs are limited. In these cases, the fused feature embeddings with CMPTs (blue) remain diffuse and overlapping, and exhibit weak alignment with the complete multimodal features (green). Although CMPTs show improvement over the baseline without CMPTs (red), the alignment remains insufficient to form consistently discriminative clusters, indicating limited or no improvement in missing modality robustness in these classes.

exhibit weak alignment with the complete multimodal features (green). This is particularly evident for visually ambiguous classes or classes with high intra-class variance in the feature embedding space. Although CMPTs show improvement over the baseline without CMPTs (red), the alignment remains insufficient to form consistently discriminative clusters, indicating limited or no improvement in missing modality robustness.

We also analyze the attention maps from the last layer of the vision encoder for the CLS token and CMPTs in the text-missing scenario, as shown in Figure A2.8. The left half of the figure includes examples when the fused features with and without CMPTs provide correct predictions (approximately 60% samples show this pattern). For these cases, the CLS token already attends to semantically-relevant regions and make the correct prediction. CMPTs further refine the attention

but do not change the model’s prediction, as it is already correct. Thus, the inclusion of CMPTs does not necessarily improve performance. The right half of the figure shows examples when the fused features with CMPTs lead to incorrect predictions (approximately 3% samples show this pattern). We observe that CMPTs attend to semantically-relevant regions of the image but still lead to incorrect predictions. For example, in the *French Onion Soup* sample, the CMPT focuses heavily on the melted cheese surface—a visually salient but ambiguous feature—leading to a misclassification as *Shrimp and Grits*. Similarly, for the *Lobster Roll Sandwich*, the CMPT fails to recover from the original misclassification and predicts *Foie Gras* based on misleading visual cues. These examples highlight scenarios where either the available modality is insufficient or the learned cross-modal associations are misleading, resulting in CMPTs contributing to incorrect predictions rather than correcting them.

A2.9 Analysis of Modality Dependence

We observe that the impact of missing modalities varies significantly across classes. Certain classes show strong dependence on a specific modality. In this experiment, we train model on modality-complete data and evaluate it on both modality-complete and incomplete test samples.

Figure A2.9 presents per-class performance on the MM-IMDb dataset, revealing heterogeneous effects of missing modalities. For instance, the *Drama* and *Thriller* classes maintain consistent performance regardless of missing modalities. In contrast, the *Biography* and *Documentary* classes experience significant decline in performance when text is missing, even if images are available. Interestingly, certain classes like *Animation* show slightly higher performance with single-

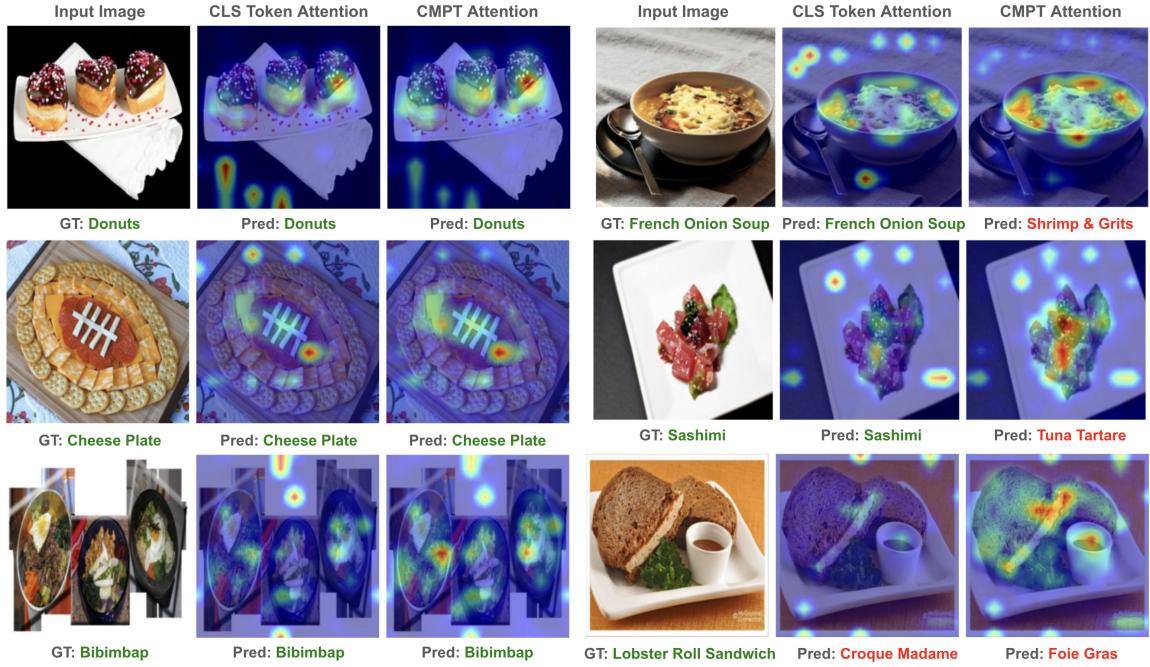


Figure A2.8: Attention map visualizations from the last layer of the vision encoder for the CLS token and CMPTs in the text-missing scenario on the UPMC Food-101 dataset. **Left:** Cases where the fused features with and without CMPTs provide correct predictions. Here, CLS tokens already attend to semantically-relevant regions and CMPTs provide further refinement, but do not change the prediction. **Right:** Cases where CMPTs attend to semantically-relevant regions but lead to incorrect predictions. The last example illustrates a scenario where both the fused features with and without CMPTs result in incorrect predictions. These examples highlight limitations of CMPTs when the available modality lacks sufficient discriminative information or when learned cross-modal associations are imprecise.

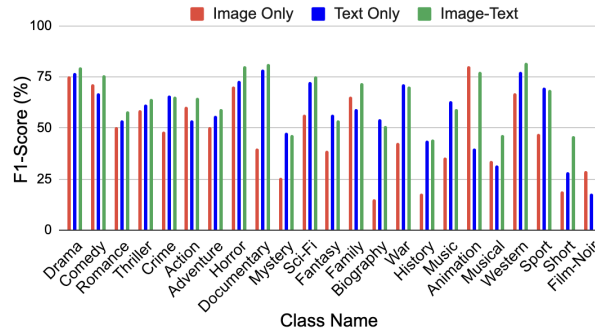


Figure A2.9: Per class performance analysis on MM-IMDb dataset. Certain classes show strong dependence on a specific modality.

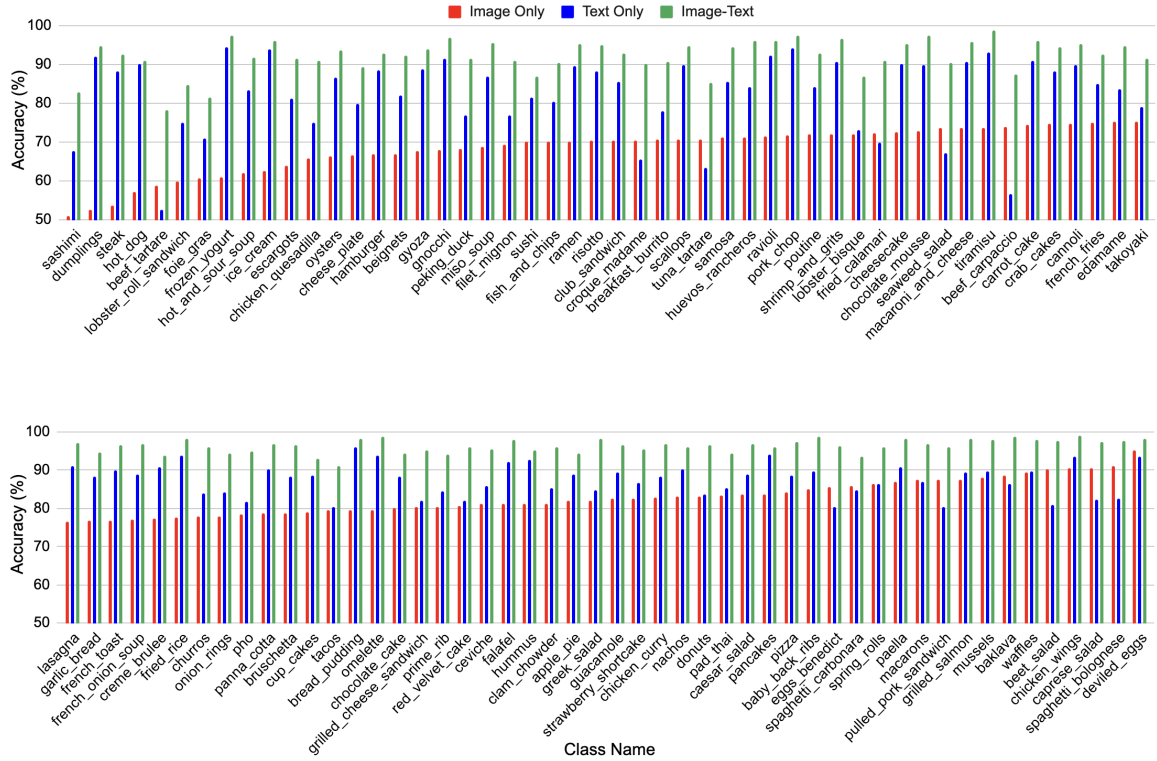


Figure A2.10: Per class performance analysis on UPMC Food-101 dataset. Certain classes show strong dependence on a specific modality.

modality inputs, particularly in image-only scenario. Similar patterns emerge in *Sport* and *Music* categories, where text-only samples performs slightly better than complete modality. This discrepancy may stem from the simple sum-fusion strategy we use, which might not effectively model complex modality interactions. These findings indicate that the effect of missing modalities is not uniform across classes, with some being more modality-dependent than others.

Extending this analysis to UPMC Food-101, we observe that multimodal (image-text) data consistently outperforms unimodal inputs across all classes as shown in Figure A2.10. However, the reliance on specific modalities varies widely among food categories. Classes such as *dumplings*, *Frozen_yogurt*, and *Ice cream* depend heavily on text, showing strong performance with text-only inputs but experiencing significant drops when text is missing. Conversely, categories like

Table A2.15: Performance comparison with three different parameter efficient adaptation methods. We used rank=1 for all the methods. The best Accuracy (Acc.) for each scenario is highlighted in **bold**. Params (M) indicates the number of learnable parameters in millions.

Datasets	Available	Missing	LoRA		DoRA		VeRA	
			Params (M)	Acc. (%)	Params (M)	Acc. (%)	Params (M)	Acc. (%)
UPMC Food-101	Image	Text		75.66		69.03		70.82
	Text	Image	0.271	85.31	0.302	83.52	0.155	82.85
	Image - Text	-		94.47		92.59		92.34
AVE	Audio	Video		85.21		86.57		89.05
	Video	Audio	0.173	84.58	0.246	91.79	0.098	90.05
	Audio - Video	-		96.77		96.27		96.27
CREMA-D	Audio	Video		67.20		67.07		66.13
	Video	Audio	0.157	76.21	0.229	73.12	0.082	71.64
	Audio - Video	-		88.84		88.04		86.02
KS	Audio	Video		68.27		68.77		69.93
	Video	Audio	0.176	85.77	0.248	85.97	0.101	83.69
	Audio - Video	-		91.21		91.32		90.71

tuna_tartare, *beef_carpaccio*, and *beet_salad* rely more on image information, with performance deteriorating notably when images are absent. This observation suggests that some classes rely more on texts, while others depend heavily on images for accurate classification.

Interestingly, some classes such as *beef_tartare*, *chicken_quesadilla*, and *croque_madame* exhibit poor performance with unimodal inputs but improve significantly with both image and text, reinforcing the importance of multimodal data. Performance on these classes improve significantly when we utilize both image and text. Overall, these findings confirm that while multimodal information universally enhances performance, the degree of modality dependence remains heterogeneous across classes.

A2.10 Generalizability to Other Parameter-Efficient Adaptation Methods

To demonstrate that CMPTs are not restricted to a particular parameter-efficient adaptation method, we extend our analysis beyond LoRA [72], which is used in our primary experiments. Specifically, we evaluate CMPTs with two recent alternatives: DoRA [115] and VeRA [92]. For a fair comparison, we fix the rank to 1 across all methods. We train the models on modality-complete data and evaluate them on both modality-complete and modality-incomplete scenarios. The results are summarized in Table A2.15.

We find that the number of learnable parameters differs across methods, with VeRA being the most parameter-efficient and DoRA the least. In terms of performance, no single method consistently outperforms the others across all datasets. LoRA achieves the best overall results on the UPMC Food-101 and CREMA-D datasets, while DoRA shows superior performance on the Kinetics-Sound dataset and remains highly competitive on AVE. VeRA, while using substantially fewer parameters, delivers strong performance but is generally slightly below LoRA and DoRA, suggesting a trade-off between parameter efficiency and maximal performance.

Most importantly, our framework maintains robust performance in missing-modality scenarios regardless of the underlying adaptation method. The relative stability observed between complete and missing-modality settings is consistent across all three adaptation techniques. These findings confirm that CMPTs function as a modular approach whose effectiveness is not tied to any specific adaptation strategy, underscoring their scalability and versatility across different parameter-efficient methods.

Table A2.16: Performance comparison of our method with a ViT versus a SigLIP model as the vision backbone. The models are trained on complete multimodal data and tested on both complete and missing modality scenarios. We report % accuracy as the performance metric, and the highest score for each scenario is highlighted in **bold**.

Datasets	Available Modality	Missing Modality	ViT [34]	SigLIP [211]
UPMC Food-101	Image	Text	75.66	77.50
	Text	Image	85.31	85.58
	Image - Text	-	94.47	94.58
AVE	Audio	Video	85.21	86.32
	Video	Audio	84.58	30.60
	Audio - Video	-	96.77	94.78
CREMA-D	Audio	Video	67.20	65.73
	Video	Audio	76.21	35.35
	Audio - Video	-	88.84	84.68
KS	Audio	Video	68.27	69.66
	Video	Audio	85.77	30.76
	Audio - Video	-	91.21	89.44

A2.11 Applicability to SigLIP-Based Models

To assess the architectural flexibility of our method, we replaced the ViT encoder [34] with a pretrained SigLIP model [211], which employs Global Attention Pooling (GAP) instead of a CLS token. For audio–video datasets, we retained the AST model as the audio encoder, and for image–text datasets, we used BERT as the text encoder. In this setup, CMPT tokens must perform a cross-architectural reconstruction: inferring missing audio or text features from SigLIP’s vision embedding, and vice versa.

We train the models on modality-complete data and evaluate them on both modality-complete and modality-incomplete scenarios. The results in Table A2.16 show that our method is

not restricted to a single encoder type. On UPMC Food-101 dataset, the SigLIP-based model consistently outperforms the ViT-based model, demonstrating that CMPTs can be effectively applied to GAP-based vision backbones for vision–language tasks. In contrast, for audio–video datasets, SigLIP exhibits sharp performance degradation when audio is missing. Here, CMPTs struggle to recover audio features from global video embeddings, likely because GAP discards the fine-grained temporal information that is critical for inferring a corresponding audio representation. The reverse process, however, is more successful. When the video modality is missing, AST’s audio features provide a strong basis for reconstructing SigLIP’s global video representation. This suggests that AST’s feature space is more compatible with SigLIP than vice versa.

Overall, these findings indicate that CMPTs remain applicable to GAP-based models like SigLIP, but their effectiveness depends on the modality. While GAP facilitates efficient vision embeddings, it limits the representational richness required for audio reconstruction, constraining CMPTs in certain cases. Nonetheless, the results confirm that CMPTs are not inherently incompatible with SigLIP, though their utility is shaped by architectural and modality-specific factors.

A2.12 Scalability with Larger Vision and Text Encoders

To assess the scalability of our framework with larger encoders, we replaced the ViT-Base and BERT-Base encoders used in our primary experiments with ViT-Large and BERT-Large. We train the models on modality-complete data and evaluate them on both modality-complete and modality-incomplete scenarios on the UPMC Food-101 and MM-IMDb datasets.

The results, summarized in Table [A2.17](#), show that our framework can leverage the increased capacity of larger backbones. On MM-IMDb, scaling yields consistent improvements in

Table A2.17: Performance comparison of our method using *Base* versus *Large* ViT and BERT models. The models are trained on complete multimodal data and tested on both complete and missing modality scenarios. We report % accuracy for UPMC Food-101 and F1-Macro score for MM-IMDb dataset. Params (M) indicates the number of learnable parameters in millions. The best performance for each scenario is highlighted in **bold**.

Datasets	Available	Missing	ViT-Base + BERT-Base		ViT-Large + BERT-Large	
	Modality	Modality	Params (M)	Performance	Params (M)	Performance
MM-IMDb (F1-Macro)	Image	Text		46.97		53.42
	Text	Image	0.211	56.32	0.540	59.92
	Image - Text	-		63.58		64.00
UPMC Food-101 (Accuracy)	Image	Text		75.66		77.17
	Text	Image	0.271	85.31	0.501	83.91
	Image - Text	-		94.47		93.74

F1-Macro across all scenarios. The gains are particularly notable in missing-modality cases, with a +6.45% increase when text modality is missing and +3.60% when image modality is missing. These results suggest that stronger backbone representations improve CMPT’s ability to reconstruct missing features and maintain robust predictions.

On UPMC Food-101, the larger backbone improves image-only accuracy by +1.51% but shows a decline of -1.41% for text-only performance. With both modalities available, performance remains comparable (94.47 vs. 93.74). This indicates a potential trade-off, where the added capacity may overfit to the text modality, reducing generalization when the image is absent.

In summary, CMPT scales effectively with larger backbones, particularly on complex datasets such as MM-IMDb. At the same time, the results highlight that the benefits of scaling depend on dataset and modality characteristics, underscoring the need for careful backbone selection to balance capacity and generalization.

A3 SSAM: Singular Subspace Alignment for Merging Multimodal Large Language Models

A3.1 Datasets and Benchmarks

We evaluate the performance of our merged model on four multimodal datasets from two representative benchmarks. In addition, we also assess performance on three domain-specific benchmarks to verify that merging preserves the capabilities of the specialist models. Detailed descriptions of the datasets and benchmarks are provided below.

MUSIC-AVQA. The MUSIC-AVQA dataset [96] is a large-scale audio–visual question answering benchmark constructed from musical performance videos. It contains 45,867 question–answer pairs derived from 9,288 videos, totaling over 150 hours of content. The dataset emphasizes fine-grained multimodal understanding and spatio-temporal reasoning, covering 9 question types and 33 question templates. Each question requires joint reasoning over different combinations of image, audio, and video modalities paired with a language query.

AVQA. The AVQA dataset [201] is a large-scale audio–visual question answering benchmark designed to evaluate multimodal understanding of real-world objects, events, and activities. It comprises 57,015 videos depicting diverse daily scenarios and 57,335 carefully constructed question–answer pairs. The questions are designed to require joint reasoning over image, audio, and video modalities, capturing temporal dependencies, causal interactions, and cross-modal relationships. Each question may rely on different combinations of the three input modalities paired with a language query.

Multimodal Commonality Understanding Benchmark (MCUB). The MCUB benchmark [14] evaluates the ability of a model to infer shared semantic concepts from heterogeneous multimodal inputs. Each sample includes inputs from multiple modalities (image, audio, video, and point cloud) along with a textual query, and the model must select the correct answer from four candidate answers. The benchmark assesses both cross-modal alignment and higher-level conceptual reasoning, making it a challenging benchmark for unified multimodal understanding. MCUB offers two evaluation datasets. **MCUB-3** contains four subsets, each comprising three of the four input modalities paired with language: image–video–point cloud–language, audio–video–point cloud–language, image–audio–point cloud–language, and image–audio–video–language. **MCUB-4** includes all four input modalities (image, audio, video, and point cloud) paired with language. Together, these settings enable a comprehensive assessment of a model’s ability to generalize across diverse and unseen modality combinations. Additional dataset details, preprocessing steps, and annotation procedures are available in [14].

Domain-Specific Benchmarks. To assess whether the merged model preserves the capabilities of the original specialist models, we evaluate it on three domain-specific benchmarks covering language only, vision–language, and audio–language tasks. Specifically, we use MMLU [68], a widely used benchmark for evaluating general language understanding across diverse academic and professional subjects; OCRBench [116], which measures vision–language reasoning with a focus on optical character recognition and text understanding in images; and MMAU [153], a benchmark designed to evaluate multimodal audio–language reasoning and understanding. We report performance on the MMAU (test-mini) subset. These evaluations verify whether the merged model retains

the domain expertise of the original specialist models. Our results show that merging preserves, and in some cases slightly improves, performance compared to the original specialist models.

A3.2 Implementation Details

We use the pretrained specialist models released by [14], which include checkpoints for vision–language, audio–language, video–language, and point cloud–language models. Each specialist model is composed of a modality-specific encoder, a projection module, and a Vicuna-7B-v1.5 [228] language decoder. Below, we summarize the encoder and projector architectures for each modality:

- **Vision-Language Model:** CLIP-ViT-L-336px [144] is used as the vision encoder, followed by a two-layer MLP projector to align image embeddings to the language embedding space.
- **Audio-Language Model:** BEATs-Iter3+ [17] is used as the audio encoder and a Q-Former [98] with 32 learnable query tokens serves as the projection module.
- **Video-Language Model:** The video encoder from LanguageBind [238] is used, coupled with an MLP projector similar to the vision projector.
- **Point Cloud-Language Model:** The point cloud encoder from PointLLM [198] is used with a corresponding MLP projector to align point cloud embeddings to the language embedding space.

All specialist models follow a two-stage training procedure as described in [14]. In stage 1, only the projectors are trained while keeping the modality-specific encoders and the language decoder completely frozen. In stage 2, the language decoder is fine-tuned using LoRA [72] with

Algorithm 1 SSAM Algorithm

Input: Pretrained model weight W_0 ;

Language-specific weights from specialist models $\{W_i^t\}_{i=1}^n$;

Rank k ; Number of layers L ;

Scaling coefficient λ .

Output: Merged language-specific weight W_{merged}^t

```
1:  $W_{\text{merged}}^t \leftarrow \{\}$ 
2: for  $l = 1$  to  $L$  do
  ▷ Step 1: Calculate language vectors for layer  $l$  ◁
3:   for  $i = 1$  to  $n$  do
4:      $\Delta_{i,l} \leftarrow W_{i,l}^t - W_{0,l}$ 

  ▷ Step 2: Compute shared subspaces for layer  $l$  ◁
5:    $A_l \leftarrow \sum_{i=1}^n \Delta_{i,l} \Delta_{i,l}^\top$ 
6:    $B_l \leftarrow \sum_{i=1}^n \Delta_{i,l}^\top \Delta_{i,l}$ 
7:    $U_l, \Sigma_{A,l}, U_l^\top \leftarrow \text{SVD}(A_l)$ 
8:    $U_{c,l} \leftarrow U_{l,[:,1:k]}$ 
9:    $V_l, \Sigma_{B,l}, V_l^\top \leftarrow \text{SVD}(B_l)$ 
10:   $V_{c,l} \leftarrow V_{l,[:,1:k]}$ 

  ▷ Step 3: Subspace projection and merging ◁
11:   $P_{u,l} \leftarrow U_{c,l} U_{c,l}^\top$ 
12:   $P_{v,l} \leftarrow V_{c,l} V_{c,l}^\top$ 
13:  for  $i = 1$  to  $n$  do
14:     $\Delta_{i,l}^p \leftarrow P_{u,l} \Delta_{i,l} P_{v,l}$ 
15:   $\Delta_{\text{merged},l} \leftarrow \lambda \sum_{i=1}^n \Delta_{i,l}^p$ 

  ▷ Step 4: Final merged language-specific weight for layer  $l$  ◁
16:   $W_{\text{merged}}^t(l) \leftarrow W_{0,l} + \Delta_{\text{merged},l}$ 
17: return  $W_{\text{merged}}^t$ 
```

rank 128 and $\alpha = 256$, while keeping all modality-specific encoders and projectors frozen. We use the publicly released checkpoints from [14] directly for all the experiments without any additional fine-tuning or modification. All merging is performed in a training-free setting.

Table A3.18: Performance of SSAM with different subspace ranks (k) on the MCUB-3 and MCUB-4 datasets. I, A, V, P, and L denote Image, Audio, Video, Point Cloud, and Language, respectively. Rank $k = 128$ empirically yields overall better performance.

Subspace	MCUB-3					MCUB-4
Rank (k)	I+V+P+L	A+V+P+L	I+A+P+L	I+A+V+L	Average	I+A+V+P+L
64	58.20	61.40	65.00	56.40	60.25	62.34
128	59.40	61.00	65.40	55.60	60.35	62.27
256	59.20	59.20	64.80	56.20	59.85	61.66
384	58.40	57.80	64.40	54.40	58.75	61.47
512	58.60	58.20	65.20	55.40	59.35	61.52

For evaluating baseline merging methods, Task Arithmetic [78], TSV [43], WUDI [22], and OptMerge [182], we apply their public implementations to the same specialist checkpoints to ensure a fair comparison. Results for other baselines are taken from their respective publications.

A3.3 Proposed SSAM Algorithm

The key idea of SSAM is threefold: (i) compute the *language vector* Δ_i for each specialist language decoder W_i , representing its fine-tuning direction in parameter space; (ii) identify a shared low-rank subspace that captures dominant and mutually consistent update directions across all language vectors; and (iii) project and merge these language vectors within this subspace to form the final language-specific decoder weight W_{merged}^t . This process integrates shared knowledge from all specialist models while minimizing parameter interference. As mentioned in Section 5.3, these language vectors are computed independently for each linear layer l of the model. The pseudocode for SSAM is shown in Algorithm 1.

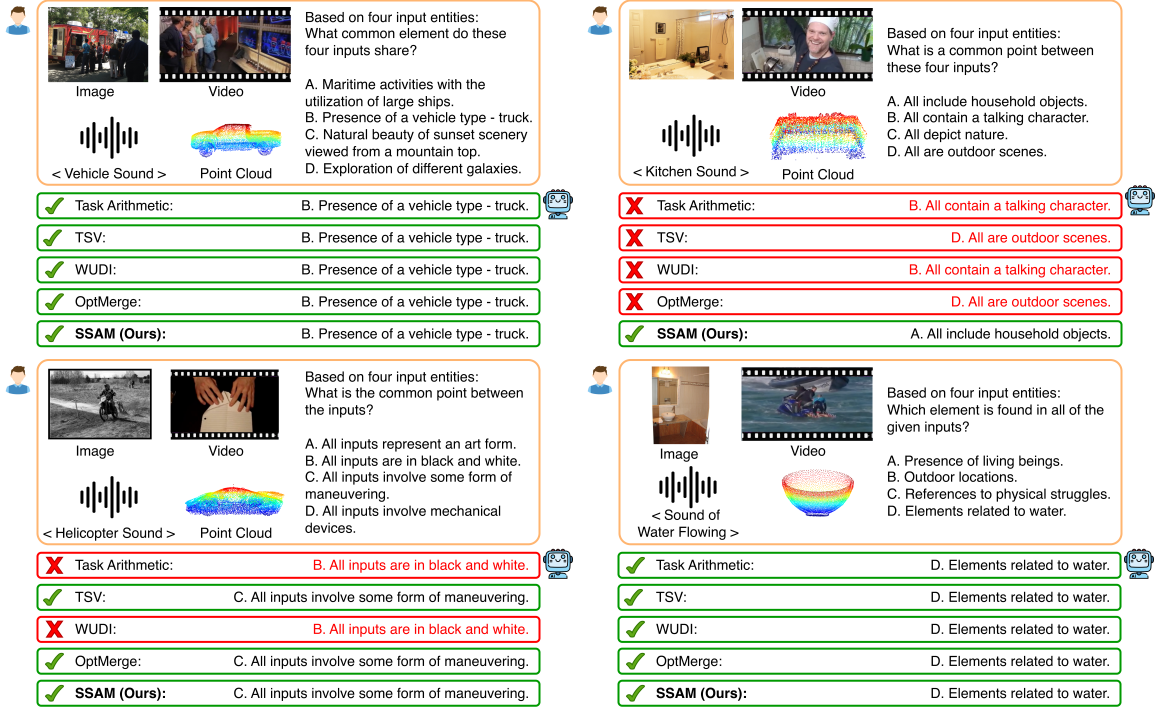


Figure A3.11: Qualitative results on the MCUB-4 [14] dataset with four input modalities (image, video, audio, and point cloud) paired with natural-language queries. SSAM demonstrates strong cross-modal reasoning and provides coherent answers in complex multimodal settings, even though none of the specialist models were trained on this modality combination.

A3.4 Selecting Optimal Subspace Rank (k)

We evaluate the performance of the proposed method, SSAM, with subspace ranks $k \in \{64, 128, 256, 384, 512\}$ on the MCUB-3 and MCUB-4 datasets to determine the optimal subspace dimensionality. In this section, we provide a detailed performance analysis for each modality combination. As shown in Table A3.18, both $k = 64$ and $k = 128$ achieve similar performance, with $k = 128$ performing marginally better. Increasing the rank beyond this point gradually decreases the performance of the merged model. We hypothesize that higher ranks introduce redundant or conflicting updates within the language vectors, leading to parameter interference. Therefore, we use $k = 128$ for all experiments.

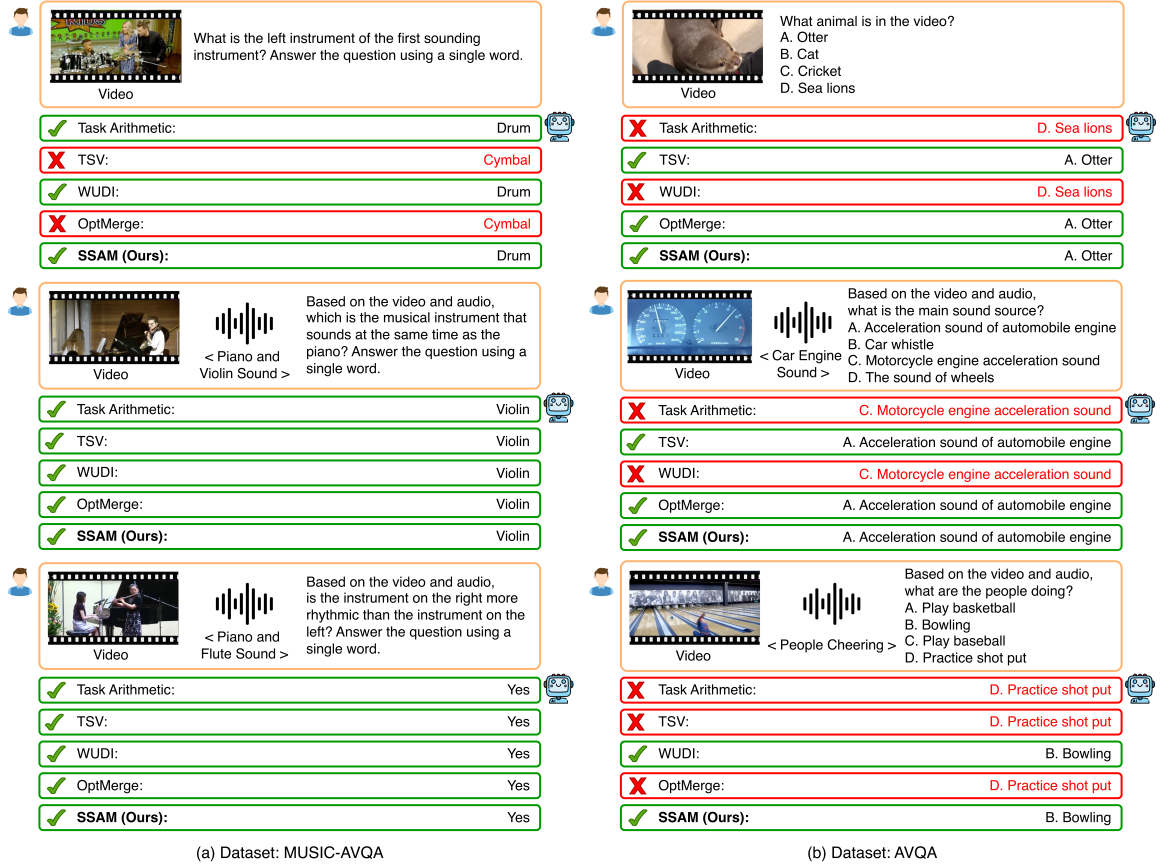


Figure A3.12: Qualitative results on the MUSIC-AVQA [96] and AVQA [201] datasets with video and audio as the input modalities paired with natural-language queries. SSAM provides consistent responses, demonstrating stable performance on multimodal queries despite the specialist models not being trained on paired video-audio-language data.

A3.5 Qualitative Results

We present additional qualitative examples from the MCUB-4 [14] dataset in Figure A3.11 and from the MUSIC-AVQA [96] and AVQA [201] datasets in Figure A3.12. Each example includes outputs from SSAM alongside baseline merging methods: Task Arithmetic [78], TSV [43], WUDI [22], and OptMerge [182]. Across these datasets, SSAM generates coherent responses across heterogeneous input modality combinations, while other baseline merging methods often fail under unseen modality combinations. These examples illustrate that SSAM effectively integrates comple-

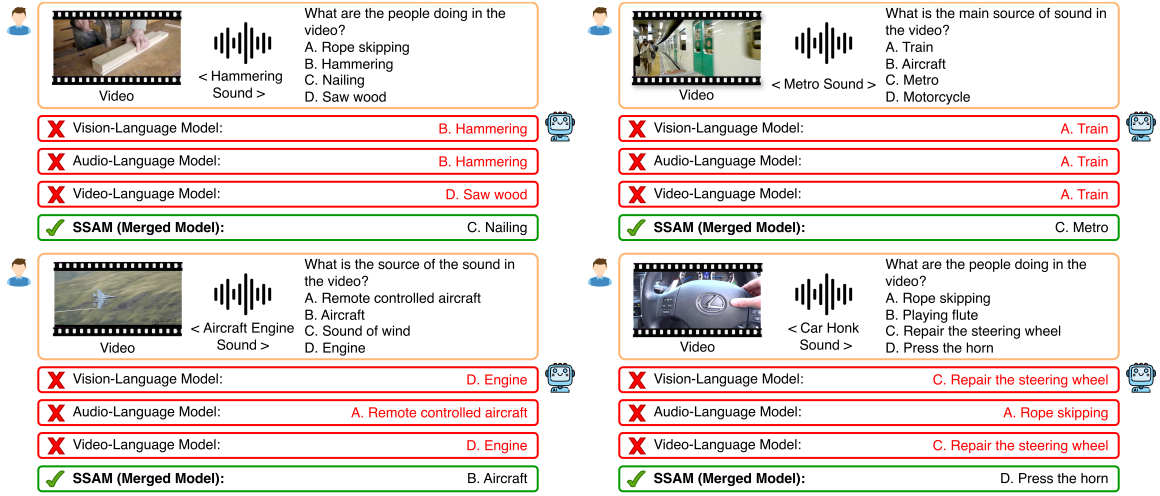


Figure A3.13: Qualitative results on the AVQA [201] dataset with video and audio as input modalities paired with natural-language queries. SSAM effectively combines information from both audio and video to produce more accurate and consistent responses, demonstrating stable performance on multimodal queries compared to individual specialist models.

mentary information from image, audio, video, and point cloud modalities. Notably, SSAM generalizes to unseen modality combinations (e.g., audio–video–language, image–audio–video–point cloud–language), despite none of the specialist models being trained to handle such combinations jointly.

Qualitative examples on the AVQA [201] dataset in Figure A3.13 demonstrate the advantage of SSAM in multimodal reasoning scenarios requiring joint audio–video understanding. We use the first frame from each video as the image modality for the vision–language model. In these examples, individual specialist models (vision–language, audio–language, and video–language) produce incorrect predictions when relying on a single modality. In contrast, the merged model (SSAM) integrates complementary cues from both audio and video to infer the correct answer. For instance, SSAM correctly identifies actions such as *nailing* by combining the visual scene with the characteristic hammering sound, recognizes a *metro* from the combination of subway visuals and

audio, distinguishes an *aircraft* by jointly reasoning about the visual appearance and engine sound, and correctly identifies a *car horn* from the steering-wheel interaction in video with the honking audio. These examples highlight SSAM’s ability to capture information across modalities and provide more accurate responses than individual specialist models without being trained on such modality combinations.

Overall, these qualitative results reinforce that the proposed subspace alignment and merging mechanism effectively combines complementary cross-modal information and supports robust multimodal reasoning across diverse and previously unseen modality combinations.