UNIVERSITY OF CALIFORNIA
RIVERSIDE

Learning to Align Multimodal Data for Static and Dynamic Tasks

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

Doctor of Philosophy

in

Electrical Engineering

by

Sudipta Paul

December 2022

Dissertation Committee:

Dr. Amit K. Roy-Chowdhury, Chairperson
Dr. Evangelos Papalexakis
Dr. Salman Asif

The Dissertation of Sudipta Paul is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

I would like to express my gratitude to some people without whom I wouldn't have been able to complete my PhD dissertation. First and foremost, I would like to thank my advisor Dr. Amit K. Roy-Chowdhury for his constant help during the course of this dissertation. My PhD journey was never smooth and the progress was slow during the first couple of years due to some initial struggles. Nevertheless, Prof. Roy-Chowdhury never stopped trusting in my abilities and constantly motivated me to make progress systematically. He gave me the freedom to choose problem statements that are of my interest and allowed me to grow as an independent researcher. From problem formulation and scientific approach to solving a problem to preparing manuscripts or presentation slides, I have learned various aspects of doing good research from him over the course of five years. I am sure it will help me in my future endeavors.

I would like to thank my undergrad thesis supervisor S. M. Mahbubur Rahman who exposed me to the amazing field of computer vision. I would also like to thank my dissertation committee members, Dr. Evangelos Papalexakis and Dr. Salman Asif for giving me thoughtful feedback and constructive comments in improving the quality of this dissertation. I had the opportunity to work with both Dr. Evangelos Papalexakis and Dr. Salman Asif. Prof. Papalexakis's insights and ideas related to the multimodal learning field helped me during my research work. Prof. Asif's ability to do a thorough investigation of the feasibility of any problem statement or approach motivated me to do the same while working on the research problems. During my PhD, I took several courses to enhance my knowledge and aptitude to do better research. I would like to thank all the course instructors

including Dr. Anastasios Mourikis, Dr. Ertem Tuncel, Dr. Hamed Mohsenian-Rad, and Dr. Jay A Farrell for their exceptional teaching and presentation.

During the course of five years, I had the opportunity to go for summer internships twice. It was a great break from the monotonous work of my PhD and a great learning opportunity. In Mayachitra, I learned a lot about the structured and industrial aspects of programming from Dr. Shiv Chandrasekaran and Dr. Carlos Torres. In MERL, Dr. Anoop Cherian introduced me to the research field of Embodied AI. I learned a lot from him on how to do good research. I am indebted to all of my internship mentors and I thank them for their valuable guidance.

I had a great time with my lab mates at UC Riverside. I especially want to thank Niluthpol Chowdhury Mithun for all the discussions and guidance throughout my dissertation as a friend and as a mentor. Among the senior members of our lab, I would also like to thank Jawadul H. Bappy, Sujoy Paul, Akash Gupta, Shasha Li, and Rameswar Panda for the amazing time we spent together. I am fortunate enough to have some contemporary and junior lab-mates who made my PhD journey more memorable. I would like to thank Ghazal Mazaheri, Abhishek Aich, Dripta Raychaudhuri, Miraj Ahmed, Sayak Nag, and Cody Simons for being there when I needed any help. Even outside the lab group, I received a lot of support from my friends- Jubair Yusuf, Rakib Hyder, Risul Islam, Md Omar Faruk Rokon, Dipan shaw, and Farzana Rahman Rimi. Without them, my life at UC Riverside would have been much more challenging.

Finally, I would like to express my heartfelt regards to my parents (Sunil Chandra Pal and Snigdha Paul) and my sister (Sudeshna Paul) for their love and affection toward

me. They never had any doubts about my abilities and cheered me and motivated me to work harder. My parents have sacrificed a lot just to make sure I lead a happier life and I am forever indebted to them. I would also like to thank Tarek Salman Suvro and Jubayer Mahmod, who had been my friend since my undergraduate days and still continues to support me in my difficult times.

I thank the National Science Foundation (IIS- 1901379), Office of Naval Research (ONR N00014-19-1-2264, ONR N00014-15-C-5113), Navy (NavAir N-6833518-C-0199), and Mitsubishi Electric Research Lab(MERL) for their grants to Dr. Amit K. Roy-Chowdhury, which partially supported my research. I also thank Victor Hill of UCR CS and Tom Gregory of UCR ECE for setting up the computing infrastructure used in most of the works presented in this thesis. Acknowledgment of previously published materials: The text of this dissertation, in part or in full, is a reprint of the material as appeared in three previously published papers that I first authored. The co-author Dr. Amit K. Roy-Chowdhury, listed in all three publications, directed and supervised the research which forms the basis for this dissertation.

The papers are as follows:

- Sudipta Paul, Niluthpol Chowdhury Mithun, Amit K. Roy-Chowdhury, "Text-based Localization of moments in a Video Corpus", IEEE Transactions on Image Processing(T-IP), 2021

- Sudipta Paul, Niluthpol Chowdhury Mithun, Amit K. Roy-Chowdhury, "Text-based Temporal Localization of Novel Events", European Conference on Computer Vision (ECCV), 2022

- Sudipta Paul*, Amit K Roy-Chowdhury, Anoop Cherian*, "AVLEN: Audio-Visual-Language Embodied Navigation in 3D Environments", Neural Information Processing Systems (NeurIPS), 2022

To my parents for all the support.

ABSTRACT OF THE DISSERTATION

Learning to Align Multimodal Data for Static and Dynamic Tasks

by

Sudipta Paul

Doctor of Philosophy, Graduate Program in Electrical Engineering
University of California, Riverside, December 2022
Dr. Amit K. Roy-Chowdhury, Chairperson

Our experience of the world is multimodal - we see objects, hear sounds, and read texts to perceive information. In order for Artificial Intelligence to make progress in understanding the world around us, it needs to be able to interpret such multimodal signals together. The heterogeneity of the data brings unique challenges while working with multimodal signals. One such challenge is to identify and understand the alignment between two different modalities. In this dissertation, we focus on learning to align vision and language modalities in static and dynamic tasks in different scenarios.

In the first dimension, we address the task of text-based video moment localization. Existing approaches assume that the relevant video is already known/given and attempt to localize the moment based on text query on that given video only. We relax this strong assumption and address the task of localizing moments in a corpus of videos for a text query. This task poses a unique challenge as the system is required to perform retrieval of the relevant/correct video and temporal localization of the moment in the detected video based on the text query simultaneously. Our proposed approach learns to distinguish subtle

differences between intra-video moments as well as distinguish inter-video global semantic concepts based on text queries.

We also consider text-based temporal localization task where both the video moments and text queries are not observed/available during training. Conventional approaches are trained and evaluated relying on the assumption that the localization system, during testing, will only encounter events that are available in the training set. As a result, these models are unlikely to generalize to the practical requirement of localizing a wider range of events, some of which may be unseen. Towards solving this problem, we formulate the inference task of text-based localization of moments as a relational prediction problem, hypothesizing a conceptual relation between semantically relevant moments. The likelihood of a candidate moment being the correct one based on an unseen text query will depend on its relevance to the moment corresponding to the semantically most relevant seen query.

Continuing in the direction of learning to align multimodal data, we extend it to the dynamic task of Audio-Visual-Language embodied navigation in 3D environments. The goal of our embodied agent is to localize an audio event via navigating the 3D visual world; however, the agent may also seek help from a human (oracle), where the assistance is provided in free-form natural language. We propose a multimodal hierarchical reinforcement learning backbone that learns: (a) high-level policies to choose either audio cues for navigation or to query the oracle and (b) lower-level policies to select navigation actions based on its audio-visual or audio-visual-language inputs. The policies are trained via rewarding for the success of the navigation task while minimizing the number of queries to the oracle.

Contents

List of Figures **xiii**

List of Tables **xvi**

1 Introduction **1**

2 Moment Localization from Video Corpus **7**
2.1 Introduction . . . 7
2.1.1 Contributions . . . 10
2.2 Related Works . . . 11
2.3 Methodology . . . 14
2.3.1 Problem Statement . . . 15
2.3.2 Framework Overview . . . 16
2.3.3 Feature Extraction Unit . . . 17
2.3.4 Moment Encoder Module . . . 17
2.3.5 Sentence Encoder Module . . . 19
2.3.6 Learning Joint Embedding Space . . . 19
2.4 Experiments . . . 24
2.4.1 Datasets . . . 25
2.4.2 Evaluation Metric . . . 26
2.4.3 Implementation Details . . . 28
2.4.4 Analysis of Results . . . 30
2.4.5 Qualitative Results . . . 42
2.5 Conclusion . . . 43

3 Temporal Localization of Novel Events **45**
3.1 Introduction . . . 45
3.2 Related Works . . . 48
3.3 Methodology . . . 50
3.3.1 Problem Statement . . . 50
3.3.2 Localization Inference Schema . . . 51
3.3.3 Framework . . . 53

3.3.4 Relational Prediction . . . 55
3.3.5 Learning Relational Inference . . . 56
3.3.6 Inference for Unseen Queries . . . 58
3.4 Experiments . . . 58
3.4.1 Reorganized Datasets . . . 58
3.4.2 Evaluation Metric . . . 61
3.4.3 Implementation Details . . . 62
3.4.4 Result Analysis . . . 64
3.5 Conclusion . . . 72

4 Audio-Visual-Language Navigation **74**
4.1 Introduction . . . 74
4.2 Related Works . . . 78
4.3 Proposed Method . . . 80
4.3.1 Problem Setup . . . 80
4.3.2 Problem Formulation . . . 81
4.3.3 Multimodal Hierarchical Deep Reinforcement Learning . . . 83
4.3.4 Navigation Using Audio Goal Policy, π_g. . . . 86
4.3.5 Navigation Using Language Policy, π_ℓ . . . 86
4.3.6 Learning When-to-Query Policy, π_q . . . 88
4.3.7 Reward Design . . . 88
4.3.8 Policy Training . . . 89
4.3.9 Generating Oracle Navigation Instructions . . . 90
4.4 Experiments and Results . . . 91
4.4.1 Dataset . . . 91
4.4.2 Evaluation Metrics . . . 92
4.4.3 Implementation Details . . . 92
4.4.4 Experimental Results and Analysis . . . 93
4.5 Conclusions . . . 101

5 Conclusions **102**
5.1 Dissertation Summary . . . 102
5.2 Future Research Directions . . . 104
5.2.1 Webly Supervised/Knowledge Transfer . . . 104
5.2.2 Bi-directional Interaction of Navigating Agent . . . 104
5.2.3 Discrete Alignment vs Continuous Alignment . . . 105
5.2.4 Few-shot Capability . . . 105
5.2.5 Learning with Noisy Annotation . . . 105
5.2.6 Scalable Vision-language Model . . . 105
5.2.7 Task Adaptation . . . 106
5.2.8 Enforcing Sparsity on Transformer . . . 106

Bibliography **107**

List of Figures

2.1 Example illustration of our proposed task. We consider localizing moments in a corpus of videos given a text query. Here, for the queried text: 'Person puts clothes into a washing machine', the system is required to identify the relevant video-*(b)* from the illustrated corpus of videos (video-*(a)*, video-*(b)*, and video-*(c)*) and temporally localize the pertinent moment (ground truth moment marked by the green dashed box) in that relevant video. 8

2.2 A brief illustration of the proposed Hierarchical Moment Alignment Network for the moment localization task in a video corpus. The framework uses the feature extraction unit to extract clip and sentence features. Hierarchical moment encoder module and sentence encoder module projects moment representations and sentence representations in the joint embedding space respectively. The network learns to align moment-sentence pairs in the joint embedding space by explicitly focusing on distinguishing intra-video moments and inter-video global semantic differences. (Details of the learning procedure in section 2.3.6) 11

2.3 A conceptual representation of our proposed learning objective. For a text query s with relevant moment m_{11} in a set of videos $\{v_1, v_2\}$ with set of moments $\{m_{11}, m_{12}, m_{21}, m_{22}\}$, we learn the joint embedding space using- (a) intra-video moments: increasing similarity for relevant pair (m_{11}, s) and decreasing similarity for non-relevant pair (m_{12}, s) from the same video, (b) global semantics of video: increasing video-sentence relevance for relevant pair (v_1, s) and decreasing for non-relevant pair (v_2, s), where the video-sentence relevance is computed in terms of moment-sentence similarity. This is also illustrated in (c), where the arrows indicate which pairs are learning to increase their similarity (moving close in the embedding space) and which pairs are learning to decrease their similarity (moving further away in the embedding space). Details can be found in section 2.3.6 15

2.4 Illustration of λ_1 sensitivity on the HMAN performance. We observe that for the set of values $\{3, 4, 5, 6, 7\}$, performance of HMAN is stable. 37

2.5 t-SNE visualization of text query representation and candidate moment representations. Different color represents different video. The color of the text representation is the same as the corresponding video. We use different markers for the representation of incorrect candidate moments, correct candidate moments and text. Here, representations of the text query and the correct candidate moment coincide. Also, the representations of candidate moments from the same video are clustered together. 39
2.6 Example illustration of the performance of HMAN for the task of localization of moments in a corpus of videos. For each query sentence, we display the top-3 retrieved moments. The retrieved moments are surrounded by gold boxes and the ground truth moments are indicated by green lines. We observe that for each of the queries, the top-3 retrieved moments are semantically related to the sentence proving the efficacy of our approach. 41

3.1 Example illustration of our proposed task. We consider the task of localizing novel moments for unseen queries. The set of verbs and nouns present in the testing set is absent in the training set, e.g., training data does not have any text with verb 'walk' or noun 'doorway'. Hence, the system is required to learn transferable knowledge from the training data to perform localization for novel events based on unseen queries. 46
3.2 A brief illustration of our novel text-based temporal localization approach. While existing works learn to encode video segments to identify the correct moment ((a) and (b)), we consider relational reasoning between two semantically relevant moment for localization purpose (c). 52
3.3 Overview of the framework and the training of the relational reasoning based temporal localization approach. Candidate moment and support moment representations are aggregated to form positive pairs (positive candidate, positive support) and negative pairs (negative candidate, positive support)/(positive candidate, negative support). The relational module is trained to estimate the relational scores based on the pairs. 53
3.4 List of selected verbs and nouns for Charades-STA Unseen. 61
3.5 List of selected verbs and nouns for ActivityNet Captions Unseen. 62
3.6 Given the query 'The person laughs' and the corresponding video, this figure shows: (a) ground truth segment of the video which corresponds to the text query, (b) predicted moment by 2D-TAN, (c) predicted moment when irrelevant moment is used as support, and (d) predicted moment using retrieved relevant support moment (TLRR). While (b) and (c) result in failure, TLRR is able to detect the correct moment using relational reasoning. 69

3.7 Example illustration from ActivityNet Captions Unseen, where splits are created based on activity annotation. Given the text query, 'An older blonde newswomen is reading a story' and the corresponding video, our proposed approach retrieves the moment corresponding to the semantically relevant query 'A blonde woman is talking in a room' from the train set, reason on that and identifies the correct moment in the video. GT indicates the ground truth timestamps and Prediction indicates the predicted temporal endpoints of our approach. 70
3.8 Given the text query 'Person walking through the doorway' and the corresponding video, our proposed approach retrieves the moment corresponding to the semantically relevant query 'Person running to the door' from the train set, reason on that, and identifies the correct moment in the video. GT indicates the ground truth timestamps and Prediction indicates the predicted temporal endpoints of our approach. 70
3.9 This figure illustrates the performance of SCDM [177] for Charades-STA Unseen and 2D-TAN [188] for ActivityNet Captions Unseen dataset for seen events and unseen events. For both datasets, performance of the trained model drops significantly for unseen events. 71

4.1 An illustration of our proposed AVLEN framework. The embodied agent starts navigating from location denoted ① guided by the audio-visual event at ③. At location ②, the learned policy for the agent decides to seek help from an oracle (e.g., because the audio stopped). The oracle provides a short natural language instruction for the agent to follow. The agent translates this instruction to produce a series of navigable steps to move towards the goal ③. 75
4.2 Architecture of our AVLEN pipeline. We show the two-level hierarchical RL policies that the model learns (offline), as well as the various input modalities and the control flow. 84
4.3 Network architecture for goal-based navigation policy π_g. The model architecture is similar to option selection/query policy π_q. However, the action space is different for these two policies. 87
4.4 Network architecture for language-based navigation policy π_ℓ 88
4.5 Performance (SPL) comparison against varying the number of allowed queries. 94
4.6 Distribution of queries triggered against the time steps in episodes. 95
4.7 Two qualitative results from AVLEN's navigation trajectories. We show egocentric views and top down maps for three different viewpoints in agent's trajectory. The agent starts from ①, receives oracle help in ②, navigates to the goal in ③. In the top episode, agent receives directional information ('Walk forward and turn right'), whereas in the bottom episode, agent receives language instruction more grounded on the scene ('Walk down the hallway'). 96
4.8 Sensitivity to the number of queries ν to the oracle that AVLEN can make. The results are for the unheard sound sceneario. Please see the main paper for plots on the success rate. 99
4.9 Robustness to silence duration analysis 100

List of Tables

2.1 Tabulated summary of the details of dataset contents 23
2.2 Tabulated summary of the implementation details regarding video processing for three datasets 24
2.3 Comparison of performance for the task of temporally localizing moments in a video corpus on DiDeMo dataset. († reported from [36]) (↓ indicates the performance is better if the score is low) 24
2.4 Comparison of performance for the task of temporally localizing moments in a video corpus on Charades-STA dataset. († reported from [36]) (↓ indicates the performance is better if the score is low) 25
2.5 Comparison of performance for the task of temporally localizing moments in a video corpus on ActivityNet Captions dataset. († reported from [36]) 27
2.6 Comparison of the performance of HMAN with/without the Hierarchical moment Encoder Module. The experiments are done for DiDeMo and Charades-STA datasets. († reported from [36]) (↓ indicates the performance is better if the score is low) 27
2.7 Ablation study for the effectiveness of learning embedding space utilizing different loss components as described in 2.3.6 for DiDeMo dataset using sum-margin set up. 28
2.8 Performance comparison for the task of retrieving correct video based on sentence query on DiDeMo and Charades-STA dataset. 29
2.9 Comparison of the performance of proposed LogSumExp pooling and average pooling. We compare the performance for the task of temporal localization of moments in video corpus for DiDeMo and Charades-STA dataset. 29
2.10 Ablation Study of the performance of HMAN (sum-margin) for Different Visual Features for DiDeMo dataset. 30
2.11 Ablation study of the performance of HMAN (sum-margin) on DiDeMo when the number of test set data is decreased. 32
2.12 Per epoch training and inference time for Charades-STA dataset. 34

3.1 Tabulated summery of **number of moment-text pairs** in Charades-STA Unseen and ActivityNet Captions Unseen dataset. 60

3.2 Tabulated summery of **number of videos** in the reorganized Charades-STA Unseen and ActivityNet Captions Unseen dataset. 61
3.3 **Number of verbs and nouns** used to create train/test splits of Charades-STA Unseen and ActivityNet Captions Unseen dataset. 61
3.4 This table reports ***unseen*** text query based temporal moment localization performance of TLRR, compared against several approaches, on Charades-STA Unseen dataset. 63
3.5 This table reports ***unseen*** text query based temporal moment localization performance of TLRR, compared against several approaches, on ActivityNet Captions Unseen dataset. 64
3.6 This table reports *unseen* text query based novel event localization performance using different types of support moments to analyze TLRR for Charades-STA Unseen dataset. 66
3.7 This table reports ***seen*** text query based temporal moment localization performance of TLRR on Charades-STA Unseen dataset. Here, Δ_{avg} refers to average performance difference for seen events and unseen events (Table 3.4) for a specific method. From the lower value of Δ_{avg}, it is evident that TLRR generalizes significantly better than other temporal localization approaches. . 68
3.8 This table reports ***seen*** text query based temporal moment localization performance of TLRR on ActivityNet Captions Unseen dataset. Δ_{avg} refers to average performance difference for seen events and unseen events (Table 3.5) for a specific method. From the lower value of Δ_{avg}, it is evident that TLRR generalizes significantly better than other temporal localization approaches. . 68
3.9 Per batch inference time of TLRR compared to SCDM and 2D-TAN in ActivityNet Captions Unseen dataset. 72
3.10 This table reports text query based temporal moment localization performance of TLRR on the original Charades-STA dataset. 72

4.1 Comparison of performances against state of the art in heard and unheard sound settings. 91
4.2 Comparisons in heard and unheard sound settings against varied query-triggering methods. 91
4.3 Comparisons against varied query-triggering methods with ground truth action as feedback. 92
4.4 Comparison of AVLEN performances against baselines and when-to-query approaches in the *presence of distractor sound* 93
4.5 Comparisons in performance for different architectural choices for language-based policy π_ℓ in heard sound setting. 95
4.6 Vision-language navigation performance. 100

Chapter 1

Introduction

Our experience of the world is multimodal - we see objects, hear sounds, and read texts to perceive information. In order for Artificial Intelligence to make progress in understanding the world around us, it needs to be able to interpret such multimodal signals together [8]. Multimodal machine learning aims to design systems that can understand and reason on multiple communicative modalities. It is a prominent research field that requires bridging knowledge from different disciplines (e.g., vision, language, speech, audio). Although multimodal machine learning encompasses a broader range of modality information, we narrow down our focus to vision and language modality. There is a wide range of application domains that involve both vision and language modalities, e.g., image/video-text retrieval [37, 153, 193, 78, 24, 74, 102, 21, 142, 32, 48], image/video captioning [7, 150, 191, 39, 171, 61], text-to-image/video generation [79, 52, 121, 107, 44, 96], referring image/video segmentation [158, 40, 170], visual question answering [168, 130, 46, 9], and vision-language navigation [4, 59, 64, 85, 97, 94, 95, 196]. Working with multiple modalities together requires

addressing challenges that arise due to the heterogeneity of data, which is unique compared to independent research of vision or language modality. One such challenge is to identify and understand the alignment between two different modalities. In this dissertation, we focus on learning to align vision and language modalities in static and dynamic tasks in different scenarios.

We define an alignment problem as finding and understanding correspondence between instances of two different modalities. For example, video moment retrieval based on the text query requires a system that has an understanding of what text query matches with which segment of the video. While the main focus of the alignment challenge is to learn the relationship between two modalities, understanding when a model fails to relate two modalities or is uncertain of the alignment is also of significant importance. In this dissertation, we consider learning and understanding the alignment of different multimodal scenarios; specifically focusing on vision and language modality. We first consider text-based temporal localization moment in video corpus [114]. With full supervision available at hand, our task is to align a text query with a particular segment of a video from a collection of videos. This is achieved by considering the intra-video subtle differences in context and inter-video global semantic differences. Then we consider text-based temporal localization of novel events where the system needs to perform well on unseen events or queries [115]. Since we don't have supervision for novel/unseen events, a conventional contrastive learning approach with available seen data is guaranteed to fail. As a result, we formulate the alignment inference as a relational reasoning problem to solve the task. Finally, we consider a dynamic task of audio-visual-language navigation [116]. Understanding when to interact with and how to

utilize natural language feedback from the interaction is of great importance for AI. Our work develops a system that can identify when to interact based on the alignment uncertainty of audio-visual modality and vision-language modality. If the audio-visual alignment is uncertain then the system queries for help and receives language feedback to assist its navigation. Along with the three above mentioned works that mainly focus on learning and understanding the alignment of multimodal data, we have worked and collaborated on several other projects [117, 124, 113, 76] that have strengthened our understanding in the field of multimodal learning.

In Chapter 2, we discuss the task of text-based temporal localization of video moments in a collection of videos [114]. Prior works on text-based video moment localization [102, 21, 142, 32, 48] focus on temporally grounding the textual query in an untrimmed video. These works assume that the relevant video is already known and attempt to localize the moment on that relevant video only. Different from such works, we relax this assumption and address the task of localizing moments in a corpus of videos for a given sentence query. This task poses a unique challenge as the system is required to perform: (i) retrieval of the relevant video where only a segment of the video corresponds with the queried sentence, and (ii) temporal localization of moment in the relevant video based on sentence query. Towards overcoming this challenge, we propose Hierarchical Moment Alignment Network (HMAN) which learns an effective joint embedding space for moments and sentences. In addition to learning subtle differences between intra-video moments, HMAN focuses on distinguishing inter-video global semantic concepts based on sentence queries. We validate our approach quantitatively and qualitatively on three benchmark text-based video moment

retrieval datasets.

In Chapter 3, we consider the task of text-based temporal localization of novel events [115]. Recent works on text-based localization of moments [102, 21, 142, 32, 48] have shown high accuracy on several benchmark datasets. However, these approaches are trained and evaluated relying on the assumption that the localization system, during testing, will only encounter events that are available in the training set (i.e., seen events). As a result, these models are optimized for a fixed set of seen events and they are unlikely to generalize to the practical requirement of localizing a wider range of events, some of which may be unseen. Moreover, acquiring videos and text comprising all possible scenarios for training is not practical. In this regard, our goal is to temporally localize video moments based on text queries, where both the video moments and text queries are not observed/available during training. Towards solving this problem, we formulate the inference task of text-based localization of moments as a relational prediction problem, hypothesizing a conceptual relation between semantically relevant moments, e.g., a temporally relevant moment corresponding to an unseen text query and a moment corresponding to a seen text query may contain shared concepts. The likelihood of a candidate moment being the correct one based on an unseen text query will depend on its relevance to the moment corresponding to the semantically most relevant seen query. Empirical results on two reorganized text-based moment localization datasets show that our proposed approach can reach up to 15% absolute improvement in performance compared to existing localization approaches.

In Chapter 4, we look into the dynamic task of audio-visual-language navigation. Recent years have seen embodied visual navigation advance in two distinct directions: (i)

in equipping the AI agent to follow natural language instructions, and (ii) in making the navigable world multimodal, e.g., audio-visual navigation. However, the real world is not only multimodal, but also often complex, and thus in spite of these advances, agents still need to understand the uncertainty in their actions and seek instructions to navigate. To this end, we present AVLEN – an interactive agent for Audio-Visual-Language Embodied Navigation [116]. Similar to audio-visual navigation tasks, the goal of our embodied agent is to localize an audio event via navigating the 3D visual world; however, the agent may also seek help from a human (oracle), where the assistance is provided in free-form natural language. To realize these abilities, AVLEN uses a multimodal hierarchical reinforcement learning backbone that learns: (a) high-level policies to choose either audio-cues for navigation or to query the oracle, and (b) lower-level policies to select navigation actions based on its audio-visual and language inputs. The policies are trained via rewarding for the success on the navigation task while minimizing the number of queries to the oracle. To empirically evaluate AVLEN, we present experiments on the SoundSpaces [16] framework for semantic audio-visual navigation tasks. Our results show that equipping the agent to ask for help leads to a clear improvement in performance, especially in challenging cases, e.g., when the sound is unheard during training or in the presence of distractor sounds.

Organization of the Dissertation. In Chapters 2 and 3, we consider the static task of text-based temporal localization of moments and in Chapter 4, we consider the dynamic task of audio-visual-language navigation. In chapter 2, we discuss the task of text-based temporal localization of video moments in a collection of videos. In Chapter 3, we consider the task of text-based temporal localization of novel events. In Chapter 4, we

look into the dynamic task of audio-visual-language navigation. In each of the chapters, we discuss the problem setting, motivation, proposed approach and experimental results. Finally, we conclude the dissertation in Chapter 5 with some interesting future directions of works in multimodal learning for static and dynamic tasks.

Chapter 2

Moment Localization from Video Corpus

2.1 Introduction

Localizing activity moments in long and untrimmed videos is a prominent video analysis problem. Early works on moment localization were mostly limited by the use of a predefined set of labels to describe an activity [73, 14, 133, 81]. However, due to the nature of the complexity of real-life activities, natural language sentences would be the appropriate choice to describe an activity rather than a predefined set of labels. Recently, there are several works [47, 5, 156, 89, 20, 49, 162, 182, 177, 83] that utilize sentence queries to temporally localize moments in untrimmed videos. All these approaches build upon an underlying assumption that the correspondence between sentences and videos is known. As a result, these approaches attempt to localize moments only in the related video. We argue

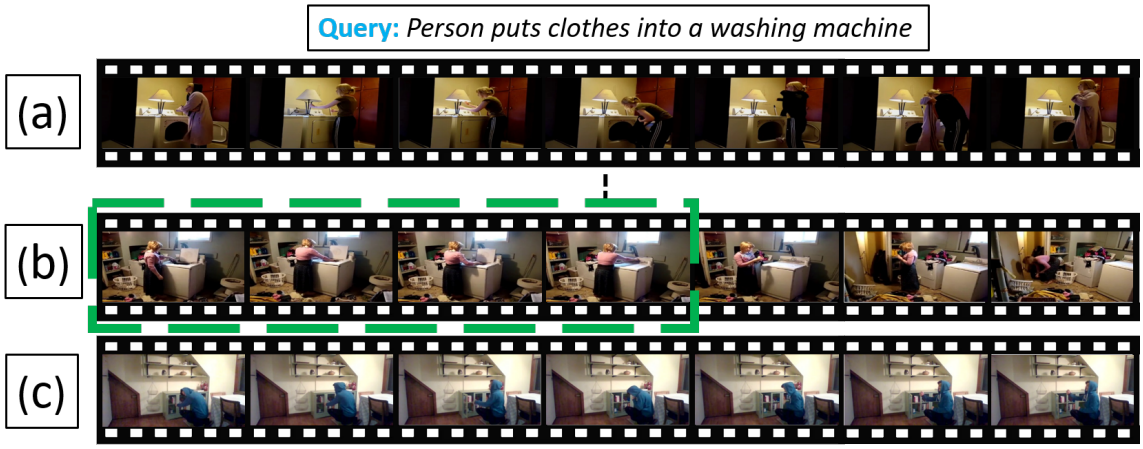


Figure 2.1: Example illustration of our proposed task. We consider localizing moments in a corpus of videos given a text query. Here, for the queried text: 'Person puts clothes into a washing machine', the system is required to identify the relevant video-*(b)* from the illustrated corpus of videos (video-*(a)*, video-*(b)*, and video-*(c)*) and temporally localize the pertinent moment (ground truth moment marked by the green dashed box) in that relevant video.

that such an assumption of knowing relevant videos a priori may not be plausible for most practical scenarios. It is more likely that a user would need to retrieve a moment from a large corpus of videos given a sentence query.

In this work, we relax the assumption of specified video-sentence correspondence of the prior works on temporal moment localization and address the more challenging task of localizing moments in a corpus of videos. For example in Figure 2.1, the moment marked by the green dashed box in video-*(b)* corresponds to the text query: *'Person puts clothes into a washing machine'*. Prior works on temporal moment localization only attempt to detect the temporal endpoints in the given video-*(b)* by learning to identify subtle changes in dynamics of the activity. However, the task of localizing the correct moment in the illustrated collection of videos (i.e., *(a)*, *(b)*, and *(c)* in Figure 2.1) imposes the additional requirement to distinguish moments from different videos and identify the correct video (video-*(b)*) based

on the differences of putting and pulling activities as well as the presence of washing machine and clothes.

To address this problem, a trivial approach would be to use an off-the-shelf video-text retrieval module to retrieve the relevant video and then localize the moment in that retrieved video. Most of the video-text retrieval approaches [181, 99, 129, 155, 34, 119, 157, 41] are designed for cases where videos and text queries have a one-to-one correspondence, i.e., a query sentence reflects a trimmed and short video or a query paragraph represents a long and untrimmed video. However, in our addressed task, the query sentence reflects a segment of a long and untrimmed video, and different segments of a video can be associated with different language annotations, resulting in one-to-many video-text correspondence. Hence, the existing video-text retrieval approaches are likely to fall short on our target task. Another trivial approach would be to scale up the temporal localization of moments approaches, i.e., instead of searching over a given video, it searches over the corpus of videos. However, these approaches are only designed to discern intra-video moments based on sentence semantics and fail to distinguish moments from different videos and identify the correct video.

In this work, based on the text query, we focus on discerning moments from different videos as well as understand the nuances of activities simultaneously to localize the correct moment in the relevant video. Our objective is to learn a joint embedding space that will align representations of corresponding video moments and sentences. For this, we propose **H**ierarchical **M**oment **A**lignment **N**etwork (**HMAN**), a novel neural network framework that effectively learns a joint embedding space to align corresponding video moments and sentences. Learning joint embedding space for retrieval or localization tasks has been addressed by

several other methods [5, 36, 41, 110, 173, 172]. Among them, [5] and [36] are closely related to our work as they try to align corresponding moment and sentence representations in the joint embedding space. However, our approach is significantly different from these works. In contrast to these works, HMAN utilizes temporal convolutional layers in a hierarchical structure to represent candidate video moments. It allows the model to generate all candidate moment representations of a video in a single pass, which is more efficient than sliding based approaches like [5, 36]. Our learning objective is also different from [5, 36], where they only try to distinguish between intra-video moments and inter-video moments. In our proposed approach, in addition to distinguishing intra-video moments, we propose a novel learning objective that utilizes text-guided global semantics to distinguish different videos. Global semantics of a video refers to the semantics that is common across most of the moments of that video. As the global semantics vary across videos, by distinguishing videos, we learn to distinguish inter-video moments. We demonstrate the advantage of our proposed approach over other baseline approaches and contemporary works on three benchmark datasets.

2.1.1 Contributions

The main contributions of the proposed work are as follows:

- We explore an important, yet under-explored, problem of text query-based localization of moments in a video corpus.
- We propose a novel framework, HMAN, that uses stacked temporal convolutional layers in a hierarchical structure to represent video moments and texts jointly in an embedding space. Combined with the proposed learning objective, the model is able to align moment

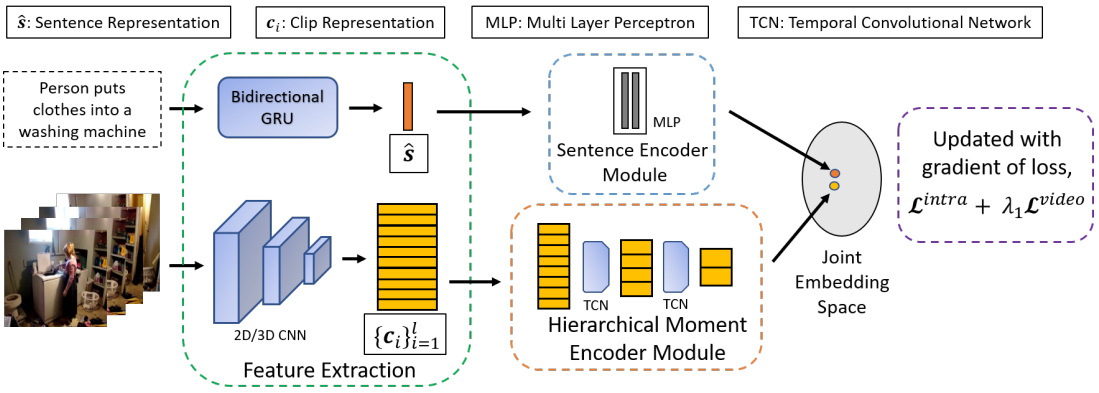


Figure 2.2: A brief illustration of the proposed Hierarchical Moment Alignment Network for the moment localization task in a video corpus. The framework uses the feature extraction unit to extract clip and sentence features. Hierarchical moment encoder module and sentence encoder module projects moment representations and sentence representations in the joint embedding space respectively. The network learns to align moment-sentence pairs in the joint embedding space by explicitly focusing on distinguishing intra-video moments and inter-video global semantic differences. (Details of the learning procedure in section 2.3.6)

and sentence representations by distinguishing both local subtle differences of the moments as well as global semantics of the videos simultaneously.

- Towards solving the problem, we propose a novel learning objective that utilizes text-guided global semantics of the videos to distinguish moments from different videos.

- We empirically show the efficacy of our proposed approach on DiDeMo, Charades-STA, and ActivityNet Captions dataset and study the significance of our proposed learning objective.

2.2 Related Works

Video-Text Retrieval. Among the cross-modal retrieval tasks [72, 88, 30, 33, 100], video-text retrieval has gained much attention recently. Emergence of datasets like the

Microsoft Research Video to Text (MSR-VTT) [163], the MPII movie description dataset as part of the Large Scale Movie Description Challenge (LSMDC) dataset [125], and Microsoft Video Description Dataset (MSVD) [18] have boosted video-text retrieval task. These datasets contain short video clips with accompanying natural language. Initial approaches for the video-text retrieval task were based on concept classification [98, 69, 145]. Recent approaches focus on directly encoding video and text in a common space and retrieving relevant instances based on some similarity measure in the common space [33, 100, 165, 175, 77, 25]. These works used Convolutional Neural Network (CNN) [175] or Long Short-Term Memory Network (LSTM) [176] for video encoding. To encode text representations, Recurrent Neural Network (RNN) [165], bidirectional LSTM [175] and GRU [99] were commonly used. Mithun et al. [99] employed multimodal cues such as image, motion, and audio for video encoding. In [34], multi-level encodings for video and text were used and both videos and sentences were encoded in a similar manner. Liu et al. [91] proposed collaborative experts model to aggregate information effectively from different pre-trained experts. Yu et al. [175] proposed a Joint Sequence Fusion model for sequential interaction of videos and texts. Song et al. [136] introduced Polysemous Instance Embedding Networks that compute multiple and diverse representations of an instance. Among the recent works, Wray et al. [155] enriched the embedding learning by disentangling parts-of-speech of captions. Chen et al. [23] used Hierarchical Graph Reasoning to improve fine-grained video-text retrieval. Another line of work considers video-paragraph retrieval. For example, Zhang et al. [181] proposed hierarchical modeling of videos, and paragraphs and Shao et al. [129] utilized top-level and part-level association for the task of video-paragraph retrieval. However, all of these

approaches have an underlying assumption that videos and text queries have one-to-one correspondence. As a result, they are not adaptable for our addressed task, where the video-text pairs have one-to-many correspondence.

Temporal Localization of Moments. The task of localizing a moment/activity in a given long and untrimmed video via text query was introduced in [47, 5]. After that, there have been a lot of works [156, 89, 20, 49, 162, 182, 177, 63, 90, 189, 101, 50, 178, 190, 188, 53, 51, 56, 84, 123] that addressed this task. All of these works on temporal localization of moments can be divided into two categories: i) two stage approaches that sample segments of videos in the first step and then try to find a semantic alignment between sentences and those segments of videos in the second step [47, 5, 156, 89, 20, 49, 162], and ii) single stage approaches that predict the association of sentences with multi-scale visual representation units as well as predict temporal boundary for each visual representation unit in a single pass [182, 177]. Among all the approaches, Gao et al. [47] developed Cross-modal Temporal Regression Localizer that jointly models text queries and video clips. A common embedding space for video temporal context features and language features was learnt in [5]. Some of the works focused on vision-language fusion techniques to improve localization performance. For example, Multimodal Circulant Fusion was incorporated in [156]. Liu et al. [89] incorporated a memory attention mechanism to emphasize the visual features mentioned in the query and simultaneously use their context. Ge et al. [49] mined activity concepts from both video and language modalities to improve the regression performance. Chen et al. [20] proposed Temporal GroundNet which captures evolving fine-grained frame-by-word interactions. Xu et al. [162] used early integration of vision and language for proposal generation and query

sentence modulation using visual features. Among the single shot approaches, candidate moment encoding and temporal structural reasoning were unified in a single shot framework in [182]. Semantic Conditioned Dynamic Modulation (SCDM) was proposed in [177] for correlating sentence and related video contents. These approaches on moment localization in a given video show promise, but fall short on realizing the requirement of identifying the correct video to address the task of moment localization in a corpus of videos.

There has been one concurrent work [36] that addressed the task of temporal localization of moments in a video corpus. They adopted the approach of Moment Context Network [5]. However, instead of directly learning moment-sentence alignment as in [5], they tried to learn clip-sentence alignment for scalability issues where a moment consists of multiple clips. Even so, a referring event is likely to consist of multiple clips, and a single clip can not reflect the complete dynamics of an event. Hence, consecutive clips with different content need to be aligned with the same sentence which results in suboptimal representation for both the clips and the sentence. We later empirically show that our approach is significantly more effective than [36] in the addressed task.

2.3 Methodology

In this section, we present our framework for the task of text-based temporal localization of moments in a corpus of untrimmed and unsegmented videos. First, we define the problem and provide an overview of the HMAN framework. Then, we present how clip-level video representations and word-level sentence representations are extracted. Then, we describe the framework in detail along with the hierarchical temporal convolutional

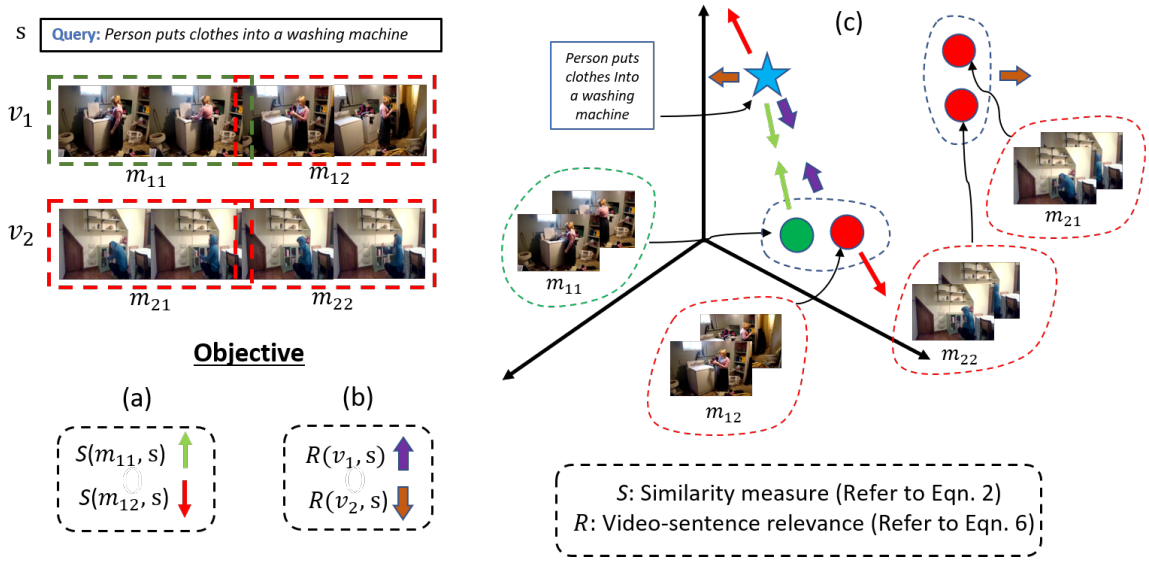


Figure 2.3: A conceptual representation of our proposed learning objective. For a text query s with relevant moment m_{11} in a set of videos $\{v_1, v_2\}$ with set of moments $\{m_{11}, m_{12}, m_{21}, m_{22}\}$, we learn the joint embedding space using- (a) intra-video moments: increasing similarity for relevant pair (m_{11}, s) and decreasing similarity for non-relevant pair (m_{12}, s) from the same video, (b) global semantics of video: increasing video-sentence relevance for relevant pair (v_1, s) and decreasing for non-relevant pair (v_2, s), where the video-sentence relevance is computed in terms of moment-sentence similarity. This is also illustrated in (c), where the arrows indicate which pairs are learning to increase their similarity (moving close in the embedding space) and which pairs are learning to decrease their similarity (moving further away in the embedding space). Details can be found in section 2.3.6

network to generate moment embeddings and sentence embeddings. Finally, we describe how we learn to encode moment and sentence representations in the joint embedding space for effective retrieval of the moment based on a text query.

2.3.1 Problem Statement

Consider that we have a set of N long and untrimmed videos $\mathcal{V} = \{v_i\}_{i=1}^{N}$, where a video v is associated with m_v temporal sentence annotations $\mathcal{T} = \{(s_j, \tau_j^s, \tau_j^e)\}_{j=1}^{m_v}$. Here, s_j is the sentence annotation and τ_j^s, τ_j^e are the starting time and ending time of the moment in

the video that corresponds with the sentence annotation s_j. The set of all temporal sentence annotations is $\mathcal{S} = \{\mathcal{T}_i\}_{i=1}^{N}$. Given a natural language query s, our task is to predict a set $s_{det} = \{v, \tau^s, \tau^e\}$ where, v is the video that contains the relevant moment and τ^s, τ^e are the temporal information of that moment.

2.3.2 Framework Overview

Our goal is to learn representations for candidate moments and sentences in such a way that the related moment-sentence pairs are aligned in the joint embedding space. Towards this goal, we propose HMAN, which is illustrated in Figure 2.2. First, we employ a feature extraction unit to extract clip level features $\{\boldsymbol{c}_i\}_{i=1}^{l}$ from a video and sentence features $\hat{\boldsymbol{s}}$ from a sentence. Clip representations and sentence representations are used to learn the semantic alignment between sentences and candidate moments. To project the moment representations and sentence representations in the joint embedding space, we use a hierarchical moment encoder module and a sentence encoder module respectively. The moment encoder module is inspired by single shot temporal action detection approach [81] where temporal convolutional layers are stacked in a hierarchical structure to obtain multi-scale moment features representing video segments of different duration. For the sentence encoder module, we use a two-layer feedforward neural network. Based on text queries, we derive the learning objective to explicitly focus on distinguishing intra-video moments and inter-video global semantics. We adopted sum-margin based triplet loss [43] and max-margin based triplet loss [43] separately in two different settings to train the model in an end-to-end fashion. In the inference stage, for a query sentence, the candidate moment with the most similar representation is retrieved from the corpus of videos.

2.3.3 Feature Extraction Unit

To work with data from different modalities, we extract feature representations using modality specific pretrained models.

Video Feature Extraction. We extract high level video features using a deep convolutional neural network. Each video v is divided into a set of l non-overlapping clips and we extract features for each clip. As a result, the video is represented by a set of features $\{\boldsymbol{c}_i\}_{i=1}^{l}$, where $\boldsymbol{c}_i$ is the feature representation of the i^{th} clip. To generate representations for all the candidate moments of a video in a single shot approach [81], we keep the input video length, i.e., number of clips, l, fixed. A video longer than the fixed length is truncated and a video shorter than the fixed length is padded with zeros.

Sentence Feature Extraction. To represent sentences, we use GloVe word embedding [118] for each word in a sentence. Then these word embedding sequences are encoded using a Bi-directional Gated Recurrent Unit (GRU) [28] with 512 hidden states. Here, words in a sentence are represented by a 512-dimensional vector, corresponding to their GRU hidden states. So, we can have a set of word-by-word representations of a sentence $S = \{\boldsymbol{h}_i\}_{i=1}^{n}$, where n is the number of words present in the sentence. The average of the word representations is used as the sentence representation $\hat{\boldsymbol{s}}$.

2.3.4 Moment Encoder Module

Existing approaches for moment localization based on learning joint visual-semantic embedding space either use a temporal sliding window with multiple scales [5] or optimize over a predefined set of consecutive clips based on clip-sentence similarity [36] to generate

candidate segments. However, sliding over a video with different scales or optimizing for all possible combinations of clips is computationally heavy. Again, in both cases, extracted candidate moments or predefined clips are projected in the joint embedding space independent of neighboring or overlapping moments/clips of the same video. Consequently, while learning the moment-sentence or clip-sentence semantic alignment, representations for neighboring or overlapping moments are not constrained to be well clustered to preserve the semantic similarity. Therefore, instead of projecting representations for candidate moments independently and inefficiently in the joint embedding space, inspired by the single shot activity detection [81], we use temporal convolutional layers [71] in a hierarchical setup to project representations for all candidate moments of a video simultaneously. We use a stack of $1D$ convolutional layers where the convolution operation can be denoted as $Conv(\sigma_k, \sigma_s, d)$. Here, σ_k, σ_s, and d indicate the kernel size, stride size, and filter numbers, respectively. The set of moment representations generated for K layers of hierarchical structure is $\{\{\boldsymbol{m}_i^k\}_{i=1}^{T_k}\}_{k=1}^{K}$. Here, T_k is the temporal dimension of the k^{th} layer, which decreases in the following layers. $\boldsymbol{m}_i^k \in \mathcal{R}^d$ is the i^{th} moment representation of the k^{th} layer and k^{th} layer generates T_k moment representations. Feature representations in the top layers of the hierarchy correspond to moments with shorter temporal duration, while the feature representations in the bottom layers correspond to moments with longer duration in a video. We keep the feature dimension of each moment representation fixed to d for all the layers of the temporal convolutional network.

Algorithm 1 Learning optimized HMAN (max-margin case)

Input: Untrimmed video set $\mathcal{V}$, Temporal sentence annotation set $\mathcal{S}$, Initialized HMAN weights θ
for t = 1 to maxIter **do**
step 1: Construct minibatch of video-sentence pairs
step 2: Extract video and sentence feature
step 3: Project candidate moment and sentence representations in the joint embedding space
step 4: Construct triplets
step 5: Compute $\mathcal{L}_{max}^{intra}$ and $\mathcal{L}_{max}^{video}$ using Eqn. 2.5 & 2.10
step 6: Optimize θ by minimizing total loss
end for
Output: Optimized HMAN weights θ

2.3.5 Sentence Encoder Module

We learn to project the textual representations in the joint embedding space keeping the inputs from different modalities with similar semantics close to each other. We use two layers of feedforward neural network with learnable parameters $\boldsymbol{W}_1^s$, $\boldsymbol{W}_2^s$, $\boldsymbol{b}_1^s$, and $\boldsymbol{b}_2^s$ to project the sentence representation $\hat{\boldsymbol{s}}$ in the joint embedding space, which can be defined as,

$$\boldsymbol{s} = \boldsymbol{W}_2^s\Big(BN\big(ReLU(\boldsymbol{W}_1^s\hat{\boldsymbol{s}} + \boldsymbol{b}_1^s)\big)\Big) + \boldsymbol{b}_2^s \tag{2.1}$$

Here, the dimension of the projected sentence representation $\boldsymbol{s}$ is kept consistent with the projected moment representation $\boldsymbol{m}$ ($\boldsymbol{m}, \boldsymbol{s} \in \mathcal{R}^d$).

2.3.6 Learning Joint Embedding Space

Projected representations in the joint embedding space from different modalities need to be close to each other if they are semantically related. Training procedures to learn projected representations in the joint embedding space mostly adopts two common loss

functions: sum-margin based triplet ranking loss [43] and max-margin based triplet ranking loss [37]. We consider both of these loss functions separately. As illustrated in Figure 2.3, we focus on distinguishing intra-video moments and inter-video global semantic concepts. In this section, we discuss our approach to learn projecting representations from different modalities in the joint embedding space for multimodal data.

Similarity Measure. We use the cosine similarity of projected representations from two modalities in the joint embedding space to infer their semantic relatedness. So, the similarity between a candidate moment m and a sentence s is,

$$S(m, s) = \frac{\boldsymbol{m}^T \boldsymbol{s}}{\|\boldsymbol{m}\| \|\boldsymbol{s}\|} \tag{2.2}$$

where $\boldsymbol{m}$ and $\boldsymbol{s}$ are the projected moment representation and sentence representation in the joint embedding space.

Learning for Intra-video Moments. To localize a sentence query in a video, the model needs to identify the subtle differences of the candidate moments from the same video and distinguish them. Among the candidate segments of a video, one or few of the moments can be considered related to the query sentence based on some IoU threshold. While training the network, we consider related moments with the queried sentence as the positive pairs and non-corresponding moments with the queried sentence as the negative pairs. Suppose, for a pair of video-sentence (v, s), we consider the set of positive moment-sentence pairs $\{(m, s)\}$ and the set of negative moment-sentence pairs $\{(m^-, s)\}$. We compute the intra-video ranking loss for all video-sentence pairs $\{(v, s)\}$. Using the sum-margin setup, the intra-video triplet loss is:

$$\mathcal{L}_{sum}^{intra} = \sum_{\{(v,s)\}} \sum_{\{(m,s)\}} \sum_{\{(m^-,s)\}} \left[\alpha_{intra} - S(m,s) + S(m^-,s)\right]_+ \tag{2.3}$$

Similarly, using the max-margin setup, we calculate the intra-video triplet loss by,

$$\hat{m} = \arg\max_{m^-} S(m^-,s) \tag{2.4}$$

$$\mathcal{L}_{max}^{intra} = \sum_{\{(v,s)\}} \sum_{\{(m,s)\}} \left[\alpha_{intra} - S(m,s) + S(\hat{m},s)\right]_+ \tag{2.5}$$

Here, $[f]_+ = max(0,f)$ and α_{intra} is the ranking loss margin for intra-video moments.

Learning for Inter-video Moments. Learning to distinguish intra-video moments only allows the model to learn subtle changes in the video. It does not allow the model to distinguish moments from different videos. However, learning to differentiate moments from different videos is important as we need to localize the correct moment in the video corpus. Hence, we also learn to distinguish moments from different videos by capitalizing on the text-guided global semantics of videos. As the global semantics varies across videos we try to distinguish videos based on these global semantics. To do so, we learn to maximize the relevance of correct video-sentence pairs. Video-sentence relevance is computed in terms of moment-sentence relevance. As a result, learning to align video-sentence pairs enforces constraints on the representation of moments from different videos to be dissimilar. Inspired by the work of [72], we compute the relevance of a video and a sentence by,

$$R(v,s) = \log\Big(\sum_{\{m\}} \exp\big(\beta S(m,s)\big)\Big)^{1/\beta}, \tag{2.6}$$

where β is a factor that determines how much to magnify the importance of the most relevant moment-sentence pair and $\{m\}$ is the set of all the moments in video v. As $\beta \to \infty$, $R(v, s)$ approximates $\max_{m_i \in v} S(m_i, s)$. This is necessary because all the segments of the video do not correspond to the sentence.

For each positive video-sentence pair (v, s) where the sentence s relates to a segment of the video v, we can consider two sets of negative pairs $\{(v^-, s)\}$ and $\{(v, s^-)\}$. Using the sum-margin setup, we calculate the triplet loss for video-sentence alignment of all the positive video-sentence pairs $\{(v, s)\}$ by,

$$\begin{aligned}\mathcal{L}_{sum}^{video} = &\sum_{\{(v,s)\}} \sum_{\{(v^-,s)\}} \left[\alpha_{video} - R(v,s) + R(v^-,s)\right]_+ \\ &+ \sum_{\{(v,s)\}} \sum_{\{(v,s^-)\}} \left[\alpha_{video} - R(v,s) + R(v,s^-)\right]_+ \end{aligned} \tag{2.7}$$

Similarly, using the max-margin setup, we compute the triplet loss for video-sentence alignment by,

$$\hat{v} = \arg\max_{v^-} R(v^-, s) \tag{2.8}$$

$$\hat{s} = \arg\max_{s^-} R(v, s^-) \tag{2.9}$$

$$\begin{aligned}\mathcal{L}_{max}^{video} = &\sum_{\{(v,s)\}} \left[\alpha_{video} - R(v,s) + R(\hat{v},s)\right]_+ \\ &+ \sum_{\{(v,s)\}} \left[\alpha_{video} - R(v,s) + R(v,\hat{s})\right]_+ \end{aligned} \tag{2.10}$$

Here, α_{video} is the ranking loss margin for learning inter-video global semantic concepts.

Table 2.1: Tabulated summary of the details of dataset contents

Dataset	Number of videos		Moment-sentence pairs
	Total	Train/Val/Test	
DiDeMo	10464	8395 / 1065 / 1004	26892
Charades-STA	6670	5336 / - / 1334	16128
ActivityNet Captions	20k	10009 / 4917 / -	71942

Overall Learning Objective. We combine the calculated loss for intra-video case and video-sentence alignment case and try to minimize it as our final objective. For the sum-margin setup, the final objective is,

$$\min_{\theta} \mathcal{L}_{sum}^{intra} + \lambda_1 \mathcal{L}_{sum}^{video} + \alpha \|\mathcal{W}\|_F^2 \tag{2.11}$$

Similarly, for the max-margin setup, the final objective is,

$$\min_{\theta} \mathcal{L}_{max}^{intra} + \lambda_1 \mathcal{L}_{max}^{video} + \alpha \|\mathcal{W}\|_F^2 \tag{2.12}$$

Here, θ represents the network weights and all the learnable weights are lumped together in $\mathcal{W}$. λ_1 balances the contribution between learning to distinguish intra-video moments and learning to distinguish videos based on a text query. α is the weight on the regularization loss. Our objective is to optimize θ to generate a proper representation for candidate moments and sentences to minimize these combined losses. During training, these losses are computed for a mini-batch where the mini-batches are sampled randomly from the entire training set. This stochastic approach yields the advantage of reducing the probability of selecting instances with high semantic relation as the negative samples.

Table 2.2: Tabulated summary of the implementation details regarding video processing for three datasets

Dataset	Video length	# of candidate moments	Per Unit duration	Temporal dimension of layers
DiDeMo	12	21	2.5s	{6,5,4,3,2,1}
Charades-STA	64	61	1s	{31,16,8,4,2,1}
ActivityNet Captions	512	1023	1s	{512, 256, 128, 64, 32, 16, 8, 4 ,2, 1}

Table 2.3: Comparison of performance for the task of temporally localizing moments in a video corpus on DiDeMo dataset. († reported from [36]) (↓ indicates the performance is better if the score is low)

	Feature used	DiDeMo					
		$IoU = 0.50$			$IoU = 0.70$		
		R@10	R@100	MR↓	R@10	R@100	MR↓
Moment Prior† [36]	-	0.22	2.34	2527	0.17	1.99	3234
MCN† [5]	RGB (ResNet-152)	2.15	12.47	1057	1.55	9.03	1423
SCDM [177]	RGB (ResNet-152) + Flow (TSN)	0.57	4.43	-	0.22	1.42	-
VSE++ [37] + SCDM [177]	RGB (ResNet-152) + Flow (TSN)	0.70	4.16	-	0.30	2.81	-
CAL† [36]	RGB (ResNet-152)	3.90	16.51	831	2.81	12.79	1148
HMAN (sum-margin, Eqn. 2.11)	RGB (ResNet-152)	5.63	26.49	412	4.51	20.82	546
HMAN (TripSiam [35])	RGB (ResNet-152) + Flow (TSN)	2.34	17.82	509	1.59	13.92	637
HMAN (DSLT [92])	RGB (ResNet-152) + Flow (TSN)	5.95	25.45	313	4.66	20.04	447
HMAN (sum-margin, Eqn. 2.11)	RGB (ResNet-152) + Flow (TSN)	**6.25**	**28.39**	**302**	**4.98**	**22.51**	**416**
HMAN (max-margin, Eqn. 2.12)	RGB (ResNet-152) + Flow (TSN)	5.47	20.82	618	3.86	16.28	905

Inference. In the inference step, for a query sentence, we compute the similarity of candidate moment representations with the query sentence representation using Eqn. 2.2. We retrieve the candidate moment from the video corpus that results in the highest similarity.

2.4 Experiments

In this section, we first discuss the datasets we use and the implementation details of the experiments. Then we report and analyze the results both quantitatively and qualitatively.

Table 2.4: Comparison of performance for the task of temporally localizing moments in a video corpus on Charades-STA dataset. († reported from [36]) (↓ indicates the performance is better if the score is low)

	Feature used	Charades-STA					
		IoU = 0.50			*IoU* = 0.70		
		R@10	R@100	MR↓	R@10	R@100	MR↓
Moment Prior† [36]	-	0.17	1.63	4906	0.05	0.56	11699
MCN† [5]	RGB (ResNet-152)	0.52	2.96	6540	0.31	1.75	10262
SCDM [177]	RGB (I3D)	0.73	6.41	-	0.56	4.23	-
VSE++ [37] + SCDM [177]	RGB (I3D)	1.02	5.06	-	0.70	3.37	-
CAL† [36]	RGB (ResNet-152)	0.75	4.39	5486	0.42	2.78	8627
HMAN (TripSiam [35])	RGB (I3D)	1.27	7.60	2821	0.70	4.49	5766
HMAN (DSLT [92])	RGB (I3D)	1.05	7.27	2390	0.54	4.61	**5496**
HMAN (sum-margin, Eqn. 2.11)	RGB (I3D)	1.29	7.73	2418	0.83	4.12	6395
HMAN (max-margin, Eqn. 2.12)	RGB (I3D)	**1.40**	**7.79**	**2183**	**1.05**	**4.69**	5812

2.4.1 Datasets

We conduct experiments and evaluate the performance on three benchmark text-based video moment retrieval datasets, namely DiDeMo [5], Charades-STA [47], and ActivityNet Captions [67]. All of these datasets contain unsegmented and untrimmed videos with natural language sentence annotations with temporal information. Table 2.1 summarizes the details of the contents of three datasets.

DiDeMo. The Distinct Describable Moments (DiDeMo) dataset [5] is one of the most diverse datasets for the temporal localization of moments in videos given natural language descriptions. The videos are collected from Flickr and each video is trimmed to a maximum of 30 seconds. The videos in the dataset are divided into 5-second segments to reduce the complexity of annotation. The dataset is split into training, validation, and test sets containing 8,395, 1,065, and 1,004 videos respectively. The dataset contains a total of 26,892 moment-sentence pairs and each natural language description is temporally grounded by multiple annotators.

Charades-STA. Charades-STA dataset is introduced in [47] to address the task of temporal localization of moments in untrimmed videos. The dataset contains a total of 6,670 videos with 16,128 moment-sentence pairs. We have used the published split of videos during training and testing (train-5,336, test-1,334). As a result, the training set and the testing set contain 12,408 and 3,720 moment-sentence pairs respectively. This dataset is originally built on the Charades [131] activity dataset with temporal activity annotation and video-level description. Authors in [47] adopted a keyword matching strategy to generate clip-level sentence annotation.

ActivityNet Captions. ActivityNet Captions [67] dataset, which is proposed for dense video captioning task, is built on the ActivityNet dataset [55]. It consists of YouTube video footage where each video contains at least two ground truth segments and each segment is paired with one ground truth caption [162]. This dataset contains around 20k videos which are split into training, validation, and testing set. We use the published splits over videos (train set – 10,009 videos, validation set – 4,917 videos), where the evaluation is done on the validation set. Videos are typically longer in length than DiDeMo and Charades-STA datasets.

2.4.2 Evaluation Metric

We use the standard evaluation criteria adopted by various previous temporal moment localization works [47, 177, 182]. These works use $R@k, IoU{=}m$ metric, which reports the percentage of cases where at least one of the top-k results have Intersection-over-Union (IoU) larger than m [47]. For a sentence query, $R@k, IoU{=}m$ reflects if one of the top-k retrieved moments has Intersection-over-Union with the ground truth moment larger

Table 2.5: Comparison of performance for the task of temporally localizing moments in a video corpus on ActivityNet Captions dataset. († reported from [36])

	Feature used	ActivityNet Captions			
		$IoU = 0.50$		$IoU = 0.70$	
		R@10	R@100	R@10	R@100
Moment Prior†	-	0.05	0.47	0.03	0.26
MCN† [5]	RGB (ResNet-152)	0.18	1.26	0.09	0.70
CAL† [36]	RGB (ResNet-152)	0.21	1.58	0.10	0.90
HMAN (sum)	RGB (C3D)	0.43	2.84	0.22	1.48
HMAN (max)	RGB (C3D)	**0.66**	**4.75**	**0.32**	**2.27**

Table 2.6: Comparison of the performance of HMAN with/without the Hierarchical moment Encoder Module. The experiments are done for DiDeMo and Charades-STA datasets. († reported from [36]) (↓ indicates the performance is better if the score is low)

	DiDeMo						Charades-STA					
	$IoU = 0.50$			$IoU = 0.70$			$IoU = 0.50$			$IoU = 0.70$		
	R@10	R@100	MR↓	R@10	R@100	MR↓	R@10	R@100	MR↓	R@10	R@100	MR↓
HMAN (sum, w/o TCN)	3.44	14.14	1168	2.14	9.91	1636	1.13	6.12	4170	0.43	4.09	8295
HMAN (sum, w/ TCN)	**6.25**	**28.39**	**302**	**4.98**	**22.51**	**416**	**1.29**	**7.73**	**2418**	**0.83**	**4.12**	**6395**
HMAN (max, w/o TCN)	3.41	12.13	1603	1.99	8.96	2214	0.70	4.71	5800	0.46	3.13	10907
HMAN (max, w/ TCN)	**5.47**	**20.82**	**618**	**3.86**	**16.28**	**905**	**1.40**	**7.79**	**2183**	**1.05**	**4.69**	**5812**

than the specified threshold m. So, for each query sentence, $R@k, IoU{=}m$ is either 1 or 0. As this metric is associated with a queried sentence, we compute it for all the sentence queries in the testing set (DiDeMo, Charades-STA) or in the validation set (ActivityNet Captions) and report the average results. We report $R@k, IoU{=}m$ over all queried sentences for $k \in \{10, 100\}$ and $m \in \{0.50, 0.70\}$. We also use median retrieval rank (MR) as an evaluation metric. MR computes the median of the rank of the correct moment for each query. Lower values of MR indicate good performance. We compute MR for $IoU \in \{0.50, 0.70\}$. Note that DiDeMo dataset provides multiple temporal annotations for each sentence. We consider a detection is correct if it overlaps with a minimum of two temporal annotations with a specified IoU.

Table 2.7: Ablation study for the effectiveness of learning embedding space utilizing different loss components as described in 2.3.6 for DiDeMo dataset using sum-margin set up.

	$IoU = 0.50$		$IoU = 0.70$	
	R@10	R@100	R@10	R@100
HMAN (intra)	0.57	6.00	0.52	4.71
HMAN (video)	1.77	10.03	0.30	2.34
HMAN (proposed)	**6.25**	**28.39**	**4.98**	**22.51**

2.4.3 Implementation Details

For DiDeMo dataset, we use ResNet-152 features [54], where pool5 features are extracted at 5 fps over the video frames. Then the features are max-pooled over $2.5s$ clips. Also, we extract optical flow features from the penultimate layer from a competitive activity recognition model [149]. We use Kinetics pretrained I3D network [11] to extract per second clip features for the Charades-STA dataset. For ActivityNet Captions dataset, we use extracted C3D features [144]. We set the number of input clips of a video, $l = 12$ for DiDeMo dataset, $l = 64$ for Charades-STA dataset, and $l = 512$ for ActivityNet Captions dataset. Here, per unit length of input video represents non-overlapping clip of $2.5s$ duration for DiDeMo and non-overlapping clip of $1s$ duration for both Charades-STA and ActivityNet Captions dataset. For DiDeMo dataset, we use a fully connected layer followed by max-pool to generate representations with temporal dimension 6 for each video. Then we use 6 temporal convolutional layers to generate representations with temporal dimensions of $\{6, 5, 4, 3, 2, 1\}$ resulting in representations for 21 candidate moments. Similarly for Charades-STA, we use a fully connected layer followed by max-pool to generate representations with temporal dimension 32 for each video. Then we use 6 temporal convolutional layers with the temporal dimension of $\{32, 16, 8, 4, 2, 1\}$ where we use the 31 candidate moment

Table 2.8: Performance comparison for the task of retrieving correct video based on sentence query on DiDeMo and Charades-STA dataset.

	DiDeMo			Charades-STA		
	R@10	R@100	R@200	R@10	R@100	R@200
VSE++ [37]	2.49	16.81	29.53	1.89	13.31	24.43
HMAN (max)	12.43	42.43	58.22	2.26	15.87	27.26
HMAN (sum)	**15.36**	**55.23**	**69.12**	**2.45**	**18.51**	**30.52**

Table 2.9: Comparison of the performance of proposed LogSumExp pooling and average pooling. We compare the performance for the task of temporal localization of moments in video corpus for DiDeMo and Charades-STA dataset.

	DiDeMo				Charades-STA			
	$IoU = 0.50$		$IoU = 0.70$		$IoU = 0.50$		$IoU = 0.70$	
	R@10	R@100	R@10	R@100	R@10	R@100	R@10	R@100
HMAN (sum, ave)	5.63	26.05	4.43	20.82	1.10	7.19	0.62	**4.47**
HMAN (sum, log)	**6.25**	**28.39**	**4.98**	**22.51**	**1.29**	**7.73**	**0.83**	4.12
HMAN (max, ave)	5.27	17.65	**4.01**	13.60	0.75	7.00	0.51	4.53
HMAN (max, log)	**5.47**	**20.82**	3.86	**16.28**	**1.40**	**7.79**	**1.05**	**4.69**

representations from the last 5 layers. Additionally, we use a branch temporal convolutional layer to generate representations of 30 overlapping candidate moments, each with $6s$ duration and $2s$ stride. Combining these, we consider 61 candidate moments for each video of Charades-STA dataset. For ActivityNet Captions dataset, we use a feedforward network followed by 10 convolutional layers to generate representations with temporal dimension of $\{512, 256, 128, 64, 32, 16, 8, 4, 2, 1\}$, resulting in 1023 candidate moment representations. Table 2.2 illustrates the implementation details for video processing for all three datasets. we consider sentences with maximum of 15 words in length. If a sentence contains more than 15 words, the tailing words are truncated.

The proposed network is implemented in TensorFlow and trained using a single RTX

Table 2.10: Ablation Study of the performance of HMAN (sum-margin) for Different Visual Features for DiDeMo dataset.

	$IoU = 0.50$		$IoU = 0.70$	
	R@10	R@100	R@10	R@100
VGGNet	2.61	16.36	1.79	12.82
VGGNet + Flow	3.98	21.29	3.14	16.76
ResNet	5.63	26.49	4.51	20.82
ResNet + Flow	**6.25**	**28.39**	**4.98**	**22.51**

2080 GPU. To train the HMAN network, we use mini-batches containing 64 video-sentence pairs for DiDeMo and Charades-STA and 32 video-sentence pairs for ActivityNet Captions. We use the learning rate with exponential decay initializing from 10^{-3} for all three datasets. ADAM optimizer is used to train the network. We use 0.9 as the exponential decay rate for the first moment estimates and 0.999 as the exponential decay rate for the second-moment estimates. We set α_{intra} and α_{video} to 0.05 and 0.20, respectively for all three datasets. λ_1 is empirically set to 5, 1, and 1.5, respectively for DiDeMo, Charades-STA, and ActivityNet Captions. α is set to 5×10^{-5} for all three datasets.

2.4.4 Analysis of Results

We conduct the following experiments to evaluate the performance of our proposed method:

- Comparison of the performance of proposed HMAN for the task of temporal localization of moments in video corpus with different baseline approaches and a concurrent work.
- Evaluation of the effectiveness of utilizing hierarchical moment encoder module.
- Investigation of the impact of learning joint embedding space by utilizing different com-

ponents of the loss function (learning for intra-video moments ($\mathcal{L}^{intra}$) and learning for videos ($\mathcal{L}^{video}$).

- Evaluation of the effectiveness of utilizing global semantics to identify the correct video.
- Analyzing the effectiveness of video relevance computation (Eqn. 2.6) for the task of temporal localization of moments in a video corpus.
- Studying the performance of proposed HMAN for different visual features.
- Performance comparison of HMAN with decreasing number of test set moment-sentence pairs.
- Evaluation of the run time efficiency.
- Analysis of the λ_1 parameter sensitivity.

Temporal Localization of Moments in Video Corpus. Table 2.3, Table 2.4, and Table 2.5 illustrate the quantitative performance of our framework for the task of temporal localization of moments in the video corpus. The evaluation setup considers $IoU \in \{0.50, 0.70\}$ and for each IoU threshold, we report R@10, R@100 and MR. For a query sentence, the task requires to search over all the videos and retrieve the relevant moment. For example, in the DiDeMo dataset, the test set consists of 1,004 videos totaling 4,016 moment-sentence pairs. Again, we consider 21 candidate moments for each video. So, for each query sentence, we need to search over $21 \times 1{,}004 = 21{,}084$ moment instances and retrieve the correct moment. This is itself a difficult task and the addition of ambiguity of similar kinds of activities in different videos makes the problem even harder. We compare the proposed method with the following baselines:

Table 2.11: Ablation study of the performance of HMAN (sum-margin) on DiDeMo when the number of test set data is decreased.

	$IoU = 0.50$			$IoU = 0.70$		
	R@10	R@100	MR↓	R@10	R@100	MR↓
HMAN (100%)	6.25	28.39	302	4.98	22.51	416
HMAN (50%)	6.90	30.15	268	5.68	23.73	372
HMAN (25%)	8.74	34.93	193	7.06	27.62	269
HMAN (10%)	13.35	45.60	102	10.30	36.65	142

- **Moment Frequency Prior:** We use Moment Frequency Prior baseline from [5], which selects moments that correspond to gifs most frequently described by the annotators.

- **MCN:** The Moment Context Network [5] for temporal localization of moments in a given video is scaled up to search moment from the entire video corpus.

- **SCDM:** The state-of-the-art Semantic Conditioned Dynamic Modulation (SCDM) network [177] for temporal localization of moments in a video is scaled up to search over the entire video corpus.

- **VSE++ + SCDM:** We use joint embedding based retrieval approach (VSE++ [37]) combined with SCDM as a baseline. In this setup, the framework first retrieves a few relevant videos (top 5%) and then localize moments on those retrieved videos using SCDM approach.

- **CAL:** We compare with Clip Alignment of Language [36]. It is a concurrent work that addresses the task of localizing moments in a video corpus by aligning clip representation with language representation in the embedding space.

Note that we do not compare with baselines that utilize temporal endpoint features from [5], as these directly correspond to dataset priors and do not reflect a model's capability [84].

We observe that MCN and CAL perform better than the state-of-the-art SCDM approach in DiDeMo dataset but perform poorly compared to the SCDM approach in Charades-STA dataset. This is due to the fact that the video contents and language queries differ a lot among different datasets [182]. MCN and CAL learn to distinguish both intra-video moments and inter-video moments locally while SCDM only learns to distinguish intra-video moments. As DiDeMo dataset contains diverse videos of different concepts and relatively less number of candidate moments, learning to differentiate inter-video moments locally improves performance significantly. However, learning to differentiate inter-video moments locally does not have much impact on Charades-STA dataset. This also indicates the importance of distinguishing moments from different videos based on global semantics for a diverse set of video datasets. We also observe that in some of the cases, VSE++ + SCDM scores drop compared to the SCDM approach. Since the performance of VSE++ + SCDM depends on retrieving correct video, the localization performance drops if the retrieval approach fails to retrieve correct videos with higher accuracy.

For HMAN, we report the performance for both sum-margin and max-margin based triplet loss setups. Additionally, for DiDeMo and Charades-STA dataset, we report the performance of HMAN for two different loss calculation setups: TripSiam [35] and DSLT [92]. In Table 2.3, Compared to baseline approaches, the performance of our proposed approach is better for all metrics and outperforms other approaches with a maximum of 11.88%

Table 2.12: Per epoch training and inference time for Charades-STA dataset.

Approach	Training time	Inference time
Sliding-based	35.05 s	90.46 s
HMAN	21.18 s	83.91 s

absolute improvement in DiDeMo dataset. We observe that the sum-margin based triplet loss setup outperforms the max-margin setup, while both of these setups perform better than other baselines in DiDeMo dataset. For a fair comparison with CAL and MCN, we report the performance of HMAN with the ResNet-152 feature computed from RGB frames only. This setup also outperforms CAL and MCN. We also conduct experiment incorporating temporal end point feature in HMAN for DiDeMo dataset. It results in $\sim 0.5\% - 1\%$ improvement over HMAN (sum-margin) in $R@k$ metrics. It indicates the bias in the dataset where different types of events are correlated with different time frames of the video. In Table 2.4, for the Charades-STA dataset, the performance of HMAN is better for all metrics and the max-margin based triplet loss setup outperforms other baseline approaches with a maximum of 3.4% absolute improvement. In Table 2.5, for ActivityNet Captions dataset, the HMAN max-margin setup outperforms other baselines with a maximum of 3.17% absolute improvement. We do not compute SCDM and VSE++ + SCDM baselines for ActivityNet Captions dataset. Moment representations in SCDM and VSE++ + SCDM approaches are conditioned on sentence queries. For each query sentence, we need to compute moment representations from all the videos, resulting in $O(n^2)$ complexity. So testing on a set of 34,160 query sentences and $4{,}917 \times 1{,}023 = 5{,}030{,}091$ moment representations is impractical using these approaches.

TripSiam [35] and DSLT [92] are two different variants of triplet loss which are used in object tracking. TripSiam defines a matching probability for each triplet to measure the possibility of assigning the positive instance to exemplar compared with the negative instance and tries to maximize the joint probability among all triplets during training. DSLT [92] utilizes modulating function to minimize the contribution of easy samples in the total loss. While both setups perform better than baseline approaches, we observe that there is a significant improvement in median retrieval rank (MR). This indicates that even if TripSiam and DSLT can not retrieve the correct moment, they are robust in terms of the semantic association between moments and sentences.

Effectiveness of Hierarchical Moment Encoder. HMAN utilizes stacked temporal convolutional layers in a hierarchical structure to represent video moments. We conduct experiments to analyze the effects of using the hierarchical moment encoder module in our proposed model. We consider two setups, i) **w/ TCN**: the hierarchical moment encoder module built using temporal convolutional network is present in the model and ii) **w/o TCN**: the hierarchical moment encoder module is replaced with a simple feedforward network to project the candidate moment representations in the joint embedding space. We consider both sum-margin based and max-margin based triplet loss to train the networks. Table 2.6 illustrates the effect of utilizing hierarchical moment encoder module. We observe that for both the learning approaches and for both datasets, there is a significant improvement in performance when the hierarchical moment encoder module is used. For example, in DiDeMo dataset, we observe $\sim 14\%$ (sum-margin) and $\sim 8\%$ (max-margin) absolute improvement in performance for $R@100, IoU = 0.50$.

Ablation Study of Learning Joint Embedding Space. We conduct experiments to analyze the impact of different components of the loss function to learn the joint embedding space for our targeted task in DiDeMo dataset and reported the results in Table 2.7. We use three setups to learn the joint embedding space:

- **HMAN (intra)**: Only uses $\mathcal{L}^{intra}$. So the network only learns to distinguish intra-video moments.
- **HMAN (video)**: Only uses $\mathcal{L}^{video}$. So the network only learns to disntinguish moments from different videos based on global semantics.
- **HMAN (proposed)**: Our proposed approach, combination of $\mathcal{L}^{intra}$ and $\mathcal{L}^{video}$.

In Table 2.7, we observe that the performance of HMAN is poor for both the case of HMAN (intra) and HMAN (video). Performance of HMAN (intra) is better compared to HMAN (video) in Table 2.7 when higher IoU threshold requirement is considered $(R@k, IoU = 0.7)$. This indicates that HMAN (intra) learns to better identify temporal boundaries in a video compared to HMAN (video), while HMAN (video) is better at distinguishing moments from different videos compared to HMAN (intra). However, when we combine both of these criteria, there is a significant performance boost as the model is able to effectively learn to identify both the correct video and the temporal boundary. All the results in Table 2.7 are reported for sum-margin based triplet loss setup.

Effectiveness of Utilizing Global Semantics. Our proposed learning objective utilizes global semantics to distinguish moments from different videos. To do so, we learn to align corresponding video-sentence pairs, where the video-sentence relevance $R(v, s)$ in the

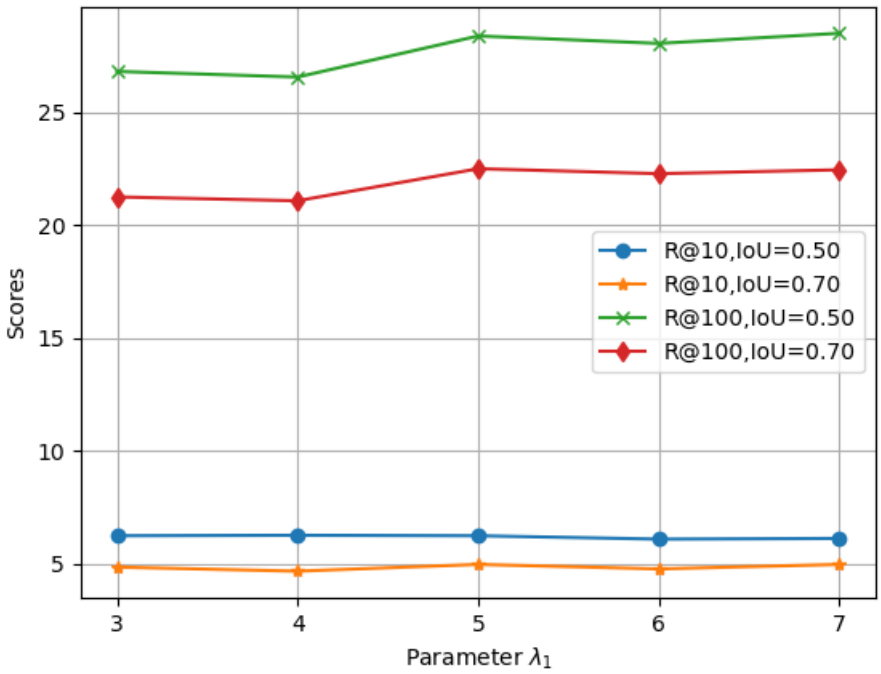


Figure 2.4: Illustration of λ_1 sensitivity on the HMAN performance. We observe that for the set of values $\{3, 4, 5, 6, 7\}$, performance of HMAN is stable.

embedding space is computed in terms of moment-sentence similarity $S(m, s)$. So we use this video-sentence relevance score $R(v, s)$ to analyse the models performance to identify or retrieve the correct video given a text query and report the results in Table 2.8. We use the standard evaluation criteria $R@k$ for video retrieval task and report $R@10$, $R@100$, and $R@200$ scores for DiDeMo and Charades-STA dataset. Here, $R@K$ calculates the percentage of query sentences for which the correct video is found in the top-K retrieved videos to the query sentence. In DiDeMo test set, there are 1,004 videos with 4,016 moment-sentence pairs ($\sim$ 4 sentences per video) and in Charades-STA testset, there are 1,334 videos with 3,720 moment-sentence pairs ($\sim$ 2.8 sentences per video). Due to the one-to-many correspondences, we consider 4,016 and 3,720 video-sentence pairs respectively for DiDeMo and Charades-STA datasets for the video retrieval task, where a single video can pair up with multiple sentences.

Table 2.8 shows that both sum-margin (HMAN (sum)) and max-margin (HMAN (max)) based triplet loss setups of our proposed approach outperforms standard Visual Semantic Embedding based retrieval approach (VSE++) for the task of retrieving the correct video. Along with the consistent improvement of performance in all metrics for both datasets, We observe $\sim$ 40% absolute improvement of retrieval performance for the metric R@200 for DiDeMo dataset. As the video-sentence relevance is computed in terms of moment-sentence similarity, this experiment validates the models capability to distinguish videos as well as moments from different videos utilizing global semantics.

Analysis of Video Relevance Computation Approach. In an untrimmed video with temporal language annotation, the segment/portion of the video mostly matches with the sentence semantics. So to compute the video-sentence relevance, it needs to focus on the moments that have higher similarity with the query sentence semantics. To tackle this issue, we compute the video-sentence relevance using LogSumExp pooling (Eqn. 2.6) of the moment-sentence similarity. In Table 2.9, we analyze the significance of the LogSumExp pooling compared to average pooling for both sum-margin and max margin based triplet loss setups. In Table 2.9, 'ave' and 'log' indicates average and LogSumExp pooling respectively, while 'sum' and 'max' indicates sum-margin based and max-margin based triplet loss respectively. For both DiDeMo and Charades-STA datasets, we observe that LogSumExp pooling performs better than average pooling for the target task of temporal localization of moments in video corpus in both sum-margin based and max-margin based triplet loss setups.

Ablation Study of Different Visual Features. We conduct experiments to study the performance of HMAN for different visual features for DiDeMo dataset and reported the

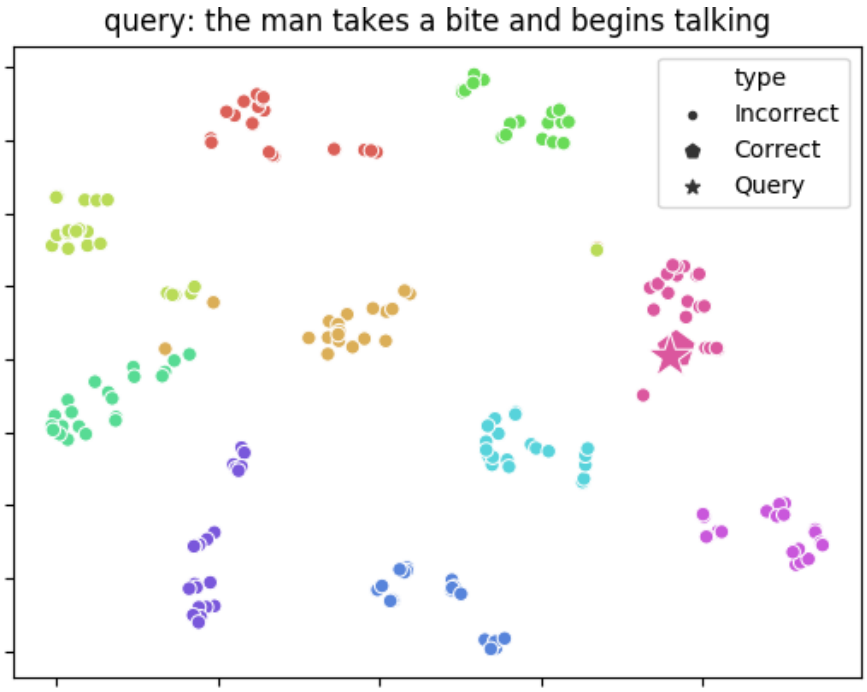


Figure 2.5: t-SNE visualization of text query representation and candidate moment representations. Different color represents different video. The color of the text representation is the same as the corresponding video. We use different markers for the representation of incorrect candidate moments, correct candidate moments and text. Here, representations of the text query and the correct candidate moment coincide. Also, the representations of candidate moments from the same video are clustered together.

results in Table 2.10. We use extracted features from VGGNet [132], ResNet-152 [54] for RGB frames and optical flow features from [149]. In Table 2.10, we observe that a combination of RGB and optical flow features perform better than using only an RGB stream. It indicates the models increased capacity due to the increase in the number of learnable weights. As a result, HMAN is suitable to work with multiple encodings of the same data together compared to the shallow embedding networks [5, 36]. We have reported the results for sum-margin based triplet loss setup.

Performance of HMAN on Decreased Number of Moment-sentence Pairs. Since HMAN searches for the correct candidate moment across all the videos in the test set during inference, the temporal localization performance of HMAN is expected to improve by decreasing the number of moment-sentence pairs in the test set. We conduct experiments on DiDeMo dataset to evaluate the performance of HMAN (learned using sum-margin based triplet loss) on the decreased number of moment-sentence pairs in the test phase. We consider four setups: **HMAN (100%)**: Model searches over the full test set during inference, **HMAN (50%)**: Model searches over each 50% of the test set separately and take the average of the scores, **HMAN (25%)**: Model searches over each 25% of test set separately and take the average of the scores, **HMAN (10%)**: Model searches over each 10% of test set separately and take the average of the scores. Table 2.11 illustrates the performance for all four setups. We observe that with decreased number of test set moment-sentence pairs, the performance of HMAN improves.

Evaluation of Run Time Efficiency. We conduct experiments on the Charades-STA dataset to compare the run time of HMAN with the sliding window-based approaches. The differences in the sliding-based approach compared to the setup of HMAN is that: i) the moment encoder module with temporal convolutional network of HMAN is replaced by a simple single layer feedforward network, ii) instead of generating candidate moment representations directly from the video, we slide over the video to extract features of different temporal durations, then use extracted features to generate candidate moment representations. Table 2.12 illustrates that for both training case and inference case, the sliding-based approach takes longer than HMAN per epoch, even though the network is much smaller in the sliding-

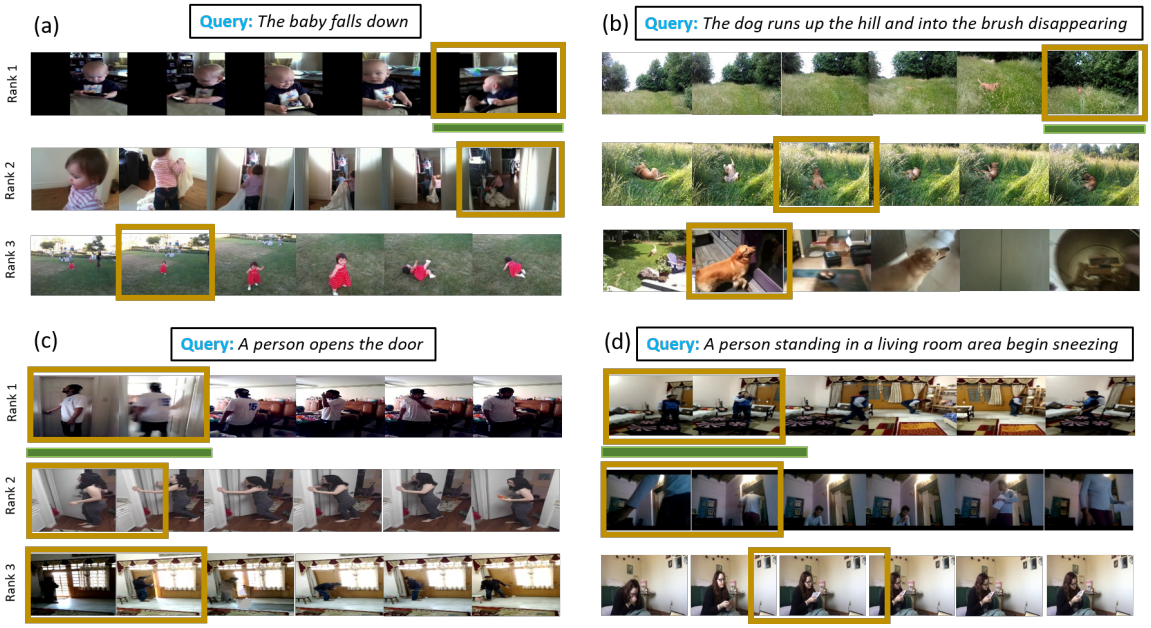


Figure 2.6: Example illustration of the performance of HMAN for the task of localization of moments in a corpus of videos. For each query sentence, we display the top-3 retrieved moments. The retrieved moments are surrounded by gold boxes and the ground truth moments are indicated by green lines. We observe that for each of the queries, the top-3 retrieved moments are semantically related to the sentence proving the efficacy of our approach.

based approach compared to HMAN. For a fair comparison, we keep the number of candidate moments the same, and similar computations (apart from hierarchical moment encoder module replaced by single layer feed forward network) are done for both the approaches. We have computed the run time for five epochs and reported the average results. Here, the inference time is higher due to the added requirement of computing the cosine distance between each text query and all the candidate moment representations.

λ_1 Parameter Sensitivity Analysis. In our framework, λ_1 balances the contribution of $\mathcal{L}^{intra}$ and $\mathcal{L}^{video}$ for both sum-margin and max-margin case. We choose the value of λ_1 empirically. We conduct an experiment to check the sensitivity of HMAN performance based on a set of values for λ_1 in the DiDeMo dataset where $\lambda_1 \in \{3, 4, 5, 6, 7\}$. In Figure 2.4 shows that for this set of values of λ_1, the performance is stable.

2.4.5 Qualitative Results

t-SNE Visualization. We provide t-SNE visualization of embedding representations of text query and candidate moments in Figure 2.5. For a text query, we consider embedding representation of the text query, representations of candidate moments from the correct video, and representations of candidate moments from randomly picked 9 other videos and visualize the distribution of representations. In Figure 2.5, different color represents different videos. Each video has 21 candidate moments. We keep the color of the text query representation the same as the color of candidate moments representation from the correct video and use separate markers for correct candidate moment and text query representation. We observe that representations of the text query and the correct candidate moment coincide. Also, the representations of candidate moments from the same video are clustered together.

Example Illustration. In Figure 2.6, we illustrate some qualitative results for our proposed approach. The two examples in the top row are for the DiDeMo dataset and the two examples in the bottom row are for the Charades-STA dataset. For each query sentence, we demonstrate the examples where the network is able to retrieve the correct moment as the rank-1 from the test set videos. We also display rank-2 and rank-3 moments retrieved by the model for each query sentence. Figure 2.6(a) shows that for the query *'The baby falls down'*, the model was able to retrieve the correct moment with the highest matching. However, the interesting fact lies in the retrieved rank-2 and rank-3 moments. For the query *'The baby falls down'*, the retrieved rank-2 and rank-3 moments also contain activity of a baby, including a baby falling down. Similar results are observed for other examples for both datasets. For example, in Figure 2.6(b), for the query sentence *'A person opens the door'*, the model was able to retrieve the correct moment with the highest matching. However, all top-3 ranked moments contain activity related to a door. In the rank-2 moment, a person is opening a door and in the rank-3 moment, a person is fixing a door. Similarly, the top retrieved moments for a query of a dog running and hiding contain activities of a dog (Figure 2.6(b)) and top retrieved moments for a query of a person standing and sneezing contain standing activity and sneezing activity (Figure 2.6(d)). These results indicate the model's capability of retrieving moments with similar semantic concepts from the corpus of videos.

2.5 Conclusion

In this work, we explore an important and under-explored task of localizing moments in a video corpus based on text query. We adapt existing temporal localization of

moments approaches and video retrieval approaches for the proposed task and identified the shortcomings of those approaches. Towards addressing the challenging task, we propose Hierarchical Moment Alignment Network (HMAN), a novel neural network that effectively learns a joint embedding space for video moments and sentences to retrieve the matching moment based on semantic closeness in the embedding space. Our proposed learning objective allows the model to identify subtle changes of intra-video moments as well as distinguish inter-video moments utilizing text-guided global semantic concepts of videos. We adopt both sum-margin based and max-margin based triplet loss setups separately and achieve performance improvement over other baseline approaches in both setups. We experimentally validate the effectiveness of our proposed approach on three standard benchmark datasets.

Chapter 3

Temporal Localization of Novel Events

3.1 Introduction

Event localization in a long and untrimmed video is an important video analysis problem. Recently, there has been a surge of works that address the task of temporal grounding of text/sentence in untrimmed videos [47, 5, 177, 83, 180, 102]. Most of these works utilize a set of fully supervised training data containing videos, text descriptions, and temporal boundary annotations. These works try to optimize over a fixed set of events and queries (which we call seen events and seen queries) that are available during training. However, in a real-world dynamic environment, a system is expected to encounter *previously unseen events and queries*, as shown in Figure 3.1, and is required to *localize corresponding moments based on unseen text queries* in the videos. As a result, a system optimized over a fixed set of events is unlikely to generalize and perform well for unseen events. Moreover, as textual annotations are expensive and time consuming [101], it is impossible to collect videos of all possible events and textual descriptions and learn models with the collected data.

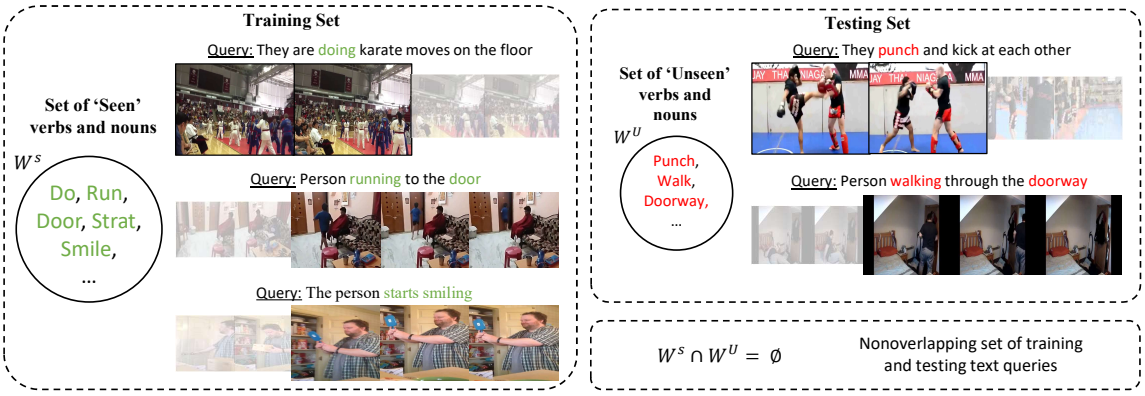


Figure 3.1: Example illustration of our proposed task. We consider the task of localizing novel moments for unseen queries. The set of verbs and nouns present in the testing set is absent in the training set, e.g., training data does not have any text with verb 'walk' or noun 'doorway'. Hence, the system is required to learn transferable knowledge from the training data to perform localization for novel events based on unseen queries.

Hence, the applicability of current text-based temporal localization systems are severely limited to a small set of events and the problem of localizing novel/unseen events based on unseen text queries remains unaddressed in the current literature.

In this work, our goal is to temporally localize video moments based on text queries, where both the video moments and text queries *are not observed/available during training*. Towards this goal, we learn transferable knowledge from seen events and queries and utilize it to localize novel/unseen events. We hypothesize that temporally relevant moments corresponding to unseen text queries and those corresponding to seen text queries are likely to contain shared concepts, if the unseen query and the seen query are semantically relevant. For instance, in Figure 3.1, moment corresponding to the unseen text query *'They punch and kick at each other'* from the testing set has similarities to the moment corresponding to seen text query *'They are doing karate moves on the floor'* from the training set. Therefore,

instead of localizing moments only based on its encoded representation, we formulate the inference task of localization as a relational prediction problem. The likelihood of a candidate moment to be the correct one based on an unseen text query depends on its relevance to the moment corresponding to the semantically most relevant seen query. We term this moment corresponding to the semantically most relevant seen query as the *support moment*. To learn a proper relational system that can localize novel events, we simulate the support moment based relational inference on the available training data during training. As a result, the system learns to localize moments based on relational reasoning, instead of directly localizing based on observed moment representations. Our motivation behind the approach is that a relational system learned on seen events/queries is transferable to the unseen events/queries [137]. We term our approach as **T**emporal **L**ocalization using **R**elational **R**easoning (**TLRR**).

Our problem is related to the zero-shot paradigm (where the objective is to adapt models to perform different tasks on the unseen or unobserved classes) as we utilize seen moment-text pairs to infer on the unseen events [160, 197, 185, 75, 108]. However, those zero-shot approaches are not directly applicable to our problem setup. For example, [186] assumes unseen classes are known in advance and uses the information to mine common semantics for seen classes and unseen classes for zero-shot temporal activity detection. However, text-based annotations of events are not limited to a fixed set of classes and the unseen queries are not known beforehand. Again, [80, 166, 26] perform retrieval across multiple modality data in the zero-shot setting. These works consider images with specific classes, and utilize the word embedding space to transfer knowledge between seen classes and unseen classes. However, in a video, textual descriptions refer to multiple entities, interactions of multiple entities,

and different activities in a combined manner that is not expressible by a single class. As a result, directly utilizing label embeddings is not enough to transfer knowledge from seen events/queries to unseen events/queries. We will demonstrate the advantage of our proposed TLRR approach over zero-shot approaches and other recent temporal localization approaches on two benchmark datasets. The following are the main **contributions** of our work.

- We address a novel and practical problem of temporal localization of video moments based on unseen text queries.
- We hypothesize a conceptual relation between semantically relevant moments and propose a relational reasoning based temporal localization approach, TLRR, which can learn transferable knowledge from seen events and localize novel events based on unseen text queries.
- We reorganize two existing text-based temporal localization datasets (Charades-STA [47] and ActivityNet Captions [67]) for our proposed novel problem setting. Empirical results on these two text-based video moment localization datasets show that our proposed approach can reach up to 15% absolute improvement in performance compared to existing localization approaches.

3.2 Related Works

Temporal Localization of Moments. Temporal localization of moments in a video based on text query was introduced by [47, 5]. Recently, there are many works that address the problem both in presence of strong supervision (temporal endpoints are known for

each query) [156, 89, 20, 49, 162, 182, 177, 63, 90, 189, 50, 178, 190, 188, 53, 51, 56, 84, 123, 180, 102, 21, 147, 87, 184, 164, 183, 62, 104, 142, 32, 187, 192, 148, 195, 86, 161, 135, 48, 114] and weak supervision (only video-text correspondence is known) [101, 82, 140, 169, 22, 152, 141]. Among the recent works on temporal localization of moments in the fully supervised setting, [177] performs semantic conditioned dynamic modulation, [180] relies on dense regression based approach, [102] utilizes both local and global interaction for video grounding. Recently, [103] proposed text-based temporal localization without query annotation. Unlike our setting, they have access to videos of all types of events and can optimize their model for such events in a weakly supervised manner. Hence, none of these works address the problem of localizing novel events based on unseen text queries.

Zero-shot Learning (ZSL). ZSL aims to do inference task on classes whose instances may not have been seen during training [160, 197, 185, 75, 108]. Initial works on ZSL were attribute-based [68, 111]. However, attribute-based ZSL has poor scalability and semantic embedding of labels are a good alternative for attributes [167]. Most of the works that utilize semantic embedding based learning focus on the association of visual and semantic information by linear compatibility [43, 1, 2, 126], non-linear compatibility [134, 159] or in a hybrid way [109]. To the best of our knowledge, only [186] works on activity detection in ZSL setup. However, [186] is limited to work on activity labels and can not be adapted directly for moment localization of unseen text queries.

Zero-shot Cross Modal Retrieval (ZS-CMR). Conventional cross modal retrieval work [34] considers similar type of events are present in both training set and testing set. However, ZS-CMR aims to perform retrieval across multiple modality data in the zero-shot setting.

They train the retrieval model with limited categories to support cross-modal retrieval on new categories [80]. There are few works that consider retrieval between visual and textual modality with ZS-CMR setting [80, 166, 26]. However, these works are limited by the use of specific class information of the images to transfer knowledge between seen classes to unseen classes.

Relational Reasoning. Relational reasoning concept has been applied to different vision applications, i.e., visual question answering [127, 122], deep reinforcement learning [179], few-shot learning [137], self supervised learning [112], activity recognition [194, 117]. [137] is the closest to the proposed TLRR and uses relational reasoning for zero-shot learning. However, our work differs in several ways: (i) we do not work with a fixed set of labels, (ii) our relational module learns to identify relations between visual information rather than learning to identify relations between visual and semantic information, and (iii) our proposed problem setup requires the model to identify intra-video subtle differences between moments, whereas [137] learns to differentiate classes.

3.3 Methodology

3.3.1 Problem Statement

Let $\mathcal{S}^{tr} = \{(v, q, (\tau_s, \tau_e))|v \in \mathcal{V}^{tr}, q \in \mathcal{Q}^{tr}, \tau_s, \tau_e \in [0, T]\}$ be the training set of video-sentence pairs for seen queries where $\mathcal{V}^{tr}$ is the set of all training videos with maximum duration T, $\mathcal{Q}^{tr}$ is the set of seen queries, (τ_s, τ_e) are the ground truth temporal endpoints for a query. For a given test-set $\mathcal{S}^{te} = \{(v, q)|v \in \mathcal{V}^{te}, q \in \mathcal{Q}^{te}\}$ with video-sentence pairs, our task is to predict the set of temporal endpoints $\{(\tau_s, \tau_e)\}$. We consider that $\mathcal{Q}^{tr} \cap \mathcal{Q}^{te} = \emptyset$,

i.e., queries in test-set are not seen during training. As a result, $\mathcal{V}^{te}$ contains events that are not present in $\mathcal{V}^{tr}$. Additionally, we consider that S^{tr} is available during inference.

3.3.2 Localization Inference Schema

Existing temporal localization approaches [177, 102, 188] learn to encode fused moment-text representations. They either follow candidate moment sampling and encoding process to predict overlap scores (Figure 3.2 (a)) [177, 188] or summarize the whole video based on query encoding and segment level encoding of video to regress temporal endpoints (Figure 3.2 (b)) [102]. In both cases, moment representations are directly optimized for available seen events. As a result, the models get tuned to the available events in the training set and do not necessarily learn to generalize for unseen events. Since, our objective is to localize events which are not available during training, we deviate from the conventional approaches and propose a novel approach on how to address the text-based temporal localization task. For our proposed TLRR, we hypothesize that the correct moment corresponding to the unseen text query and the moments corresponding to the semantically relevant seen queries will contain shared concepts or similarities. Therefore, to identify the correct moment in a video based on an unseen text query, instead of directly predicting based on the moment-text representation, we utilize semantically relevant seen events. In that regard, we formulate the localization inference as a relational reasoning problem between two semantically relevant moments.

For a given video and an unseen text query, semantically relevant moments can be identified based on the semantics of the text query. Recent advances in Natural Language

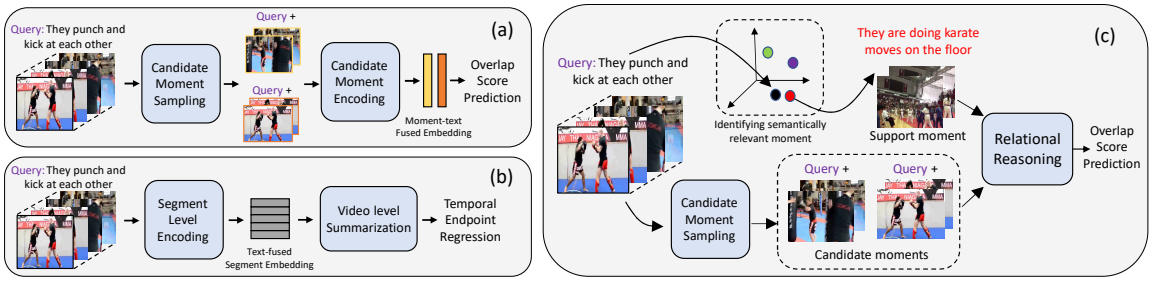


Figure 3.2: A brief illustration of our novel text-based temporal localization approach. While existing works learn to encode video segments to identify the correct moment ((a) and (b)), we consider relational reasoning between two semantically relevant moment for localization purpose (c).

Processing (NLP) unfold many sentence encoder models which are trained on large corpus of text data in self-supervised or unsupervised manner. These models are able to capture wide range of sentence semantics and can be transferred to other NLP tasks. Our idea is to use these sentence encoders to find semantically relevant moments. In our work, we utilize universal sentence encoder [12], which is also able to capture sentence semantics, to find semantically relevant moments. Figure 3.2 (c) clearly illustrates our localization inference scheme. Given the unseen query, instead of directly inferring overlap scores from moment-text fused representation, we first identify semantically relevant query and its corresponding moment using universal sentence encoder. We utilize this semantically relevant moment as the support moment and consider relational reasoning between the support moment and the candidate moments to identify the correct moment. Our motivation behind this approach is that this relational inference system can be learned using available training data and the learned relational model is transferable to unseen cases [137]. Our framework consists of candidate moment encoder, fusion network, support moment encoder and relational reasoning module.

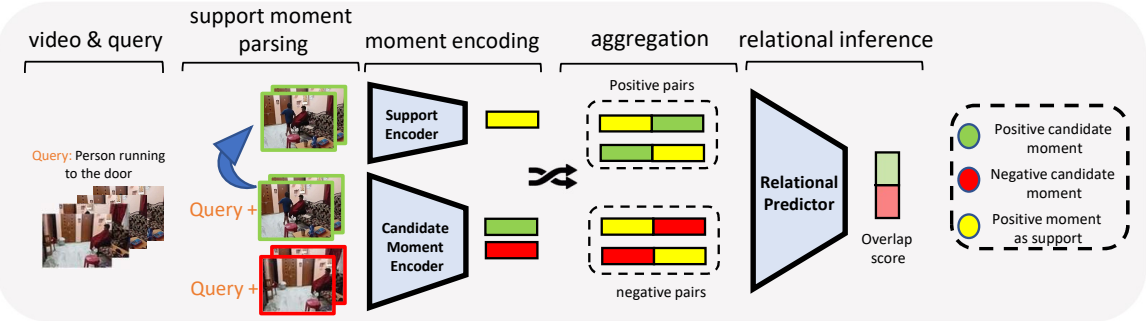


Figure 3.3: Overview of the framework and the training of the relational reasoning based temporal localization approach. Candidate moment and support moment representations are aggregated to form positive pairs (positive candidate, positive support) and negative pairs (negative candidate, positive support)/(positive candidate, negative support). The relational module is trained to estimate the relational scores based on the pairs.

3.3.3 Framework

As illustrated in Figure 3.3, our framework consists of a candidate moment encoder that generates a text-fused representation of candidate moments, a support moment encoder that encodes the support moment, and a relational prediction module to infer based on the relational reasoning between candidate moment and support moment. To learn the relational reasoning system utilizing available training samples, we mimic the relational inference task during training. At train-time, for seen queries in training set, we infer the overlap scores based on the relation between candidate moment and support moment, where the ground truth moment is used as the positive support moment. All the modules and the learning procedure are described in the following sections.

Visual Feature Extraction. We perform fixed interval sampling over the frames of the videos and sample l non-overlapping clips per video. For each clip, we extract 2D/3D convolutional feature, resulting in a set of l clip features $\{\boldsymbol{c}_i\}_{i=1}^{l}$. Here, $\boldsymbol{c}_i$ is the feature representation of the i^{th} clip.

Text Feature Extraction. We use GloVe word embedding [118] and Bi-directional LSTM network [57] for representing text queries. For each word s of the query sentence q, we use Glove word embeddings to obtain its initial embedding vectors, which are fed sequentially into a three-layer bidirectional LSTM network. The last hidden state $\hat{\boldsymbol{q}}$ is used as the feature representation of the input sentence.

Candidate Moment Encoding and Modality Fusion. Clip representations $\{\boldsymbol{c}_i\}_{i=1}^{l}$, sampled from each video is used to construct candidate moment representations. For each candidate moment, we max-pool the corresponding clip features across the specific time span. For example, moment corresponding to i^{th} to $(i+n)^{th}$ clips will be represented by $\boldsymbol{f}_{i:i+n} = MaxPool(\boldsymbol{c}_i, \ldots, \boldsymbol{c}_{i+n})$, where $\boldsymbol{f} \in \mathcal{R}^{d_f}$ (d_f is the feature dimension). Moment encodings and text encodings are projected in the same subspace and their dot product is taken as the fused moment-text representation by $\boldsymbol{e} = (\boldsymbol{W}^q\hat{\boldsymbol{q}}).(\boldsymbol{W}^f\boldsymbol{f})$. Here, $\boldsymbol{W}^q$ and $\boldsymbol{W}^f$ are the learnable parameters. We stack all moment-text representations of a video as a 2D feature map, similar to [188], and use L convolutional layers to further encode the representations. As a result, we obtain a set of candidate moment representations $\{\boldsymbol{m}_i\}_{i=0}^{N}$, where N is the total number of candidate moments from a video and $\boldsymbol{m}_i \in R^{d_m}$, where d_m is the feature dimension of the candidate moment representations.

Support Moment Encoder. We use a feed-forward network as the support moment encoder. For a support moment consisting of n consecutive clips $\{\boldsymbol{c}_i\}_{i=1}^{n}$, where $\boldsymbol{c}_i \in \mathcal{R}^{d_m}$, we first average pool the n clip representations to a single representation $\boldsymbol{s'} \in \mathcal{R}^{d_m}$. If we have multiple support moments, then we average pool all the support moment representations into a single representation. Then we use a feed-forward network to obtain the final support

representation $\boldsymbol{s}$ by

$$\boldsymbol{s} = ReLU(\boldsymbol{W}^s \boldsymbol{s'} + \boldsymbol{b}^s). \tag{3.1}$$

Here, $\boldsymbol{W}^s$ and $\boldsymbol{b}^s$ are the learnable parameters and $\boldsymbol{s} \in \mathcal{R}^{d_m}$. We keep the feature dimension of support moment same as the candidate moment feature dimension d_m. The input to the support moment encoder varies in the training stage and inference stage. In the training stage, the correct candidate moment is used as the support moment. In the inference/testing stage, based on the unseen test query, most semantically relevant moments from the training set are used as the support moments. These moments work as the helper to find the correct moment from the video.

3.3.4 Relational Prediction

The relational module is a function $\mathcal{Z}_\theta(\cdot)$ parameterized by learnable weights θ and modeled by a feed forward neural network. Input to the relational module is a pair of two representations $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, where one element represents the selected support moment $\boldsymbol{s}$ and the other element represents a candidate moment $\boldsymbol{m}_i$ from the set of candidate moment representations $\{\boldsymbol{m}_i\}_{i=1}^N$. We use concatenation as the aggregation function to get aggregated representation of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ as $a_{cat}(\boldsymbol{x}_i, \boldsymbol{x}_j)$. For a pair of support moment representation $\boldsymbol{s}$ and i^{th} candidate moment representation $\boldsymbol{m}_i$, the relational module outputs a overlap score ϕ_i by

$$\phi_i = \mathcal{Z}_\theta(a_{cat}(\boldsymbol{s}, \boldsymbol{m}_i)). \tag{3.2}$$

To confirm that the relational reasoning module $\mathcal{Z}_\theta$ predicts based on the relation between pair of representations and not based on a single representation, $\mathcal{Z}_\theta$ requires to

maintain the commutative property, i.e., $\mathcal{Z}_\theta(a_{cat}(\boldsymbol{s}, \boldsymbol{m}_i)) = \mathcal{Z}_\theta(a_{cat}(\boldsymbol{m}_i, \boldsymbol{s}))$. However, the concatenation operation $a_{cat}(\cdot, \cdot)$ is not commutative. Therefore, to enforce the commutative property of the relational module, we compute the overlap score for the pair of elements $\boldsymbol{s}$ and $\boldsymbol{m}_i$ by

$$\phi_i = \frac{1}{2}\big[\mathcal{Z}_\theta(a_{cat}(\boldsymbol{m}_i, \boldsymbol{s})) + \mathcal{Z}_\theta(a_{cat}(\boldsymbol{s}, \boldsymbol{m}_i))\big]. \tag{3.3}$$

3.3.5 Learning Relational Inference

In our learning setup, a training sample consists of a video v, a text query q, and temporal ground truth information for the query (τ_s, τ_e). Instead of learning to directly predict the overlap score for each candidate moment, we learn to infer the overlap scores based on the relation with most relevant support moments. To train this relational inference system, we sample two types of support moment: i) positive support moment and ii) negative support moment. For each query in a video, we extract the ground truth segment of the video and use it as the positive support moment $\boldsymbol{s}^+$. Again, for each query in a video, we select semantically unrelated query in the trainset and use its corresponding moment as the negative support moment $\boldsymbol{s}^-$. Our objective is to distinguish intra-video candidate moments based on the support moment. To do so, we compute overlap prediction loss $\mathcal{L}^{intra}$ for a set of pairs $\mathcal{X}^1$=$\{(\boldsymbol{m}_i, \boldsymbol{s}^+)\}$, which consists of pairs of all candidate moments and positive support moment in a video. To guide the learning of distinguishing intra-video candidate moments through relational inference system, we use scaled $tIoU$ (temporal Intersection-over-Union)

value with ground-truth segment as the supervision signal. We compute the scaled $tIoU$ by

$$y_i = \begin{cases} 0 & g_i \leq t_{min}, \\ \frac{g_i - t_{min}}{t_{max} - t_{min}} & t_{min} < g_i < t_{max}, \\ 1 & g_i > t_{max}. \end{cases} \tag{3.4}$$

Here, g_i is the ground truth $tIoU$ for the i^{th} candidate moment and t_{min}, t_{max} are two thresholds to compute y_i. For a video with N candidate moments, $\mathcal{L}^{intra}$ is realized by binary cross entropy loss as

$$\mathcal{L}^{intra} = -\frac{1}{N} \sum_{\mathcal{X}^1} \big[y_i \log(\phi_i) + (1 - y_i) \log(1 - \phi_i) \big]. \tag{3.5}$$

Here, ϕ_i is the overlap score computed using Eqn. 3.3. To ensure that the model predicts the overlap score based on the relationship between the candidate moment and the support moment, we use the sampled negative support moments s^- to train the model. In each video, candidate moments with $tIoU > t_{min}$ are considered as positive candidate moment m^+. For each video with P positive candidate moments, we formulate a set of pairs $\mathcal{X}^2$=$\{(\boldsymbol{m}_i^+, \boldsymbol{s}^-)\}$ and compute negative relational loss $\mathcal{L}^{neg}$ by

$$\mathcal{L}^{neg} = -\frac{1}{P} \sum_{\mathcal{X}^2} \log(1 - \phi_i). \tag{3.6}$$

The two losses are jointly considered for training our relational inference model,

with λ balancing contributions as in

$$\mathcal{L}^{total} = \mathcal{L}^{intra} + \lambda \mathcal{L}^{neg}. \tag{3.7}$$

We compute $\mathcal{L}_{total}$ for all seen video-text query pairs in the training set and optimize the relational inference model by minimizing the total loss.

3.3.6 Inference for Unseen Queries

During inference, given a video and an unseen text query, we are required to localize the correct moment. We use the universal sentence encoder [12] to find semantically relevant queries from the training set. Then the corresponding moment to the relevant query is used as a support moment. Based on the video, support moments, and the unseen query, the learned relational model predicts overlap score ϕ for different temporal granularities in one forward pass. All the predicted segments are ranked and refined with non-maximum suppression (NMS) according to the predicted ϕ. Afterwards, the final temporal grounding result is obtained.

3.4 Experiments

3.4.1 Reorganized Datasets

Existing benchmark temporal moment localization dataset splits are not designed for the task of temporal localization of novel events based on unseen text queries. Instead, training set (trainset for short) and testing set (testset for short) data are sampled from the

same distribution, and text queries in the testset overlap with text queries in the trainset. We reorganize two of the benchmark datasets namely Charades-STA [47] and ActivityNet Captions [67] to create splits according to our problem setting. For both datasets, we create splits based on the verbs and nouns present in the text queries. First, we combine all the annotations of the trainset and testset videos of the dataset. To create the splits, we consider a set of n_V verbs and n_N nouns present in the combined annotation. We consider it the set of seen verbs and seen nouns. Then, we identify videos that contain at least a single query that has a verb or noun not present in the mentioned set. In the selected videos, queries which do not have verbs or nouns from the mentioned set are collected as unseen testset split and, queries which have verbs or nouns from the mentioned set are collected as seen testset split. The training set is created from the rest of the videos, with queries that contain either verb or noun present in the mentioned set. We exclude queries which contains verb or noun from both seen set and unseen set. We use spaCy [60] to parse verbs and nouns from text queries. These reorganized datasets reflect a realistic setting as datasets are usually composed of recurring events of limited concepts. However, a localization system may encounter varied types of events in real-world applications. Excluding queries which contains verb or noun from both seen set and unseen set results in reduced number of moment-sentence pairs in the reorganized dataset. However, the size of the dataset doesn't have impact on the significance of our proposed problem setup.

Charades-STA Unseen. Charades-STA dataset contains a total of 6,670 videos where 5,336 and 1,334 are the number of training and testing videos. Textual annotations in Charades-STA has direct temporal correspondence with activity annotation of the Charades

Table 3.1: Tabulated summery of **number of moment-text pairs** in Charades-STA Unseen and ActivityNet Captions Unseen dataset.

Dataset	**Training**	**Unseen Testing**	**Seen Testing**
Charades-STA Unseen	5525	1665	867
ActivityNet Captions Unseen	5669	2553	710

dataset [131]. We combine training and testing set annotations and consider $n_V = 20$ and $n_N = 40$ (excluding 'person' noun) for creating Charades-STA Unseen dataset. In this way, we have Charades-STA Unseen dataset with 5525, 1665, and 867 training, unseen testing, and seen testing moment-sentence pairs respectively.

ActivityNet Captions Unseen. ActivityNet Captions [67] dataset is proposed for dense video captioning task. Each video contains at least two ground truth segments and each segment is paired with one ground truth caption [162]. This dataset contains around 20k videos which are split into training, validation, and testing set with 50%, 25%, and 25% ratio respectively. Textual description for only the training and validation set is given. We combine training and validation set and consider $n_V = 70$ and $n_N = 250$ for creating ActivityNet Captions Unseen dataset. In this way, we have ActivityNet Captions Unseen dataset with 5669, 2553, and 710 training, unseen testing, and seen testing moment-sentence pairs respectively.

Table 3.1 reports the number of moment-text pairs for training, unseen testing, and seen testing splits of both datasets. Table 3.2 reports the number of videos in each split of both datasets. Table 3.3 reports the number of verbs and nouns used to create the splits of both datasets. The list of verbs and nouns used for Charades-STA unseen and ActivityNet

Table 3.2: Tabulated summery of **number of videos** in the reorganized Charades-STA Unseen and ActivityNet Captions Unseen dataset.

Dataset	**Training**	**Unseen Testing**	**Seen Testing**
Charades-STA Unseen	3366	1271	486
ActivityNet Captions Unseen	3939	1993	513

Table 3.3: **Number of verbs and nouns** used to create train/test splits of Charades-STA Unseen and ActivityNet Captions Unseen dataset.

Dataset	**Number of Verbs**	**Number of Nouns**
Charades-STA Unseen	20	40
ActivityNet Captions Unseen	70	250

Figure 3.4: List of selected verbs and nouns for Charades-STA Unseen.

Charades-STA Unseen	
Selected Verbs	Selected Nouns
put, begin, play, start, pour, watch, take, sneeze, awaken, hold, sit, open, tidy, smile, cook, run, closet, see, drink, eat	book, shelf, phone, glass, water, television, cup, fridge, mirror, camera, front, computer, notebook, bag, door, shoe, wardrobe, entryway, stove, coffee, table, room, man, sofa, couch, hallway, closet, bed, laptop, dish, medicine, guy, chair, refrigerator, clothe, sandwich, food, blanket, light, knob

Captions unseen are given in Figure 3.4 and Figure 3.5 respectively.

3.4.2 Evaluation Metric

We use "$R@k, IoU@m$", which reports the percentage of at least one of the top-k results having Intersection-over-Union (IoU) larger than m [47]. For a text query, "$R@k, IoU@m$" reflects if one of the top-k retrieved moments has IoU with the ground

Figure 3.5: List of selected verbs and nouns for ActivityNet Captions Unseen.

ActivityNet Captions Unseen	
Selected Verbs	Selected Nouns
see, stand, lead, dance, capture, continue, end, lay, start, demonstrate, point, do, begin, move, perform, sit, wax, walk, film, turn, go, watch, play, hold, pose, pierce, follow, rub, show, ride, lean, cover, mop, set, kneel, speak, mix, measure, cut, look, twist, bend, grab, place, pick, hit, throw, attempt, picture, lie, be, flash, wear, talk, wrap, tape, block, climb, wave, jump, zoom, slide, land, hang, smile, cross, get, pop, make, put	woman, room, dancing, girl, camera, movement, floor, video, title, logo, sequence, man, living, exercise, ground, area, body, sit, up, people, kitchen, task, ski, hallway, dog, sock, lady, sidewalk, playing, music, people, boy, ball, picture, front, chair, person, ear, lotion, piercing, camel, pyramid, hand, lens, harness, child, house, mop, family, member, bedroom, ingredient, plaster, tile, piece, line, side, road, field, object, baseball, game, penalty, player, goal, head, screen, word, end, overall, wrapping, paper, toy, suit, desk, grass, uniform, set, monkey, bar, way, pan, snowboard, mountain, hill, playroom, slide, time, back, couch, sport, jersey, wall, middle, clipboard, smooth, top, leg, clip, part, city, soccer, sand, play, president, crowd, speech, other, beer, kid, beach, right, castle, circle, water, work, midway, float, pile, leave, shot, blower, machine, distance, basketball, basket, transition, stool, color, frame, speed, bagpipe, canoe, angle, group, blackjack, table, place, card, costume, tug, rope, slope, course, filmer, waif, platform, triangular, obstacle, crash, railing, bowl, noodle, broth, pair, shoe, office, close, bike, wheel, tire, tool, liquid, tip, glass, sugar, plate, mixer, mixture, drink, corner, building, bow, move, bowing, cartwheel, flip, flute, fingering, octave, note, salad, dish, information, trip, canopy, food, market, customer, purchase, money, seller, thumb, chef, counter, hulte, bite, size, cilantro, product, credit, dancer, dance, river, row, tree, bunch, intertube, tuber, fall, stunt, terrain, range, sweat, dirt, sort, acrobatic, action, variety, stun, landscape, track, run, mat, lime, board, blender, juice, jar, straw, wedge, rim, sink, brush, faucet, nozzle, dealer, chip, equipment, number, pace, seam, point, cheer, background, harmonica, detail, regard, feature, coat	

truth moment larger than the specified threshold m. So, "$R@k, IoU@m$" is either 1 or 0 for each text query. We compute it for all the text queries in the testing sets and report the average results for $k \in \{1, 5\}$ and $m \in \{0.50, 0.70\}$. We also compute mIoU where mIoU is the average IoU over all testing samples.

3.4.3 Implementation Details

We use VGG feature [132] for Charades-STA Unseen dataset. For ActivityNet Captions Unseen dataset, we use extracted C3D features [144]. The number of frames in a clip is set to 4 for Charades-STA Unseen, and 16 for ActivityNet Captions Unseen and

Table 3.4: This table reports ***unseen*** text query based temporal moment localization performance of TLRR, compared against several approaches, on Charades-STA Unseen dataset.

Method	**R@1, IoU@0.5**	**R@1, IoU@0.7**	**R@5, IoU@0.5**	**R@5, IoU@0.7**	**mIoU**
DeViSE [43]	29.98	11.29	71.42	39.81	-
ESZSL [126]	23.90	10.13	60.50	34.53	-
SCDM [177]	28.22	11.89	54.25	32.95	28.63
LGI [102]	29.01	12.85	-	-	29.62
2D-TAN [188]	31.05	13.33	70.75	36.94	29.88
TLRR	**33.15**	**16.22**	**77.66**	**42.40**	**31.29**

we use non-overlapping clips for both datasets. The number of sampled clips N is set to 16 for Charades-STA Unseen, 64 for ActivityNet Captions Unseen. For the candidate moment encoder, we adopt a 4-layer convolution network with a kernel size of 5 for Charades-STA Unseen and a 4-layer convolution network with a kernel size of 9 for ActivityNet Captions Unseen. For both datasets, the support moment encoder is a single-layer feed-forward network and the relational prediction network is a two-layer feed-forward network. The proposed network is implemented in TensorFlow and trained using a single RTX 2080 GPU. We use mini-batches containing 32 video-sentence pairs and use Adam [65] optimizer with a learning rate of 0.0001. The dimension of both candidate moment representation d_m and support moment representation d_s is set to 512 for both datasets. We set λ=3 empirically in Eqn 3.7 for both datasets. The scaling thresholds t_{min} and t_{max} of Eqn. 3.4 are set to 0.5 and 1.0 respectively for both datasets. Non-maximum suppression (NMS) with a threshold of 0.5 is applied during the inference. We train TLRR for 50 epochs. We select the checkpoint which has the best average performance across metrics for seen queries.

Table 3.5: This table reports ***unseen*** text query based temporal moment localization performance of TLRR, compared against several approaches, on ActivityNet Captions Unseen dataset.

Method	**R@1, IoU@0.5**	**R@1, IoU@0.7**	**R@5, IoU@0.5**	**R@5 IoU@0.7**	**mIoU**
DeViSE [43]	5.07	2.00	10.46	4.05	-
ESZSL [126]	4.72	1.85	11.83	4.48	-
SCDM [177]	19.22	8.22	46.38	23.58	23.97
2D-TAN [188]	19.15	10.26	38.78	24.01	21.70
VSLNet [184]	19.23	9.99	-	-	25.32
TLRR	**23.19**	**13.24**	**53.31**	**36.66**	**26.35**

3.4.4 Result Analysis

Temporal Localization Performance of Novel/Unseen Events. Since ours is the first work on temporal localization of novel events, there are no existing approaches to directly compare with. As our problem setup is closely related to zero-shot settings, we adapt two zero-shot learning approaches namely **DeViSE** [43] and **ESZSL** [126] for this problem setup. We also compare with some of the state-of-the-art temporal localization approaches with publicly available codes, e.g., **2D-TAN** [188], **SCDM** [177], **LGI** [102], and **VSLNet** [184], by training those models using our reorganized training splits.

Table 3.4 and Table 3.5 illustrate the TLRRs' performance for temporal localization of novel event based on unseen text query and compare it with other approaches for Charades-STA Unseen and ActivityNet Captions Unseen dataset respectively. For the Charades-STA Unseen dataset, the performance of different baseline approaches are comparable among them. However, TLRR provides $2\% - 7\%$ absolute improvement over the best scores of compared approaches over all the reported metrics. In Table 3.5, baseline zero-shot approaches (DeViSE, ESZSL) are performing poorly for ActivityNet Captions Unseen dataset. This is because

the text queries are complex compared to Charades-STA Unseen and it requires fine-grained analysis of longer videos in ActivityNet Caption Unseen. We observe $3\% - 15\%$ absolute improvement over best scores of compared approaches in the ActivityNet Captions Unseen dataset.

Relational Reasoning Performance Analysis. Since TLRR's performance is dependent on its ability to reason on the relationship of two different moments, in Table 3.6, we analyze the competence of our relational reasoning module $\mathcal{Z}_\theta$ for Charades-STA Unseen dataset. We consider three scenarios: i) **Irrelevant**: based on the unseen text query, retrieve the seen query from the semantic embedding space that are furthest away or most irrelevant and use the corresponding moment as the support information, ii) **Random**: retrieve random seen query from the training set and use the corresponding moment as the support information, and iii) **Relevant**: retrieve the nearest/most relevant seen query from the semantic embedding space and use the corresponding moment as the support information (i.e., our proposed TLRR). We observe that when irrelevant queries are retrieved and their corresponding moment is used as the support, the performance goes down. Since the moment corresponding to a irrelevant query does not contain shared concept/ similarities with the correct moment, the relational module expectedly fails to identify the correct moment. When random seen queries are selected, the performance is better compared to the irrelevant case. We obtain the best performance when the closest seen query is selected from the semantic embedding space.

Temporal Localization Performance of Seen Events. We further report the performance of different approaches when evaluated on the testing split of seen queries in both the datasets

Table 3.6: This table reports *unseen* text query based novel event localization performance using different types of support moments to analyze TLRR for Charades-STA Unseen dataset.

Support Moment	**R@1, IoU@0.5**	**R@1, IoU@0.7**	**R@5, IoU@0.5**	**R@5, IoU@0.7**	**mIoU**
Irrelevant	20.30	11.05	62.58	33.93	22.48
Random	28.71	14.47	73.57	40.24	28.40
Relevant	**33.15**	**16.22**	**77.66**	**42.40**	**31.29**

on Table 3.7 and Table 3.8. Although the main focus of this work is temporal localization of unseen events, this experiment is presented to evaluate how the performance of different methods changes for seen events compared to localization of unseen events (Table 3.4 and Table 3.5). We expect any method to work slightly better on localizing the seen events compared to the unseen ones; however, a drastic/large change would indicate poor generalization ability of the model.

For the compared methods and baselines, we observe that there is a significant difference in performance when the same model is evaluated in the testing split of seen queries and testing split of unseen queries for both datasets comparing Table 3.4 and Table 3.5 with Table 3.7 and Table 3.8 respectively. Not surprisingly, both the conventional temporal localization approaches (i.e., SCDM and 2D-TAN) show a drastic change in performance across metrics in both datasets. The average difference in performance is reported by Δ_{avg} in Table 3.7 and Table 3.8. SCDM shows 19.80% average difference in Charades-STA and 13.24% average difference across metrics in ActivityNet in localization performance of seen queries compared to localization performance of unseen queries. Similarly, 2D-TAN shows average difference (across metrics) of 5.89% in Charades-STA and 16.18% in ActivityNet in localizing seen queries compared to unseen. Though the zero-shot based approaches (DeViSE and ESZSL) show small gap in performance between seen and unseen events, which

is expected due to the approaches generalization ability, they are unable to maintain a proper level of localization performance compared to other methods. However, the proposed TLRR approach shows a significantly lower change in performance, e.g., 3.37% average in Charades-STA and 7.13% average in ActivityNet Captions.

This indicates the significance of the problem setup and generalization ability of TLRR. Unlike the conventional temporal localization approaches, TLRR is not designed to specifically focus on the seen events. In Table 3.7 and Table 3.8, we observe that model optimized to do localization inference directly based on the candidate moment representation overall performs better compared to TLRR for types of events that are already seen in training. However, direct localization limits these models' capacity to a small set of events which is evident by the significant gap between performances for seen and unseen events. Instead, our proposed TLRR approach is able to retain a competitive performance for the seen queries and boost the performance for unseen queries resulting in reducing the performance gap between seen and unseen events. Also, our proposed TLRR is able to show comparable performance on the original temporal localization dataset, even though TLRR is not optimized for seen events and have a relatively simple base architecture.

Effect of $\mathcal{L}^{neg}$ in learning TLRR. TLRR uses $\mathcal{L}^{intra}$ and $\mathcal{L}^{neg}$ to learn relational localization system. Effectiveness of these two loss components for distinguishing intra-video moments by relational prediction is evident from Table 3.4, Table 3.5, and Table 3.6. We consider two setups, i) TLRR trained with $\mathcal{L}^{intra}$ and ii) TLRR trained with $\mathcal{L}^{intra} + \lambda\mathcal{L}^{neg}$. We observe that when only $\mathcal{L}^{intra}$ is used to train TLRR, there is almost no difference in performance (difference within 1%) for using relevant or irrelevant moments as input to the

Table 3.7: This table reports ***seen*** text query based temporal moment localization performance of TLRR on Charades-STA Unseen dataset. Here, Δ_{avg} refers to average performance difference for seen events and unseen events (Table 3.4) for a specific method. From the lower value of Δ_{avg}, it is evident that TLRR generalizes significantly better than other temporal localization approaches.

Method	**R@1, IoU@0.5**	**R@1, IoU@0.7**	**R@5, IoU@0.5**	**R@5, IoU@0.7**	$\boldsymbol{\Delta_{avg} \downarrow}$
DeViSE [43]	36.34	15.86	77.66	44.10	5.36
ESZSL [126]	37.50	18.40	72.34	42.13	10.34
SCDM [177]	50.46	28.00	73.49	54.86	19.80
2D-TAN [188]	37.95	18.45	76.70	42.56	5.89
TLRR	34.83	20.76	78.78	48.56	**3.37**

Table 3.8: This table reports ***seen*** text query based temporal moment localization performance of TLRR on ActivityNet Captions Unseen dataset. Δ_{avg} refers to average performance difference for seen events and unseen events (Table 3.5) for a specific method. From the lower value of Δ_{avg}, it is evident that TLRR generalizes significantly better than other temporal localization approaches.

Method	**R@1, IoU@0.5**	**R@1, IoU@0.7**	**R@5, IoU@0.5**	**R@5, IoU@0.7**	$\boldsymbol{\Delta_{avg} \downarrow}$
DeViSE [43]	12.07	5.40	18.18	8.52	**5.64**
ESZSL [126]	12.64	5.40	19.74	8.66	5.89
SCDM [177]	34.66	20.74	59.51	35.37	13.24
2D-TAN [188]	34.65	22.39	57.18	42.68	16.18
TLRR	27.46	17.61	60.42	49.44	7.13

support encoder. However, there is $5\% - 15\%$ difference in Charades-STA Unseen dataset for using relevant or irrelevant moments as input to the support encoder when $\mathcal{L}^{intra} + \lambda\mathcal{L}^{neg}$ is used to train TLRR. So, $\mathcal{L}^{neg}$ enforces the model to predict based on the relation.

Qualitative Results. In Figure 3.6, 3.7, and 3.8, we illustrate some example cases of our system's success. Given the query 'The person laughs' and the corresponding video, Figure 3.6 shows: (a) the ground truth segment of the video which corresponds to the text query, (b) predicted moment by 2D-TAN, (c) predicted moment when the irrelevant moment is used as support, and (d) predicted moment using retrieved relevant support moment. Person

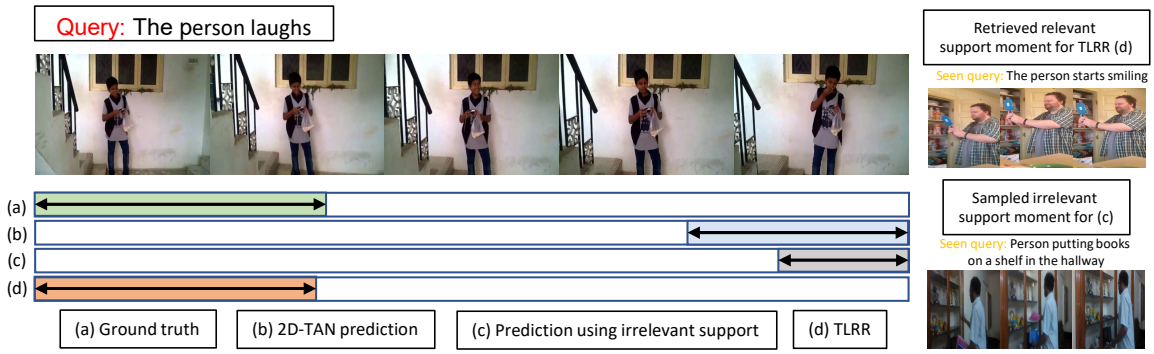


Figure 3.6: Given the query 'The person laughs' and the corresponding video, this figure shows: (a) ground truth segment of the video which corresponds to the text query, (b) predicted moment by 2D-TAN, (c) predicted moment when irrelevant moment is used as support, and (d) predicted moment using retrieved relevant support moment (TLRR). While (b) and (c) result in failure, TLRR is able to detect the correct moment using relational reasoning.

laughing is a difficult event to detect as it encompasses a small region of the frame and results in a small temporal variation in the feature. Without any notion/previous knowledge of how the activity/event is, it becomes even harder, which is reflected by the failure case of (b) and (c). However, TLRR is able to detect the correct moment using relational reasoning. Figure 3.7 shows an example from the ActivityNet Captions Unseen dataset. Given the unseen text query 'An older blonde newswomen is reading a story' and the corresponding video, our approach retrieves the semantically most relevant query and its corresponding moment as the support moment. Then based on reasoning with the support moment, our approach identifies the correct moment in the given video. Figure 3.8 shows an example from the Charades-STA Unseen dataset. It also illustrates that our approach is able to identify correct moments based on relational reasoning for unseen text queries.

Figure 3.7: Example illustration from ActivityNet Captions Unseen, where splits are created based on activity annotation. Given the text query, 'An older blonde newswomen is reading a story' and the corresponding video, our proposed approach retrieves the moment corresponding to the semantically relevant query 'A blonde woman is talking in a room' from the train set, reason on that and identifies the correct moment in the video. GT indicates the ground truth timestamps and Prediction indicates the predicted temporal endpoints of our approach.

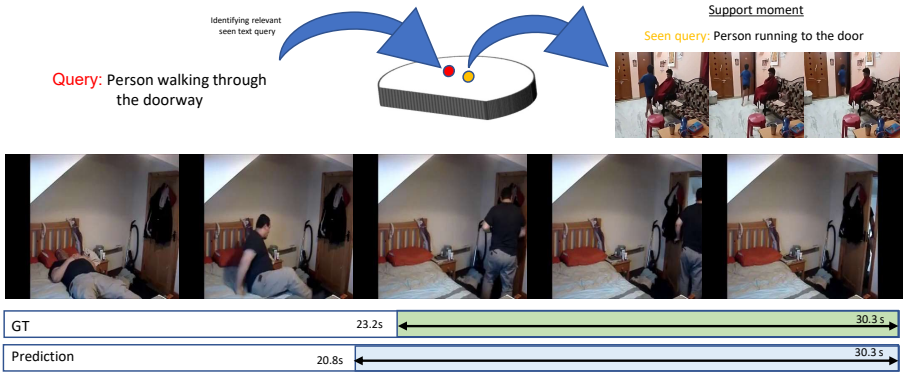


Figure 3.8: Given the text query 'Person walking through the doorway' and the corresponding video, our proposed approach retrieves the moment corresponding to the semantically relevant query 'Person running to the door' from the train set, reason on that, and identifies the correct moment in the video. GT indicates the ground truth timestamps and Prediction indicates the predicted temporal endpoints of our approach.

Significance of the Problem Setting. Figure 3.9 illustrates the significance of our problem setting. We evaluate the performance of a trained text-based temporal localization model for both seen events/queries and unseen events/queries. For Charades-STA Unseen, we consider SCDM [177], which predicts the overlap score and temporal offset directly based

on candidate moment representation. For ActivityNet Captions Unseen dataset, we consider 2D-TAN [188], which also predicts overlap scores based on candidate moment representation directly. We observe that there is a significant difference of performance between seen events/queries and unseen events/queries for both datasets. It demonstrates the requirement of a system that can retain the performance for seen queries and improve the performance for unseen queries.

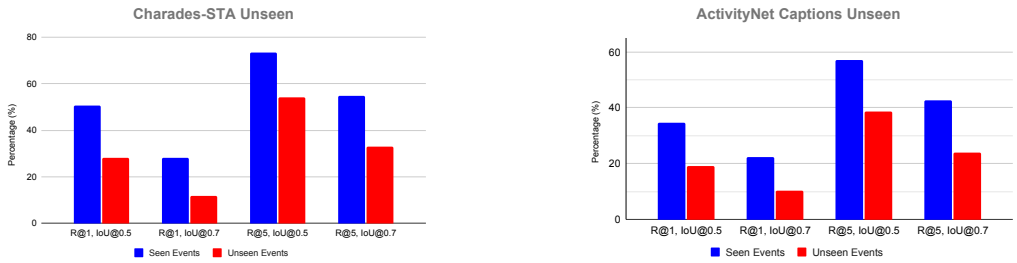


Figure 3.9: This figure illustrates the performance of SCDM [177] for Charades-STA Unseen and 2D-TAN [188] for ActivityNet Captions Unseen dataset for seen events and unseen events. For both datasets, performance of the trained model drops significantly for unseen events.

Efficiency of TLRR. We compare the run-time of our proposed TLRR with conventional temporal localization approaches SCDM and 2D-TAN. It is expected that TLRR would require more inference time due to the extra steps of computation of relevant moments and relational reasoning. We observe from Table 3.9 that compared to 2D-TAN and SCDM, proposed TLRR takes slightly more time in inference (i.e., 1.76s for proposed vs., 1.30s for 2D-TAN and 1.23s for SCDM).

Performance of TLRR in Original Charades-STA. We conduct an experiment on the original Charades-STA dataset which is reported in Table 3.10. Our proposed TLRR is able to show comparable performance on the original temporal localization dataset, even

Table 3.9: Per batch inference time of TLRR compared to SCDM and 2D-TAN in ActivityNet Captions Unseen dataset.

Method	**Inference Time**
SCDM [177]	1.23 s
2D-TAN [188]	1.30 s
TLRR	1.76 s

though TLRR is not optimized for seen events (since similar events are available in the trainset, all events in the testset can be considered as seen events) and have a relatively simple base architecture.

Table 3.10: This table reports text query based temporal moment localization performance of TLRR on the original Charades-STA dataset.

Method	**R@1, IoU@0.5**	**R@1, IoU@0.7**	**R@5, IoU@0.5**	**R@5, IoU@0.7**
CTRL [47]	23.63	8.89	58.92	29.52
2D-TAN [188]	**39.70**	**23.31**	80.32	**51.26**
TLRR	37.63	21.48	**82.61**	49.27

3.5 Conclusion

In this work, we address the novel problem of temporal localization of unseen/novel events based on unseen text queries. The problem of identifying novel events in video is important and practical because not every kind of event can be expected to be within the training set. This allows for generalization of temporal localization methods to novel scenarios. We propose a relational reasoning based framework hypothesizing a conceptual relation between moments corresponding to semantically relevant queries. Extensive experiments on reorganized Charades-STA and ActivityNet Captions datasets demonstrate the effectiveness

of the proposed framework compared to several baselines in localizing video moments from text queries. Our code and dataset splits will be publicly available. Though support moment based relational prediction can reduce the performance gap between seen and unseen events, it is burdened with the extra computation of relevant moments, which is computationally expensive. Future work can focus on this issue.

Chapter 4

Audio-Visual-Language Navigation

4.1 Introduction

Building embodied robotic agents that can harmoniously co-habit and assist humans has been one of the early dreams of AI. A recent incarnation of this dream has been in designing agents that are capable of autonomously navigating realistic virtual worlds for solving pre-defined tasks. For example, in vision-and-language navigation (VLN) tasks [4], the goal is for the AI agent to either navigate to a goal location following the instructions provided in natural language, or to explore the visual world seeking answers to a given natural language question [29, 154, 174]. Typical VLN agents are assumed deaf; i.e., they cannot hear any audio events in the scene – an unnatural restriction, especially when the agent is expected to operate in the real world. To address this shortcoming, SoundSpaces [16] reformulated the navigation task with the goal of localizing an audio source in the virtual scene; however without any language instructions for the agent to follow.

Figure 4.1: An illustration of our proposed AVLEN framework. The embodied agent starts navigating from location denoted ① guided by the audio-visual event at ③. At location ②, the learned policy for the agent decides to seek help from an oracle (e.g., because the audio stopped). The oracle provides a short natural language instruction for the agent to follow. The agent translates this instruction to produce a series of navigable steps to move towards the goal ③.

Real-world navigation is not only audio-visual, but also is often complex and stochastic, so the agent must inevitably seek a synergy between the audio, visual, and language modalities for successful navigation. Consider, for example, a robotic agent that needs to find where the *"thud of a falling person"* or the *"intermittent dripping sound of water"* is heard from. On the one hand, such a sound may not last long and may not be continuously audible, and thus the agent must use semantic knowledge of the audio-visual modality [15] to reach the goal. On the other hand, such events need to be catered to timely and the agent should minimize the number of navigation mistakes it makes – a situation that can be efficiently dealt with if the agent can seek human help when it is uncertain of its navigation actions. Motivated by this insight, we present AVLEN – a first of its kind embodied navigation agent for localizing an audio source in a realistic visual world. Our agent not only learns to use the audio-visual cues to navigate to the audio source, but also

learns to implicitly model its uncertainty in deciding the navigation steps and seeks help from an oracle for navigation instructions, where the instructions are provided in short natural language sentences. Figure 4.1 illustrates our task.

To implement AVLEN, we build on the realistic virtual navigation engine provided by the Matterport 3D simulator [4] and enriched with audio events via the SoundSpaces framework [16]. A key challenge in our setup is for the agent to decide when to query the oracle, and when to follow the audio-visual cues to reach the audio goal. Note that asking for help too many times may hurt agent autonomy (and is perhaps less preferred if the oracle is a human), while asking too few questions may make the agent explore the scene endlessly without reaching the goal. Further, note that we assume the navigation instruction provided to the agent is in natural language, and thus is often abstract and short (see Figure 4.1 above), making it difficult to be correctly translated to agent actions (as seen in VLN tasks [4]). Thus, the agent needs to learn the stochasticity involved in the guidance provided to it, as well as the uncertainty involved in the audio-visual cues, before selecting which modality to choose. To address these challenges, we propose a novel multimodal hierarchical options based deep reinforcement learning framework, consisting of learning a high-level policy to select which modality to use for navigation, among (i) the audio-visual cues, or (ii) natural language cues, and two lower-level policies that learn (i) to select navigation actions using the audio-visual features, or (ii) learns to transform natural language instructions to navigable actions conditioned on the audio-visual context. All the policies are end-to-end trainable and is trained offline. During inference, the agent uses its current state, the audio-visual cues, and the learned policies to decide if it needs oracle

help or can continue navigation using the learned audio goal navigation policies. Closely related to our motivations, a few recent works [27, 105, 106, 198] propose tasks involving interactions with an oracle for navigation. Specifically, in [27, 198] model uncertainty is used to decide when to query the oracle, where the uncertainty is either quantified in terms of the gap between the action prediction probabilities [27] to be less than a heuristically chosen threshold, or use manually-derived conditions to decide when an agent is lost in its navigation path [106]. In [105], the future actions of the policy of interest are required to be fully observed to identify when the agent is making mistakes and uses this information to decide when to query. Instead of resorting to heuristics, we propose a data-driven scheme to learn policies to decide when to query the oracle, these policies thus automatically learning the navigation uncertainty in the various modalities.

To empirically demonstrate the efficacy of our approach, we present extensive experiments on the language-augmented semantic audio-visual navigation (SAVi) task within the SoundSpaces framework for three very challenging scenarios, i.e., when: (i) the sound source is sporadic, however familiar to the agent, (ii) sporadic but unheard of during training, and (iii) unheard and ambiguous due to the presence of simultaneous distractor sounds. Our results show clear benefits when the agent knows when to query and how to use the received instruction for navigation, as substantiated by improvements in success rate by nearly 3% when using the language instructions directly and by more than 10% when using the ground truth navigation actions after the query, even when the agent is constrained to trigger help only to a maximum of three times in a long navigation episode.

Before proceeding to detail our framework, we summarize below the main contributions of this work.

- We are the first to unify and generalize audio-visual navigation with natural language instructions towards building a complete audio-visual-language embodied AI navigation interactive agent.
- We introduce a novel multimodal hierarchical reinforcement learning framework that jointly learns policies for the agent to decide: (i) when to query the oracle, (ii) how to navigate using audio-goal, and (iii) how to use the provided natural language instructions.
- Our approach shows state-of-the-art performances on the semantic audio-visual navigation dataset [15] with 85 large-scale real-world environments with a variety of semantic objects and their sounds, and under a variety of challenging acoustic settings.

4.2 Related Works

Audio-Visual Navigation. The SoundSpaces simulator, introduced in [16] to render realistic audio in 3D visual environments, pioneered research into the realm of audio-visual embodied navigation. The AudioGoal task in [16] consists of two sub-tasks, namely to localize an object: i) that sounds continuously throughout the navigation episode [16, 17] and ii) that sounds sporadically or can go mute at any time [15]. In this work, we go beyond this audio-visual setting, proposing a new task that equips the agent with the ability to use natural language – a setup that is realistic and practically useful.

Instruction-Following Navigation. There are recent works that attempt to solve the problem of navigation following instructions [4, 59, 64, 85, 97]. The instruction

can be of many forms; e.g., structured commands [106], natural language sentences [4], goal images [93], or a combination of different modalities [105]). The task in vision-and-language navigation (VLN) for example is to execute free-form natural language instructions to reach a target location. To embody this task, several simulators have been used [4, 16, 128] to render real or photo-realistic images and perform agent navigation through a discrete graph [4, 19] or continuous environment [66]. One important aspect of vision and language navigation is to learn the correspondence between visual and textual information. To achieve this, [151] uses cross-modal attention to focus on the relevant parts of both the modalities. In [94] and [95], an additional module is used to estimate the progress, which is then used as a regularizer. In [42] and [139], augmented instruction-trajectory pairs is used to improve the VLN performance. In [196], long instructions are learned to be decomposed into shorter ones, executing them sequentially (via e.g., navigation). Recently, there are works using Transformer-based architectures for VLN [59, 97]. In [59], the BERT [31] architecture is used in a recurrent manner maintaining cross-modal state information. These works only consider the language-based navigation task. Different from these works, our proposed AVLEN framework solves vision-and-language navigation as a sub-task within the original semantic audio-visual navigation task [16].

Interactive Navigation. Recently, there have been works where an agent is allowed to interact with an oracle or a different agent, receiving feedback, and utilizing this information for navigation [27, 105, 106, 143]. The oracle instructions in these works are limited to either ground truth actions [27] or a direct mapping of a specific number of actions to consecutive phrases [106] is needed. Though [105] uses a fixed set of natural language

instructions as the oracle feedback, it is coupled with the target image that the agent will face after completion of the sub-goal task. In Nguyen et al. [105], the agent needs to reach a specific location to query, which may be infeasible practically or sub-optimal if these locations are not chosen properly. Our approach differs fundamentally from these previous works in that we consider free-form natural language instructions and the agent can query the oracle from any navigable point in the environment, making our setup very natural and flexible.

4.3 Proposed Method

In this section, we will first formally define our task and our objective. This will be followed by details of our multimodal hierarchical reinforcement learning framework and our training setup.

4.3.1 Problem Setup

Consider an agent in a previously unseen 3D world navigable along a densely-sampled finite grid. At each vertex of this grid, the agent could potentially take one of a subset of actions from an action set $A = \{\text{stop}, \text{move\_forward}, \text{turn\_right}, \text{turn\_left}\}$. Further, the agent is assumed to be equipped with sensors for audio-visual perception via a binaural microphone and an ego-centric RGBD camera. The task of the agent in AVLEN is to navigate the grid from its starting location to find the location of an object that produces a sound (*AudioGoal*), where the sound is assumed to be produced by a static object and is semantically unique, however can be unfamiliar, sporadic, or ambiguous (due to distractors). We assume the agent calls the stop action only at the *AudioGoal* that

terminates the episode. In contrast to the task in SoundSpaces, an AVLEN agent is also equipped with a language interface to invoke a *query* to an *oracle* under a budget (e.g., a limit on the maximum number of such queries). The oracle responds to the query of the agent via providing a natural language short navigation instruction; this instruction describing (in natural language) an initial *segment* along the shortest path trajectory from the current location of the agent to the goal. For example, for a navigation trajectory given by the actions $\langle \text{move\_forward}, \text{turn\_right}, \text{turn\_right}, \text{move\_forward}, \text{turn\_left} \rangle$, the corresponding language instruction provided by the oracle to the agent could be *"go around the sofa and turn to the door"*. As is clear, to use this instruction to produce navigable actions, the agent must learn to associate the language constructs with objects in the scene and their spatial relations, as well as their connection with the nodes in the navigation grid. Further, given the limited budget to make queries, the agent must learn to balance between when to invoke the query and when to navigate using its audio-visual cues. In the following, we present a multimodal hierarchical options approach to solve these challenges in a deep reinforcement learning framework.

4.3.2 Problem Formulation

We formulate the AVLEN task as a partially-observable Markov decision process (POMDP) characterized by the tuple $(\mathcal{S}, \mathcal{A}, T, R, \mathcal{O}, P, \mathcal{V}, \gamma)$, where $\mathcal{S}$ represents the set of agent states, $\mathcal{A} = A \cup \{\text{query}\}$ with the navigation actions A defined above combined with an action to query the oracle, $T(s'|s, a)$ is the transition probability for mapping a state-action pair (s, a) to a state s', $R(s, a)$ is the immediate reward for the state-action pair, $\mathcal{O}$ represents a set of environment observations o, $P(o|s', a)$ captures the probability of observing $o \in \mathcal{O}$

in a new state s' after taking action a, and $\gamma \in [0, 1]$ is the reward discount factor for long-horizon trajectories. Our POMDP also incorporates a language vocabulary $\mathcal{V}$ consisting of a dictionary of words that the oracle uses to produce the natural language instruction. As our environment is only partially-observable, the agent may not have information regarding its exact state, instead maintains a belief distribution b over $\mathcal{S}$ as an estimate of its current state. Using this belief distribution, the expected reward for taking an action a at belief state b can be written as $R'(b, a) = \sum_{s \in \mathcal{S}} b(s)R(s, a)$. With this notation, the objective of the agent in this work is to learn a policy $\pi : \mathbb{R}^{|\mathcal{S}|} \times \mathcal{A} \to [0, 1]$ that maximizes the expected return defined by the value function V^π, while minimizing the number of queries made to the oracle; i.e.,

$$\arg\max_{\pi} V^\pi(b_0) \text{ where } V^\pi(b) = \mathbb{E}\left[\sum_{i=0}^{\infty} \gamma^i \left(R'(b_{t+i}, a_{t+i}) - \zeta(t+i)\mathbb{I}(a_{t+i} = \text{query})\right) | b_t = b, \pi\right], \tag{4.1}$$

where $\mathbb{I}$ denotes the indicator function, and the updated belief $b_{t+1} = \text{update}(o_{t+1}, b_t, a_t)$ defined for state s' as $b_{t+1}(s') = \eta P(o_{t+1}|s', a_t) \sum_{s \in \mathcal{S}} b_t(s)T(s'|s, a_t)$ for a normalization factor $\eta > 0$. The function $\zeta \colon \mathbb{R}_+ \to \mathbb{R}$ produces a score balancing between the frequency of queries and the expected return from navigation.

At any time step t, the agent (in belief state b_t) receives an observation $o_{t+1} \in \mathcal{O}$ from the environment and selects an action a_t according to a learned policy π; this action transitioning the agent to the new belief state b_{t+1} as per the transition function $T'(b_{t+1}|a_t, b_t) = \sum_{o \in \mathcal{O}} \mathbb{I}(b_{t+1} = \text{update}(o, b_t, a_t))P(o|s_t, a_t)$ while receiving an immediate reward $R'(b_t, a_t)$. As the navigation state space $\mathcal{S}$ of our agent is enormous, keeping a belief distribution on all states might be computationally infeasible. Instead, similar to [15], we

keep a history of past K observations in a memory module M, where an observation o_t at time step t is encoded via the tuple $e_t^o = (F_t^V, F_t^B, F_{t-1}^A, p_t)$ comprising neural embeddings of egocentric visual observation (RGB and depth) represented by F_t^V, the binaural audio waveform of the *AudioGoal* heard by the agent represented as a two channel spectrogram F_t^B, and the previous action taken F_{t-1}^A, alongside the pose of the agent p_t with respect to its starting pose (consisting of the 3 spatial coordinates and the yaw angle). The memory M is initialized to an empty set at the beginning of an episode, and at a time step t, is updated as $M = \{e_i^o : i = \max\{0, t-K\}, \ldots, t\}$. Apart from these embeddings, AVLEN also incorporates a goal estimation network f_g characterized by a convolutional neural network that produces a step-wise estimate $\hat{g}_t = f_g(F_t^B)$ of the sounding *AudioGoal*; $\hat{g}_t$ consisting of: (i) the (x, y) goal location estimate L_t from the current pose of the agent, and (ii) the goal category estimate $c_t \in \mathbb{R}^C$ for C semantic sounding object classes. The agent updates the current goal estimate combining the previous estimates as $g_t = \lambda \hat{g}_t + (1-\lambda) f_p(g_{t-1}, \Delta p_t)$ where f_p is a linear transformation of g_{t-1} using the pose change Δp_t. We use $\lambda = 0.5$, unless the sound is inaudible in which case it is set to zero. We will use $g \in G \subset \mathbb{R}^{C+2}$ to denote the space of goal estimates.

4.3.3 Multimodal Hierarchical Deep Reinforcement Learning

As is clear, the diverse input modalities used in AVLEN have distinctly varied levels of semantic granularity, and thus a single monolithic end-to-end RL policy for navigation might be sub-optimal. For example, the natural language navigation instructions received from the oracle might comprise rich details for navigating the scene that the agent need not

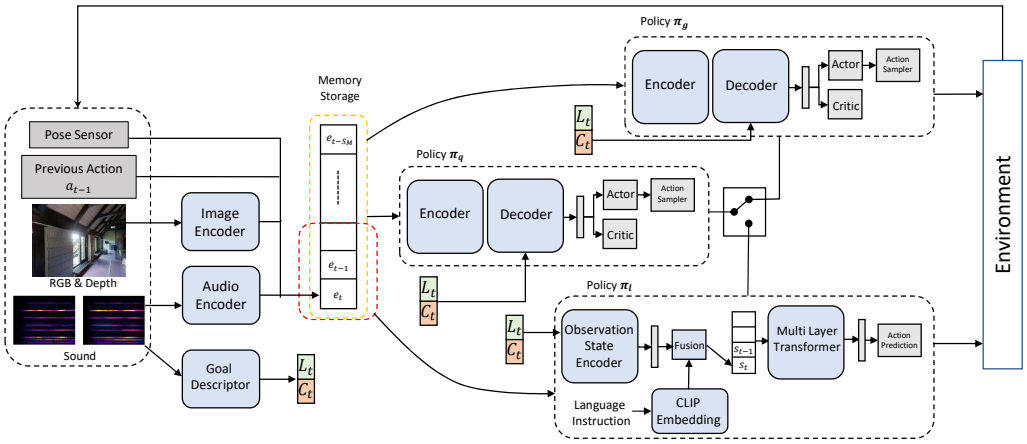


Figure 4.2: Architecture of our AVLEN pipeline. We show the two-level hierarchical RL policies that the model learns (offline), as well as the various input modalities and the control flow.

have to resort to any of the audio-visual inputs for say $\nu > 1$ steps. However, there is a budget on such queries and thus the agent must know when the query needs to be initiated (e.g., when the agent repeats sub-trajectories). Further, from a practical sense, each modality might involve different neural network architectures for processing, can have their own strategies for (pre-)training, involve distinct inductive baises, or incorporate heterogeneous navigation uncertainties.

All the above challenges naturally suggest to abstract the policy learning in the context of *hierarchical options* semi-Markov framework [6, 70, 138] consisting of low-level options corresponding to the navigation using either the *AudioGoal* or the language model, and a high-level policy to select among the options. More formally, an *option* is a triplet consisting of a respective policy ξ, a termination condition, and a set of belief states in which the option is valid. In our context, we assume a *multi-time* option policy for language-

based navigation spanning ν navigation steps[1] and a *primitive* policy [138] for *AudioGoal*. We also assume these options may be invoked independent of the agent state. Suppose $\pi_q : \mathbb{R}^{|\mathcal{S}|\times|M|} \times G \times \{\text{query}\} \to [0,1]$ represent the high-level policy deciding whether to query the oracle or not, using the current belief, the history M, and the goal estimate g. Further, let the two lower-level policies be: (i) $\pi_g : \mathbb{R}^{|\mathcal{S}|\times|M|} \times G \times A \to [0,1]$, that is tasked with choosing the navigation actions based on the audio-visual features, and (ii) $\pi_\ell : \mathbb{R}^{|\mathcal{S}|\times\nu} \times \mathcal{V}^N \times G \times A \to [0,1]$, that navigates based on the received natural language instruction formed using N words from the vocabulary $\mathcal{V}$, assuming ν steps are taken after each such query. Let R'_g and R'_ℓ denote the rewards (as defined in (4.1)) corresponding to the π_g and π_ℓ options, respectively, where we have the multi-time discounted cumulative reward (with penalty ζ) for $R'_\ell(b_t, a_t) = \mathbb{E}\left(\sum_{i=t}^{t+\nu-1} \gamma^{i-t} R'(b_i, a_i) | \pi_q = \pi_\ell, a_i \in A\right) - \zeta(t)$, while R'_g is, being a primitive option, as in (4.1) except that the actions are constrained to A. Then, we have the Bellman equation for using the options given by:

$$V^\pi(b) = \pi_q(\xi_g|b)\left[R'_g + \sum_{o'\in\mathcal{O}} P'(o'|b,\xi_g)V^\pi(b')\right] + \pi_q(\xi_\ell|b)\left[R'_\ell + \sum_{o'\in\mathcal{O}} P(o'|b,\xi_\ell)V^\pi(b')\right], \tag{4.2}$$

where ξ_g and ξ_ℓ are shorthands for $\xi = \pi_g$ and $\xi = \pi_\ell$, respectively, and $\pi = \{\pi_q, \pi_g, \pi_\ell\}$. Further, P' is the multi-time transition function given by: $P'(o'|b,\xi) = \sum_{j=1}^{\infty} \sum_{s'} \sum_{s} \gamma^j P(s', o', j|s, \xi) b(s)$, where with a slight abuse of notation, we assume $P(s', o', j)$ is the probability to observe o' in j steps using option ξ [138]. Our objective is to find the policies π that maximizes the value function in (4.2) for $V^\pi(b_0)$. Figure 4.2 shows a block diagram of the interplay between the

[1]Otherwise terminated if the stop action is called before the option policy is completed.

various policies and architectural components. Note that by using such a two-stage policy, we assume that the top-level policy π_q implicitly learns the uncertainty in the audio-visual and language inputs as well as the predictive uncertainty in the respective low-level options π_g and π_ℓ for reaching the goal state. In the next subsections, we detail the neural architectures for each of these options policies.

4.3.4 Navigation Using Audio Goal Policy, π_g.

Our policy network for π_g follows an architecture similar to [15], consisting of a Transformer encoder-decoder model [146] as shown in Figure 4.3. The encoder sub-module takes in the embedded features e^o from the current observation as well as such features from history stored in the memory M, while the decoder module takes in the output of the encoder concatenated with the goal descriptor g to produce a fixed dimensional feature vector, characterizing the current belief state b. An actor-critic network (consisting of a linear layer) then predicts an action distribution $\pi_g(b, .)$ and the value of this state. The agent then takes an action $a \sim \pi_g(b, .)$, takes a step, and receives a new observation. The goal descriptor network f_g outputs the object category c and the relative goal location estimation L. Following SAVi [15], we use off-policy category-level predictions and an on-policy location estimator.

4.3.5 Navigation Using Language Policy, π_ℓ

When an agent queries, it receives natural language instruction instr $\in \mathcal{V}^N$ from the oracle. Using instr and the current observation $e^o_t = (F^V_t, F^B_t, F^A_{t-1}, p_t)$, our language-based

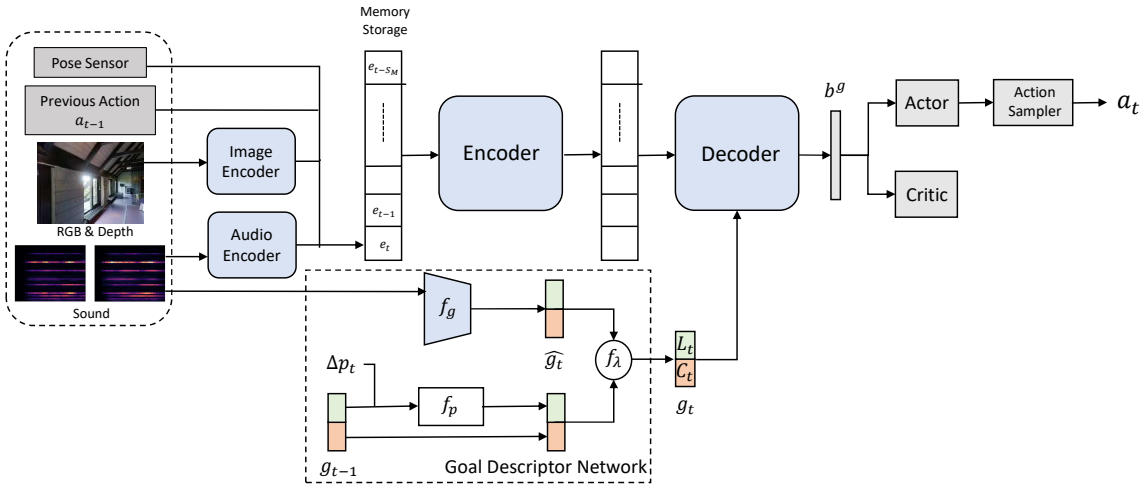


Figure 4.3: Network architecture for goal-based navigation policy π_g. The model architecture is similar to option selection/query policy π_q. However, the action space is different for these two policies.

navigation policy performs a sequence of actions $\langle a_t, a_{t+1}, \ldots, a_{t+\nu} \rangle$ as per π_ℓ option, where each $a_i \in A$. Specifically, for any step $\tau \in \langle t, \ldots, t+\nu-1 \rangle$, π_ℓ first encodes $\{e^o_\tau, g_\tau\}$ using a Transformer encoder-decoder network T_1[2], the output of this Transformer is then concatenated with CLIP [120] embeddings of the instruction, and fused using a fully-connected layer FC_1. The output of this layer is then concatenated with previous belief embeddings using a second multi-layer Transformer encoder-decoder T_2 to produce the new belief state b_τ, i.e.,

$$b_\tau = T_2\left(\text{FC}_1\left(T_1(e^o_\tau, g_\tau), \text{CLIP}(\text{instr})\right), \left\{b'_\tau : t < \tau' < \tau\right\}\right) \text{ and } \pi_\ell(b_\tau, .) = \text{softmax}(\text{FC}_2(b_\tau)). \tag{4.3}$$

Figure 4.4 illustrates the language-based navigation policy architecture.

[2]This is different Transformer from the one used for π_g, however taking g_τ as input to the decoder.

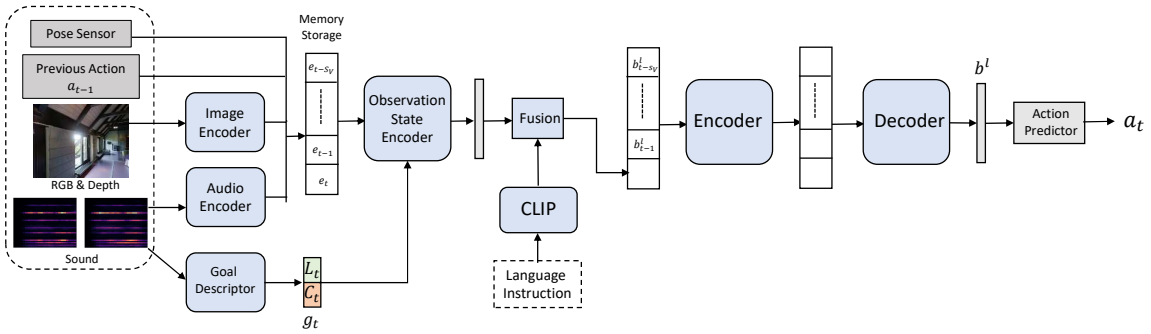


Figure 4.4: Network architecture for language-based navigation policy π_ℓ

4.3.6 Learning When-to-Query Policy, π_q

As alluded to above, the π_q policy decides when to query, i.e., when to use π_ℓ. Instead of directly utilizing model uncertainty [27], we use the reinforcement learning framework to be train this policy in an end-to-end manner, guided by the rewards ζ.

4.3.7 Reward Design

For the π_g policy, we assign a reward of $+1$ to reduce the geodesic distance towards the goal, a $+10$ reward to complete an episode successfully, i.e., calling the stop action near the *AudioGoal*, and a penalty of -0.01 per time step to encourage efficiency. As for the π_ℓ policy, we set a negative reward each time the agent queries the oracle, denoted ζ_q, as well as when the query is made within τ steps from previous query, denoted ζ_f. If the (softly)-allowed number of queries is K, then our combined negative reward function is given

by $\zeta_q + \zeta_f$, where

$$\zeta_q(k) = \begin{cases} \frac{k \times (r_{neg} + \exp(-\nu))}{\nu} & k < K \\ r_{neg} + \exp(-k) & k \geq K, \end{cases} \quad \text{and} \quad \zeta_f(j) = \begin{cases} \frac{r_f}{j} & 0 < j < \tau \\ 0 & \text{otherwise,} \end{cases} \tag{4.4}$$

where r_{neg} is set to -1.2, and r_f is set to -0.5. As a result, the agent learns when to interact with the oracle directly based on its current observation and history information. In the RL framework, the actor-critic model also predicts the value function of each state. Policy training is done using decentralized distributed proximal policy optimization (DD-PPO).

4.3.8 Policy Training

Learning π_g uses a two-stage training; in the first stage, the memory M is not used, while in the second stage, observation encoders are frozen, and the policy network is trained using both the current observation and the history in M. The training loss consists of (i) the value-function loss, (ii) policy network loss to estimate the actions correctly, and (iii) an entropy loss to encourage exploration. Our language-based navigation policy π_ℓ follows a two stage training as well. The first stage consists of an off-policy training. We re-purpose the fine-grained instructions provided by [58] to learn the language-based navigation policy π_ℓ. The second stage consists of on-policy training. During roll outs in our hierarchical framework, as the agent interacts with the oracle and receives language instructions, we use these instructions with the shortest path trajectory towards the goal to finetune π_ℓ. In both cases, it is trained with an imitation learning objective. Specifically, we allow the agent to navigate on the ground-truth trajectory by following teacher actions and calculate a

cross-entropy loss for each action in each step by, given by $-\sum_t a_{t^*} \log(\pi_\ell(b_t, a_t)$ where a^* is the ground truth action and $\pi_\ell(b_t, a_t)$ is the action probability predicted by π_ℓ.

4.3.9 Generating Oracle Navigation Instructions

The publicly available datasets for vision-and-language tasks contain a fixed number of route-and-instruction pairs at handpicked locations in the navigation grid. However, in our setup, a navigating agent can query an oracle at any point in the grid. To this end, we assume the oracle knows the shortest path trajectory s_path from the current agent location to the *AudioGoal*, and from which the oracle selects a segment consisting of n observation-action pairs (we use $n = 4$ in our experiments), i.e., s_path $= \langle (o_0, a_0), (o_1, a_1), \ldots, (o_{n-1}, a_{n-1}) \rangle$. With this assumption, we propose to mimic the oracle by a *speaker* model [42], which can generate a distribution of words $P^S(w|\,\text{s\_path})$[3]. The observation and action pairs are sequentially encoded using an LSTM encoder, $\left\langle F_0^S, F_1^S, \ldots, F_n^S \right\rangle = \text{SpeakerEncoder}(\text{s\_path})$ and decoded by another LSTM predicting the next word in the instruction by: $w_t = \text{SpeakerDecoder}(w_{t-1}, \left\langle F_0^S, F_1^S, \ldots, F_n^S \right\rangle)$. The instruction generation model is trained using the available (instruction, trajectory) pairs from the VLN dataset [4]. We use cross entropy loss and teacher forcing during training.

[3] s_path is approximated in the discrete Room-to-Room [4] environment and then used to generate instruction.

Table 4.1: Comparison of performances against state of the art in heard and unheard sound settings.

		Heard Sound					Unheard Sound				
	Feedback	Success ↑	SPL ↑	SNA ↑	DTG ↓	SWS ↑	Success ↑	SPL ↑	SNA ↑	DTG ↓	SWS ↑
Random Nav.	✗	1.4	3.5	1.2	17.0	1.4	1.4	3.5	1.2	17.0	1.4
ObjectGoal RL	✗	1.5	0.8	0.6	16.7	1.1	1.5	0.8	0.6	16.7	1.1
Gan et al. [45]	✗	29.3	23.7	**23.0**	11.3	14.4	15.9	12.3	11.6	12.7	8.0
Chen et al. [16]	✗	21.6	15.1	12.1	11.2	10.7	18.0	13.4	12.9	12.9	6.9
AV-WaN [17]	✗	20.9	16.8	16.2	10.3	8.3	17.2	13.2	12.7	11.0	6.9
SMT[38]+Audio	✗	22.0	16.8	16.0	12.4	8.7	16.7	11.9	10.0	12.1	8.5
SAVi [15]	✗	33.9	24.0	18.3	8.8	21.5	24.8	17.2	13.2	9.9	14.7
AVLEN	Language	**36.1**	**24.6**	19.7	**8.5**	**23.1**	**26.2**	**17.6**	**14.2**	**9.2**	**15.8**
AVLEN	GT Actions	**48.2**	**34.3**	**26.7**	**7.5**	**36.0**	**36.7**	**24.1**	**18.7**	**8.3**	**26.6**

Table 4.2: Comparisons in heard and unheard sound settings against varied query-triggering methods.

		Heard Sound					Unheard Sound				
	Feedback	Success ↑	SPL ↑	SNA ↑	DTG ↓	SWS ↑	Success ↑	SPL ↑	SNA ↑	DTG ↓	SWS ↑
Random	Language	32.5	21.1	16.1	8.93	21.8	23.5	14.8	11.5	9.9	14.3
Uniform	Language	33.2	22.4	17.8	9.1	22.0	22.1	14.6	11.5	9.8	13.3
Model Uncertainty	Language	34.2	24.0	19.5	8.7	20.5	24.9	16.1	13.5	9.3	15.2
AVLEN	Language	**36.1**	**24.6**	**19.7**	**8.5**	**23.1**	**26.2**	**17.6**	**14.2**	**9.2**	**15.8**

4.4 Experiments and Results

4.4.1 Dataset

We use the SoundSpaces platform [16] for simulating the world in which our AVLEN agent conducts the navigation tasks. Powered by Matterport3D scans [13], SoundSpaces facilitates a realistic simulation of a potentially-complex 3D space navigable by the agent along a densely sampled grid with 1m grid-cell sides. The platform also provides access to panoramic ego-centric views of the scene in front of the agent both as RGB and as depth images, while also allowing the agent to hear realistic binaural audio of acoustic events in the 3D space. To benchmark our experiments, we use the semantic audio-visual navigation dataset from Chen et al. [15] built over SoundSpaces. This dataset consists of sounds from 21 semantic categories of objects that are visually present in the Matterport3D scans. The object-specific sounds are generated at the location of the Matterport3D objects. In each

Table 4.3: Comparisons against varied query-triggering methods with ground truth action as feedback.

		Heard Sound					Unheard Sound				
	Feedback	Success ↑	SPL ↑	SNA ↑	DTG ↓	SWS ↑	Success ↑	SPL ↑	SNA ↑	DTG ↓	SWS ↑
Random	GT Actions	39.8	30.0	24.5	7.6	23.3	29.6	22.1	18.4	8.2	16.3
Uniform	GT Actions	38.8	29.9	25.5	7.4	20.3	28.6	21.3	18.0	**7.8**	14.8
Model Uncertainty	GT Actions	41.3	30.6	24.8	**7.3**	26.3	31.4	22.7	18.4	8.2	19.3
AVLEN	GT Actions	**48.2**	**34.3**	**26.7**	7.5	**36.0**	**36.7**	**24.1**	**18.7**	8.3	**26.6**

navigation episode, the duration of the sounds are variable and is normal-distributed with a mean 15s and deviation 9s, clipped for a minimum 5s and maximum 500s [15]. There are 0.5M/500/1000 episodes available in this dataset for train/val/test splits respectively from 85 Matterport3D scans.

4.4.2 Evaluation Metrics

Similar to [15], we use the following standard metrics for evaluating our navigation performances on this dataset: i) *success rate* for reaching the *AudioGoal*, ii) *success weighted by inverse path length* (SPL) [3], iii) *success weighted by inverse number of actions* (SNA) [17], iv) *average distance to goal* (DTG), and v) *success when silent* (SWS). SWS refers to the fraction of successful episodes when the agent reaches the goal after the end of the acoustic event.

4.4.3 Implementation Details

Similar to prior works, we use RGB and depth images, center-cropped to 64×64. The agent receives binaural audio clip as 65×26 spectrograms. The memory size for π_g and π_q is 150 and for π_ℓ is 3. All the experiments consider maximum $K = 3$ allowed queries (unless otherwise specified). For each query, the agent will take $\nu = 3$ navigation steps in the

Table 4.4: Comparison of AVLEN performances against baselines and when-to-query approaches in the *presence of distractor sound*

	Feedback	Success ↑	SPL ↑	SNA ↑	DTG ↓	SWS ↑
Chen et al. [16]	✗	4.0	2.4	2.0	14.7	2.3
AV-WaN [17]	✗	3.0	2.0	1.8	14.0	1.6
SMT[38]+Audio	✗	4.2	2.9	2.1	14.9	2.8
SAVi [15]	✗	11.8	7.4	5.0	13.1	8.4
Random	Language	11.6	6.6	4.8	12.9	7.8
Uniform	Language	11.6	6.8	5.1	13.3	7.7
Model Uncertainty	Language	12.4	6.7	5.0	12.8	8.4
AVLEN	Language	**14.0**	**8.4**	**5.9**	**12.8**	**11.1**
AVLEN	GT Actions	**24.4**	**15.3**	**11.3**	**11.3**	**21.5**

environment using the natural language instruction. We use a vocabulary with 1621 words. Training uses ADAM [65] with learning rate 2.5×10^{-4}. Refer to the Appendix for more details.

4.4.4 Experimental Results and Analysis

The main objective of our AVLEN agent in the semantic audio-visual navigation task is to navigate towards a sounding object in an unmapped 3D environment when the sound is sporadic. Since we are the first to integrate oracle interactions (in natural language) within this problem setting, we compare with existing state-of-the-art semantic audio-visual navigation approaches, namely Gan et al. [45], Chen et al. [16], AV-WaN [17], SMT [38] + Audio, and SAVi [15]. Following the protocol used in [15] and [16], we evaluate performance of the same trained model on two different sound settings: i) *heard sound*, in which the sounds used during test are heard by the agent during training, and ii) *unheard sound*, in which the train and test sets use distinct sounds. In both experimental settings, the test

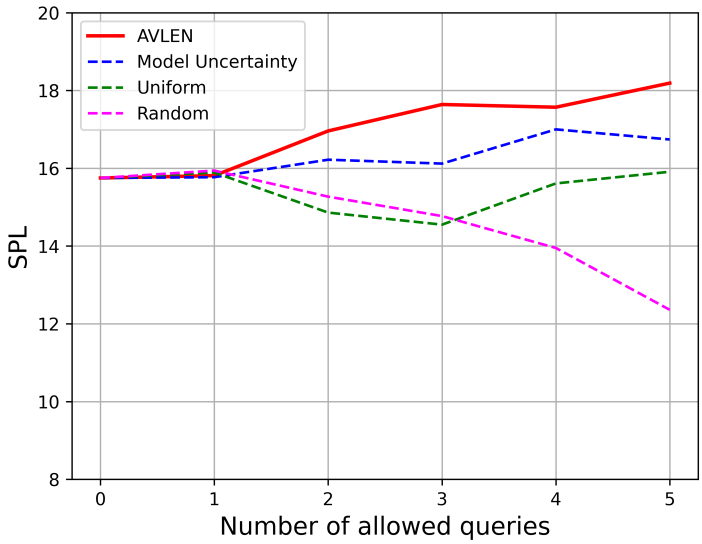


Figure 4.5: Performance (SPL) comparison against varying the number of allowed queries.

environments are unseen.

Table 4.1 provides the results of our experiments using heard and unheard sounds. The table shows that AVLEN (language) – which is our full model based on language feedback – shows a +2.2% and +1.6% absolute gain in success rate and success-when-silent (SWS) respectively, compared to the best performing baseline SAVi [15] for heard sound. Moreover, we obtain 1.4% and 1.1% absolute gain in success rate and SWS respectively for unheard sound compared to the next best method, SAVi. Our results clearly demonstrate that the agent is indeed able to use the short natural language instructions for improving the navigation.

A natural question to ask in this setting is: *Why are the improvements not so dramatic, given the agent is receiving guidance from an oracle?* Generally, navigation based on language instructions is a challenging task in itself ([4, 97]) since language incorporates strong

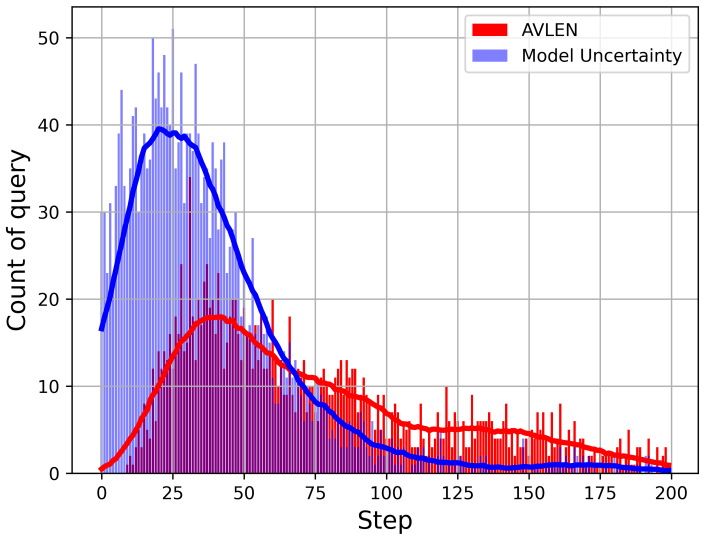


Figure 4.6: Distribution of queries triggered against the time steps in episodes.

	Feedback	Success ↑	SPL ↑	SNA ↑	DTG ↓	SWS ↑
SAVi	✗	33.9	24.0	18.3	8.8	21.5
AVLEN (Glove + GRU)	Language	36.1	24.6	19.5	8.5	23.3
AVLEN (Glove + Transformer)	Language	36.2	25.4	20.0	8.4	23.8
AVLEN (CLIP+Transformer)	Language	36.1	24.6	19.7	8.5	23.1
AVLEN (CLIP + GRU)	Language	37.7	25.5	19.9	8.5	25.1

Table 4.5: Comparisons in performance for different architectural choices for language-based policy π_ℓ in heard sound setting.

inductive biases and usually spans large vocabularies; as a result the action predictions can be extremely noisy, imprecise, and misguiding. However, the key for improved performance is to identify *when to query*. Our experiments show that AVLEN is able to identify when to query correctly (also see Table 4.2) and thus improve performance. To further substantiate this insight, we designed an experiment in which the agent is provided the ground truth (GT) navigation actions as feedback (instead of providing the corresponding language instruction) whenever a query is triggered. The results of this experiment – marked AVLEN (GT Actions)

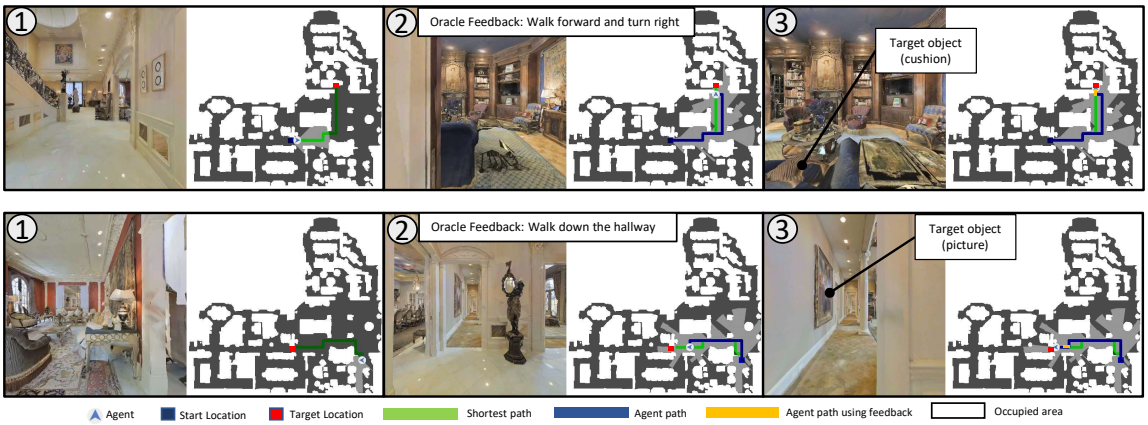


Figure 4.7: Two qualitative results from AVLEN's navigation trajectories. We show egocentric views and top down maps for three different viewpoints in agent's trajectory. The agent starts from ①, receives oracle help in ②, navigates to the goal in ③. In the top episode, agent receives directional information ('Walk forward and turn right'), whereas in the bottom episode, agent receives language instruction more grounded on the scene ('Walk down the hallway').

in Table 4.1 – clearly show an improvement in success rate by nearly +15% for heard sounds and +12% for unheard sounds, suggesting future work to consider improving language-based navigation policy π_ℓ.

Navigation Under Distractor Sounds. Next, we consider the presence of distractor sounds while navigating towards an "unheard" sound source as provided in SAVi [15]. In this setting, the agent must know which sound its target is. Thus, a one hot encoding of the target is also provided as an input to the agent, if there are multiple sounds in the environment. The presence of distractor sounds makes the navigation task even more challenging, and intuitively, the agent's ability to interact with the oracle could come useful. This insight is clearly reflected in Figure 4.4, where AVLEN shows 2.2% and 2.7% higher success rate and SWS respectively compared to baseline approaches. Figure 4.4 shows that

AVLEN outperforms other query procedures as well and the performance difference is more significant compared to when there is no distractor sound.

Analyzing When-to-Query. To evaluate if AVLEN is able to query at the appropriate moments, we designed potential baselines that one could consider for deciding when-to-query and compare their performances in Table 4.2. Specifically, the considered baselines are: i) *Uniform*: queries after every 15 steps, ii) *Random*: queries randomly within the first 50 steps of each episode, and iii) *Model Uncertainty (MU)* based on [27]: queries when the softmax action probabilities of the top two action predictions of π_g have an absolute difference are less than a predefined threshold (≤ 0.1). Table 4.2 shows our results. Unsurprisingly, we see that Random and Uniform perform poorly, however using MU happen to be a strong baseline. However, MU is a computationally expensive baseline as it requires a forward pass on the π_g model to decide when to query. Even so, we observe that compared to all the baselines, AVLEN shows better performance in all metrics. To further understand this, in Figure 4.6 we plot the distribution of the episode time step when the query is triggered for the unheard sound setting. As is clear, MU is more likely to query in the early stages of the episode, exhausting the budget, while AVLEN learns to distribute the queries throughout the steps, suggesting our hierarchical policy π_q is learning to be conservative, and judicial in its task. Interestingly we also find reasonable overlap between the peaks of the two curves, suggesting that π_q is considering the predictive uncertainty of π_g implicitly.

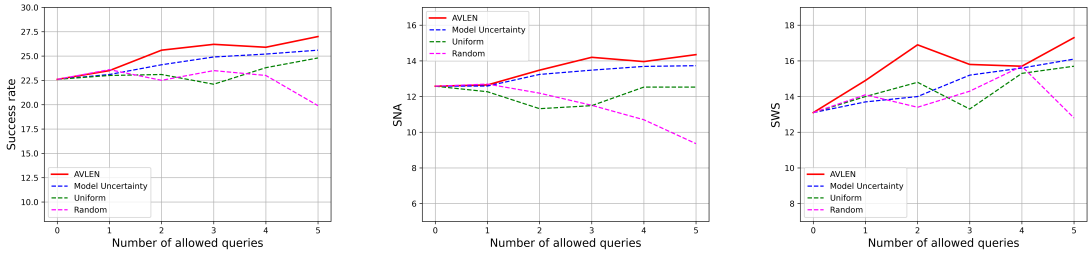
Is the AVLEN reward strategy effective to train the model to query appropriately? To answer this question, we analyzed the effectiveness of our training reward strategy (i.e., to make the agent learn when to query) by comparing the performances of AVLEN

with the other baseline querying methods (i.e.,*Random*, *Uniform*, and *Model Uncertainty*) when the oracle feedback provides the ground truth navigation actions, instead of language instructions. The results are provided in Table 4.3 and clearly show the superior performances of AVLEN against the alternatives suggesting that the proposed reward signal is effective for appropriately biasing the query selection policy.

Are the improvements in performance from better Transformer and CLIP models? To understand the architectural design choices (e.g., modules used in π_ℓ) versus learning the appropriate policy to query (learning of π_q), we conducted an ablation study replacing the sub-modules in π_ℓ neural network with close alternatives. Specifically, we consider four architectural variants for π_ℓ for comparisons, namely: (i) AVLEN (Glove + GRU), (ii) AVLEN (Glove + Transformer), (iii) AVLEN (CLIP + GRU), and (iv) AVLEN (CLIP + Transformer). We compare the performances in Table 4.5. Our results show that the improvements when using AVLEN is not due to CLIP context or Transformer representations alone, instead it is from our querying framework, that consistently performs better than the baseline SAVi model, irrespective of the various ablations.

Sensitivity to Allowed Number of Queries, ν. To check the sensitivity AVLEN for different number of allowed queries (and thus the number of natural language instructions received), we consider $\nu \in \{0, 1, 2, 3, 4, 5\}$ and evaluate the performances. Figure 4.5 shows the SPL scores this experiment in the unheard sound setting. As is expected, increasing the number of queries leads to an increase in SPL for AVLEN, while alternatives e.g., Random drops quickly; the surprising behaviour of the latter is perhaps a mix of querying at times when the π_g model is confident and the π_ℓ instructions being noisy.

Sensitivity to Allowed Number of Queries. To check the sensitivity of AVLEN for a different number of allowed queries, we consider a set of allowed query numbers $\nu = \{0, 1, 2, 3, 4, 5\}$ and evaluate performance. Figure 4.8 shows the success rate, SNA and SWS metric for allowed queries $\in \{0, 1, 2, 3, 4, 5\}$ in presence of unheard sound. For the metrics, AVLEN retains an advantage over other approaches.

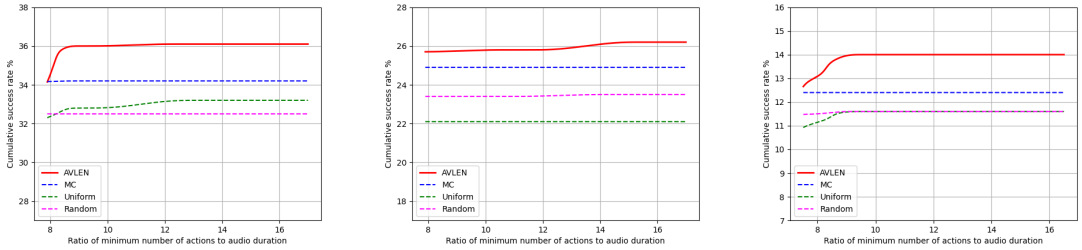


(a) Query sensitivity (SR) (b) Query sensitivity (SNA) (c) Query sensitivity (SWS)

Figure 4.8: Sensitivity to the number of queries ν to the oracle that AVLEN can make. The results are for the unheard sound sceneario. Please see the main paper for plots on the success rate.

Robustness to Silence Duration. Figure 4.9 shows the cumulative success of different approaches. The x axis represents the silent ratio (ratio of the minimum number of actions required to reach the goal to the duration of audio). A point (x, y) on this plot means the fraction of successful episodes with ratios up to x among all episodes is y. When this ratio is greater than 1, no agent can reach the goal before the audio stops. The greater this ratio is, the longer the fraction of silence, and hence the harder the episode. We observe that AVLEN results in higher cumulative success when sound is silent for longer period.

Vision-Language Navigation Performance. In our setting, an agent receives natural language instruction when it queries. It needs to "comprehend" this instruction



(a) Heard sound (b) Unheard sound (c) Distractor sound

Figure 4.9: Robustness to silence duration analysis

properly and should take navigation steps grounded on this instruction. To analyze if π_ℓ (the language policy) takes navigation steps well-grounded on the instruction, we created a VLN test-set of $7,031$ short instruction-trajectory pairs. These short trajectories aligns/overlaps with segments of test-set trajectories from semantic audio-visual navigation dataset. We analyzed the performance of **VLN-b**: trained on repurposed fine-grained instruction from [58], **VLN-f**: fine-tuned π_ℓ with collected trajectory-instruction pairs in AVLEN training, and **VLN-b (w/o instruction)** (language instruction masked) in the VLN test-set. In Table 4.6, evaluation metric $step-n$ reflects the percentage of episodes that took n sequential steps correctly. Table 4.6 shows that there is a significant drop in performance if the language is masked out (removed), which indicates π_ℓ predictions are grounded on the instruction. Also, fine-tuning π_ℓ policy with collected trajectory-instruction pairs in an online manner helps improve the performance.

	$Step-1$	$Step-2$	$Step-3$
VLN-b (w/o instruction)	51.3	22.2	17.0
VLN-b	62.8	47.3	37.8
VLN-f	**65.9**	**55.5**	**45.3**

Table 4.6: Vision-language navigation performance.

Qualitative Results. Figure 4.7 provides snapshots from two example episodes of semantic audio-visual navigation using AVLEN. The sounding object is a 'cushion' in the first episode (first row) and a 'picture' in the second episode (second row). ② of both episodes shows the viewpoint where agent queries and receive natural language instructions. In the first row, agent receives directional information: 'Walk forward and turn right', while in the second row episode, the agent receives language instruction ground in the scene: 'Walk down the hallway'. In both cases, the agent uses the instruction to assist its navigation task and reach around the vicinity of target object ③.

4.5 Conclusions

The ability to interact with oracle/human using natural language instructions to solve difficult tasks is of great importance from a human-machine interaction standpoint. In this work, we considered such a task in the context of audio-visual-language embodied navigation in a realistic virtual world, enabled by the SoundSpaces simulator. The agent, visualy navigating the scene to localize an audio goal, is also equipped with the possibility of asking an oracle for help. We modeled the problem as one of learning a multimodal hierarchical reinforcement learning policy, with a two-level policy model: higher-level policy to decide when to ask questions, and lower-level policies to either navigate using the audio-goal or follow the oracle instructions. We presented experiments using our proposed framework; our results show that using the proposed policy allows the agent achieve higher success rates on the semantic audio-visual navigation task, especially in cases when the navigation task is more difficult in presence of distractor sounds.

Chapter 5

Conclusions

5.1 Dissertation Summary

Even with the significant strides toward multimodal learning in recent times, there are many challenges in the multimodal domain that require rigorous investigation and the development of newer appropriate approaches. One such challenge is to understand and learn the alignment of multimodal information together. In this dissertation, our focus was to learn and understand the alignment of multiple modality information for static and dynamic tasks. By alignment, we refer to the problem of finding and understanding correspondence between instances of two different modalities. It also includes the realization of uncertainty when a model fails to relate two modalities. We address the alignment issue for two static tasks and one dynamic task.

In Chapter 2, we discussed the task of text-based temporal localization of video moments in a collection of videos. This task poses a unique challenge as the system is required to perform: (i) retrieval of the relevant video where only a segment of the video corresponds

with the queried sentence, and (ii) temporal localization of moment in the relevant video based on sentence query. We proposed a Hierarchical Moment Alignment Network (HMAN) which learns an effective joint embedding space for moments and sentences. In addition to learning subtle differences between intra-video moments, HMAN focuses on distinguishing inter-video global semantic concepts based on sentence queries.

In Chapter 3, we considered the task of text-based temporal localization of novel events. Models optimized for a fixed set of seen events are unlikely to generalize to the practical requirement of localizing a wider range of events, some of which may be unseen. In this regard, we formulated the inference task of text-based localization of moments as a relational prediction problem, hypothesizing a conceptual relation between semantically relevant moments, e.g., a temporally relevant moment corresponding to an unseen text query and a moment corresponding to a seen text query may contain shared concepts. The likelihood of a candidate moment being the correct one based on an unseen text query will depend on its relevance to the moment corresponding to the semantically most relevant seen query.

In Chapter 4, we looked into the dynamic task of audio-visual-language navigation. The real world is not only multimodal, but also often complex. Thus agents need to understand the uncertainty in their actions and seek instructions to navigate. To this end, we have presented AVLEN – an interactive agent for Audio-Visual-Language Embodied Navigation. The goal of our embodied agent is to localize an audio event via navigating the 3D visual world; however, the agent may also seek help from a human (oracle), where the assistance is provided in free-form natural language. To realize these abilities, AVLEN uses a

multimodal hierarchical reinforcement learning backbone that learns: (a) high-level policies to choose either audio-cues for navigation or to query the oracle, and (b) lower-level policies to select navigation actions based on its audio-visual and language inputs. The policies are trained via rewarding for the success on the navigation task while minimizing the number of queries to the oracle.

5.2 Future Research Directions

Continuing in the line of the aforementioned works, we discuss some interesting research directions for future works.

5.2.1 Webly Supervised/Knowledge Transfer

Continuing our work on text-based temporal localization for novel events where data available in the training set is used as the support information, we can improve it further utilizing a vast amount of web information. How to utilize the knowledge available on the web would be an important future direction of work.

5.2.2 Bi-directional Interaction of Navigating Agent

In our considered audio-visual-language navigation task, the agent was only able to query for help. However, a more natural scenario would be where the agent can ask questions and be able to have bi-directional interaction with an oracle for navigation purposes. This would be an interesting and challenging task to solve considering the requirement of bridging efforts from conversation AI, uncertainty estimation, and multimodal machine

learning knowledge.

5.2.3 Discrete Alignment vs Continuous Alignment

Contrary to matching discrete elements across modalities, adversarial training can be used to warp the representation of one modality into another modality. This is likely to be an important research direction if the task at hand considers domain shifts in data.

5.2.4 Few-shot Capability

Large language model [10] shows few-shot capability, where it can achieve strong performance using a few examples of a task provided to the model. Utilizing the abundance of data on the web, vision-language models can be pretrained to show few-shot ability.

5.2.5 Learning with Noisy Annotation

Human annotations are expensive and tedious effort is required to collect clean annotations. On the other hand, web-crawled data can be an unlimited source of annotated information. However, web annotation is noisy and systems are required to be robust to the noise. Again, every vision-language task is unique and requires independent investigation of how to learn in presence of noisy annotation.

5.2.6 Scalable Vision-language Model

While it is shown that cross-modal interaction results in better representation learning, a model dependent on cross-modal interaction might not be scalable for different tasks (e.g., text-based moment retrieval from video collection). In that case, an important

research direction would be how to utilize a powerful cross-modal model for representation learning and then distill the information to smaller modality-specific models.

5.2.7 Task Adaptation

Recently vision-language pretraining (VLP) has gained a lot of interest. These models are trained for representation learning where the learned representation can be utilized for different tasks. These VLP models are further augmented with more network weights for different tasks. These augmentations are task-dependent and vary from one task to another. As a result, an important research direction would be how to adapt networks specifically designed for a particular task to other tasks without modifying the network. In that manner, same model can be used for different tasks.

5.2.8 Enforcing Sparsity on Transformer

Transformer architecture utilizes an attention mechanism where each input is connected to other inputs using self-attention. However, these densely connected graphs might be overkill for many applications and easily overfit if the amount of data is low. In that sense, how to develop a Transformer architecture with sparse connections which can be trained with a limited amount of data would be an interesting research direction to explore.

Bibliography

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(7):1425–1438, 2015.

[2] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2927–2936, 2015.

[3] Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, et al. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018.

[4] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683, 2018.

[5] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017.

[6] Pierre-Luc Bacon, Jean Harb, and Doina Precup. The option-critic architecture. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[7] Yang Bai, Junyan Wang, Yang Long, Bingzhang Hu, Yang Song, Maurice Pagnucco, and Yu Guan. Discriminative latent semantic graph for video captioning. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3556–3564, 2021.

[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.

[9] Pratyay Banerjee, Tejas Gokhale, Yezhou Yang, and Chitta Baral. Weakly supervised relative spatial reasoning for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1908–1918, 2021.

[10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[11] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733. IEEE, 2017.

[12] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

[13] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017.

[14] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar. Rethinking the faster r-cnn architecture for temporal action localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1130–1139, 2018.

[15] Changan Chen, Ziad Al-Halah, and Kristen Grauman. Semantic audio-visual navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15516–15525, 2021.

[16] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *European Conference on Computer Vision*, pages 17–36. Springer, 2020.

[17] Changan Chen, Sagnik Majumder, Ziad Al-Halah, Ruohan Gao, Santhosh Kumar Ramakrishnan, and Kristen Grauman. Learning to set waypoints for audio-visual navigation. In *International Conference on Learning Representations*, 2021.

[18] David L Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 190–200. Association for Computational Linguistics, 2011.

[19] Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12538–12547, 2019.

[20] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. Temporally grounding natural sentence in video. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 162–171, 2018.

[21] Shaoxiang Chen, Wenhao Jiang, Wei Liu, and Yu-Gang Jiang. Learning modality interaction for temporal sentence localization and event captioning in videos. In *European Conference on Computer Vision*, pages 333–351. Springer, 2020.

[22] Shaoxiang Chen and Yu-Gang Jiang. Towards bridging event captioner and sentence localizer for weakly supervised dense event captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8425–8435, 2021.

[23] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10638–10647, 2020.

[24] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. 2019.

[25] Zhenfang Chen, Lin Ma, Wenhan Luo, and Kwan-Yee K Wong. Weakly-supervised spatio-temporally grounding natural sentence in video. *arXiv preprint arXiv:1906.02549*, 2019.

[26] Jingze Chi and Yuxin Peng. Dual adversarial networks for zero-shot cross-media retrieval. In *IJCAI*, pages 663–669, 2018.

[27] Ta-Chung Chi, Minmin Shen, Mihail Eric, Seokhwan Kim, and Dilek Hakkani-tur. Just ask: An interactive learning framework for vision and language navigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2459–2466, 2020.

[28] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[29] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–10, 2018.

[30] Cheng Deng, Zhaojia Chen, Xianglong Liu, Xinbo Gao, and Dacheng Tao. Triplet-based deep hashing network for cross-modal retrieval. *IEEE Transactions on Image Processing*, 27(8):3893–3903, 2018.

[31] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[32] Xinpeng Ding, Nannan Wang, Shiwei Zhang, De Cheng, Xiaomeng Li, Ziyuan Huang, Mingqian Tang, and Xinbo Gao. Support-set based cross-supervision for video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11573–11582, 2021.

[33] Jianfeng Dong, Xirong Li, and Cees GM Snoek. Predicting visual features from text for image and video caption retrieval. *IEEE Transactions on Multimedia*, 20(12):3377–3388, 2018.

[34] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang. Dual encoding for zero-example video retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9346–9355, 2019.

[35] Xingping Dong and Jianbing Shen. Triplet loss in siamese network for object tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 459–474, 2018.

[36] Victor Escorcia, Mattia Soldan, Josef Sivic, Bernard Ghanem, and Bryan Russell. Temporal localization of moments in video collections with natural language. *arXiv preprint arXiv:1907.12763*, 2019.

[37] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017.

[38] Kuan Fang, Alexander Toshev, Li Fei-Fei, and Silvio Savarese. Scene memory transformer for embodied agents in long-horizon tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 538–547, 2019.

[39] Zhengcong Fei. Attention-aligned transformer for image captioning. 2022.

[40] Guang Feng, Zhiwei Hu, Lihe Zhang, and Huchuan Lu. Encoder fusion network with co-attention embedding for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15506–15515, 2021.

[41] Zerun Feng, Zhimin Zeng, Caili Guo, and Zheng Li. Exploiting visual semantic reasoning for video-text retrieval. *arXiv preprint arXiv:2006.08889*, 2020.

[42] Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. Speaker-follower models for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 31, 2018.

[43] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2121–2129, 2013.

[44] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. *arXiv preprint arXiv:2203.13131*, 2022.

[45] Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B Tenenbaum. Look, listen, and act: Towards audio-visual embodied navigation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9701–9707. IEEE, 2020.

[46] Feng Gao, Qing Ping, Govind Thattai, Aishwarya Reganti, Ying Nian Wu, and Prem Natarajan. A thousand words are worth more than a picture: Natural language-centric outside-knowledge visual question answering. *arXiv preprint arXiv:2201.05299*, 2022.

[47] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017.

[48] Junyu Gao and Changsheng Xu. Fast video moment retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1523–1532, 2021.

[49] Runzhou Ge, Jiyang Gao, Kan Chen, and Ram Nevatia. Mac: Mining activity concepts for language-based temporal localization. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 245–253. IEEE, 2019.

[50] Soham Ghosh, Anuva Agarwal, Zarana Parekh, and Alexander Hauptmann. Excl: Extractive clip localization using natural language descriptions. *arXiv preprint arXiv:1904.02755*, 2019.

[51] Meera Hahn, Asim Kadav, James M Rehg, and Hans Peter Graf. Tripping through time: Efficient localization of activities in videos. *arXiv preprint arXiv:1904.09936*, 2019.

[52] Ligong Han, Jian Ren, Hsin-Ying Lee, Francesco Barbieri, Kyle Olszewski, Shervin Minaee, Dimitris Metaxas, and Sergey Tulyakov. Show me what and tell me how: Video synthesis via multimodal conditioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3615–3625, 2022.

[53] Dongliang He, Xiang Zhao, Jizhou Huang, Fu Li, Xiao Liu, and Shilei Wen. Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8393–8400, 2019.

[54] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[55] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970. IEEE, 2015.

[56] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with temporal language. In *Proceedings*

of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1380–1390, 2018.

[57] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[58] Yicong Hong, Cristian Rodriguez-Opazo, Qi Wu, and Stephen Gould. Sub-instruction aware vision-and-language navigation. *arXiv preprint arXiv:2004.02707*, 2020.

[59] Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. Vln bert: A recurrent vision-and-language bert for navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1643–1653, June 2021.

[60] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

[61] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17980–17989, 2022.

[62] Jiabo Huang, Yang Liu, Shaogang Gong, and Hailin Jin. Cross-sentence temporal and semantic relations in video activity localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7199–7208, 2021.

[63] Bin Jiang, Xin Huang, Chao Yang, and Junsong Yuan. Cross-modal video moment retrieval with spatial and language-temporal attention. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 217–225, 2019.

[64] Liyiming Ke, Xiujun Li, Yonatan Bisk, Ari Holtzman, Zhe Gan, Jingjing Liu, Jianfeng Gao, Yejin Choi, and Siddhartha Srinivasa. Tactical rewind: Self-correction via backtracking in vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6741–6749, 2019.

[65] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[66] Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *European Conference on Computer Vision*, pages 104–120. Springer, 2020.

[67] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *International Conference on Computer Vision (ICCV)*, 2017.

[68] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE transactions on pattern analysis and machine intelligence*, 36(3):453–465, 2013.

[69] Duy-Dinh Le, Sang Phan, Vinh-Tiep Nguyen, Benjamin Renoust, Tuan A Nguyen, Van-Nam Hoang, Thanh Duc Ngo, Minh-Triet Tran, Yuki Watanabe, Martin Klinkigt, et al. Nii-hitachi-uit at trecvid 2016. 2016.

[70] Tuyen P Le, Ngo Anh Vien, and TaeChoong Chung. A deep hierarchical reinforcement learning algorithm in partially observable markov decision processes. *Ieee Access*, 6:49089–49102, 2018.

[71] Colin Lea, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks: A unified approach to action segmentation. In *European Conference on Computer Vision*, pages 47–54. Springer, 2016.

[72] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018.

[73] Peng Lei and Sinisa Todorovic. Temporal deformable residual networks for action segmentation in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6742–6751, 2018.

[74] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11336–11344, 2020.

[75] Jingjing Li, Mengmeng Jing, Ke Lu, Zhengming Ding, Lei Zhu, and Zi Huang. Leveraging the invariant side of generative zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7402–7411, 2019.

[76] Shasha Li, Shitong Zhu, Sudipta Paul, Amit Roy-Chowdhury, Chengyu Song, Srikanth Krishnamurthy, Ananthram Swami, and Kevin S Chan. Connecting the dots: Detecting adversarial perturbations using context inconsistency. In *European Conference on Computer Vision*, pages 396–413. Springer, 2020.

[77] Xirong Li, Chaoxi Xu, Gang Yang, Zhineng Chen, and Jianfeng Dong. W2vv++ fully deep learning for ad-hoc video search. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1786–1794, 2019.

[78] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.

[79] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. Storygan: A sequential conditional gan for story visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6329–6338, 2019.

[80] Kaiyi Lin, Xing Xu, Lianli Gao, Zheng Wang, and Heng Tao Shen. Learning cross-aligned latent embeddings for zero-shot cross-modal retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11515–11522, 2020.

[81] Tianwei Lin, Xu Zhao, and Zheng Shou. Single shot temporal action detection. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 988–996, 2017.

[82] Zhijie Lin, Zhou Zhao, Zhu Zhang, Qi Wang, and Huasheng Liu. Weakly-supervised video moment retrieval via semantic completion network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11539–11546, 2020.

[83] Zhijie Lin, Zhou Zhao, Zhu Zhang, Zijian Zhang, and Deng Cai. Moment retrieval via cross-modal interaction networks with query reconstruction. *IEEE Transactions on Image Processing*, 29:3750–3762, 2020.

[84] Bingbin Liu, Serena Yeung, Edward Chou, De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. Temporal modular networks for retrieving complex compositional activities in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 552–568, 2018.

[85] Chong Liu, Fengda Zhu, Xiaojun Chang, Xiaodan Liang, Zongyuan Ge, and Yi-Dong Shen. Vision-language navigation with random environmental mixup. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1644–1654, 2021.

[86] Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. Context-aware biaffine localizing network for temporal sentence grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11235–11244, 2021.

[87] Daizong Liu, Xiaoye Qu, Xiao-Yang Liu, Jianfeng Dong, Pan Zhou, and Zichuan Xu. Jointly cross-and self-modal graph attention network for query-based moment localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4070–4078, 2020.

[88] Li Liu, Zijia Lin, Ling Shao, Fumin Shen, Guiguang Ding, and Jungong Han. Sequential discrete hashing for scalable cross-modality similarity retrieval. *IEEE Transactions on Image Processing*, 26(1):107–118, 2016.

[89] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. Attentive moment retrieval in videos. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 15–24, 2018.

[90] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. Cross-modal moment localization in videos. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 843–851, 2018.

[91] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487*, 2019.

[92] Xiankai Lu, Chao Ma, Jianbing Shen, Xiaokang Yang, Ian Reid, and Ming-Hsuan Yang. Deep object tracking with shrinkage loss. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[93] Corey Lynch, Mohi Khansari, Ted Xiao, Vikash Kumar, Jonathan Tompson, Sergey Levine, and Pierre Sermanet. Learning latent plans from play. In *Conference on robot learning*, pages 1113–1132. PMLR, 2020.

[94] Chih-Yao Ma, Jiasen Lu, Zuxuan Wu, Ghassan AlRegib, Zsolt Kira, Richard Socher, and Caiming Xiong. Self-monitoring navigation agent via auxiliary progress estimation. *arXiv preprint arXiv:1901.03035*, 2019.

[95] Chih-Yao Ma, Zuxuan Wu, Ghassan AlRegib, Caiming Xiong, and Zsolt Kira. The regretful agent: Heuristic-aided navigation through progress estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6732–6740, 2019.

[96] Yiyang Ma, Huan Yang, Bei Liu, Jianlong Fu, and Jiaying Liu. Ai illustrator: Translating raw descriptions into images by prompt-based cross-modal generation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4282–4290, 2022.

[97] Arjun Majumdar, Ayush Shrivastava, Stefan Lee, Peter Anderson, Devi Parikh, and Dhruv Batra. Improving vision-and-language navigation with image-text pairs from the web. In *European Conference on Computer Vision*, pages 259–274. Springer, 2020.

[98] Foteini Markatopoulou, Damianos Galanopoulos, Vasileios Mezaris, and Ioannis Patras. Query and keyframe representations for ad-hoc video search. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pages 407–411, 2017.

[99] Niluthpol C Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. Joint embeddings with multimodal cues for video-text retrieval. *International Journal of Multimedia Information Retrieval*, pages 1–16, 2019.

[100] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *ACM International Conference on Multimedia Retrieval (ICMR)*. ACM, 2018.

[101] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K. Roy-Chowdhury. Weakly supervised video moment retrieval from text queries. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[102] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10810–10819, 2020.

[103] Jinwoo Nam, Daechul Ahn, Dongyeop Kang, Seong Jong Ha, and Jonghyun Choi. Zero-shot natural language video localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1470–1479, October 2021.

[104] Guoshun Nan, Rui Qiao, Yao Xiao, Jun Liu, Sicong Leng, Hao Zhang, and Wei Lu. Interventional video grounding with dual contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2765–2775, 2021.

[105] Khanh Nguyen and Hal Daumé III. Help, anna! visual navigation with natural multimodal assistance via retrospective curiosity-encouraging imitation learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 684–695, Hong Kong, China, November 2019. Association for Computational Linguistics.

[106] Khanh Nguyen, Debadeepta Dey, Chris Brockett, and Bill Dolan. Vision-based navigation with language-based assistance via imitation learning with indirect intervention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12527–12537, 2019.

[107] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[108] Li Niu, Jianfei Cai, Ashok Veeraraghavan, and Liqing Zhang. Zero-shot learning via category-specific visual-semantic mapping and label refinement. *IEEE Transactions on Image Processing*, 28(2):965–979, 2018.

[109] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013.

[110] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. Jointly modeling embedding and translation to bridge video and language. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4594–4602. IEEE, 2016.

[111] Devi Parikh and Kristen Grauman. Relative attributes. In *2011 International Conference on Computer Vision*, pages 503–510. IEEE, 2011.

[112] Massimiliano Patacchiola and Amos Storkey. Self-supervised relational reasoning for representation learning. *arXiv preprint arXiv:2006.05849*, 2020.

[113] Sudipta Paul, Shivkumar Chandrasekaran, BS Manjunath, and Amit K Roy-Chowdhury. Exploiting context for robustness to label noise in active learning. *arXiv preprint arXiv:2010.09066*, 2020.

[114] Sudipta Paul, Niluthpol Chowdhury Mithun, and Amit K Roy-Chowdhury. Text-based localization of moments in a video corpus. *IEEE Transactions on Image Processing*, 30:8886–8899, 2021.

[115] Sudipta Paul, Niluthpol Chowdhury Mithun, and Amit K Roy-Chowdhury. Text-based temporal localization of novel events. In *European Conference on Computer Vision*, pages 567–587. Springer, 2022.

[116] Sudipta Paul, Amit K Roy-Chowdhury, and Anoop Cherian. Avlen: Audio-visual-language embodied navigation in 3d environments. *arXiv preprint arXiv:2210.07940*, 2022.

[117] Sudipta Paul, Carlos Torres, Shivkumar Chandrasekaran, and Amit K Roy-Chowdhury. Complex pairwise activity analysis via instance level evolution reasoning. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2378–2382. IEEE, 2020.

[118] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[119] Mengshi Qi, Jie Qin, Yi Yang, Yunhong Wang, and Jiebo Luo. Semantics-aware spatial-temporal binaries for cross-modal video retrieval. *IEEE Transactions on Image Processing*, 30:2989–3004, 2021.

[120] 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.

[121] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[122] David Raposo, Adam Santoro, David Barrett, Razvan Pascanu, Timothy Lillicrap, and Peter Battaglia. Discovering objects and their relations from entangled scene representations. *arXiv preprint arXiv:1702.05068*, 2017.

[123] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013.

[124] Ryan Rivas, Sudipta Paul, Vagelis Hristidis, Evangelos E Papalexakis, and Amit K Roy-Chowdhury. Task-agnostic representation learning of multimodal twitter data for downstream applications. *Journal of Big Data*, 9(1):1–19, 2022.

[125] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212, 2015.

[126] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *International conference on machine learning*, pages 2152–2161. PMLR, 2015.

[127] Adam Santoro, David Raposo, David GT Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. *arXiv preprint arXiv:1706.01427*, 2017.

[128] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[129] Dian Shao, Yu Xiong, Yue Zhao, Qingqiu Huang, Yu Qiao, and Dahua Lin. Find and focus: Retrieve and localize video events with natural language queries. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 200–216, 2018.

[130] Sasha Sheng, Amanpreet Singh, Vedanuj Goswami, Jose Magana, Tristan Thrush, Wojciech Galuba, Devi Parikh, and Douwe Kiela. Human-adversarial visual question answering. *Advances in Neural Information Processing Systems*, 34:20346–20359, 2021.

[131] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer, 2016.

[132] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[133] Bharat Singh, Tim K Marks, Michael Jones, Oncel Tuzel, and Ming Shao. A multi-stream bi-directional recurrent neural network for fine-grained action detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1961–1970, 2016.

[134] Richard Socher, Milind Ganjoo, Hamsa Sridhar, Osbert Bastani, Christopher D Manning, and Andrew Y Ng. Zero-shot learning through cross-modal transfer. *arXiv preprint arXiv:1301.3666*, 2013.

[135] Mattia Soldan, Mengmeng Xu, Sisi Qu, Jesper Tegner, and Bernard Ghanem. Vlg-net: Video-language graph matching network for video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3224–3234, 2021.

[136] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1979–1988, 2019.

[137] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings*

of the IEEE conference on computer vision and pattern recognition, pages 1199–1208, 2018.

[138] Richard S Sutton, Doina Precup, and Satinder Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.

[139] Hao Tan, Licheng Yu, and Mohit Bansal. Learning to navigate unseen environments: Back translation with environmental dropout. *arXiv preprint arXiv:1904.04195*, 2019.

[140] Reuben Tan, Huijuan Xu, Kate Saenko, and Bryan A Plummer. Logan: Latent graph co-attention network for weakly-supervised video moment retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2083–2092, 2021.

[141] Haoyu Tang, Jihua Zhu, Zan Gao, Tao Zhuo, and Zhiyong Cheng. Attention feature matching for weakly-supervised video relocalization. In *Proceedings of the 2nd ACM International Conference on Multimedia in Asia*, pages 1–7, 2021.

[142] Haoyu Tang, Jihua Zhu, Lin Wang, Qinghai Zheng, and Tianwei Zhang. Multi-level query interaction for temporal language grounding. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[143] Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. Vision-and-dialog navigation. In *Conference on Robot Learning*, pages 394–406. PMLR, 2020.

[144] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *International Conference on Computer Vision (ICCV)*, pages 4489–4497. IEEE, 2015.

[145] Kazuya Ueki. Waseda meisei at trecvid 2017: Ad-hoc video search. 2017.

[146] 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.

[147] Hao Wang, Zheng-Jun Zha, Xuejin Chen, Zhiwei Xiong, and Jiebo Luo. Dual path interaction network for video moment localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4116–4124, 2020.

[148] Hao Wang, Zheng-Jun Zha, Liang Li, Dong Liu, and Jiebo Luo. Structured multi-level interaction network for video moment localization via language query. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7026–7035, 2021.

[149] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision (ECCV)*, pages 20–36. Springer, 2016.

[150] Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. End-to-end dense video captioning with parallel decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6847–6857, 2021.

[151] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6629–6638, 2019.

[152] Yuechen Wang, Jiajun Deng, Wengang Zhou, and Houqiang Li. Weakly supervised temporal adjacent network for language grounding. *IEEE Transactions on Multimedia*, 2021.

[153] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. Camp: Cross-modal adaptive message passing for text-image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5764–5773, 2019.

[154] Erik Wijmans, Samyak Datta, Oleksandr Maksymets, Abhishek Das, Georgia Gkioxari, Stefan Lee, Irfan Essa, Devi Parikh, and Dhruv Batra. Embodied question answering in photorealistic environments with point cloud perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6659–6668, 2019.

[155] Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen. Fine-grained action retrieval through multiple parts-of-speech embeddings. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 450–459, 2019.

[156] Aming Wu and Yahong Han. Multi-modal circulant fusion for video-to-language and backward. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 1029–1035, 2018.

[157] Gengshen Wu, Jungong Han, Yuchen Guo, Li Liu, Guiguang Ding, Qiang Ni, and Ling Shao. Unsupervised deep video hashing via balanced code for large-scale video retrieval. *IEEE Transactions on Image Processing*, 28(4):1993–2007, 2018.

[158] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4984, 2022.

[159] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 69–77, 2016.

[160] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4582–4591, 2017.

[161] Shaoning Xiao, Long Chen, Songyang Zhang, Wei Ji, Jian Shao, Lu Ye, and Jun Xiao. Boundary proposal network for two-stage natural language video localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2986–2994, 2021.

[162] Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. Multilevel language and vision integration for text-to-clip retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9062–9069, 2019.

[163] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016.

[164] Mengmeng Xu, Juan-Manuel Pérez-Rúa, Victor Escorcia, Brais Martinez, Xiatian Zhu, Li Zhang, Bernard Ghanem, and Tao Xiang. Boundary-sensitive pre-training for temporal localization in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7220–7230, 2021.

[165] Ran Xu, Caiming Xiong, Wei Chen, and Jason J Corso. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *AAAI*, volume 5, page 6, 2015.

[166] Xing Xu, Jingkuan Song, Huimin Lu, Yang Yang, Fumin Shen, and Zi Huang. Modal-adversarial semantic learning network for extendable cross-modal retrieval. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 46–54, 2018.

[167] Xun Xu, Timothy Hospedales, and Shaogang Gong. Semantic embedding space for zero-shot action recognition. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 63–67. IEEE, 2015.

[168] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1686–1697, 2021.

[169] Wenfei Yang, Tianzhu Zhang, Yongdong Zhang, and Feng Wu. Local correspondence network for weakly supervised temporal sentence grounding. *IEEE Transactions on Image Processing*, 30:3252–3262, 2021.

[170] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18155–18165, 2022.

[171] Hanhua Ye, Guorong Li, Yuankai Qi, Shuhui Wang, Qingming Huang, and Ming-Hsuan Yang. Hierarchical modular network for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17939–17948, 2022.

[172] Mang Ye and Jianbing Shen. Probabilistic structural latent representation for unsupervised embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5457–5466, 2020.

[173] Mang Ye, Jianbing Shen, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. Augmentation invariant and instance spreading feature for softmax embedding. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[174] Licheng Yu, Xinlei Chen, Georgia Gkioxari, Mohit Bansal, Tamara L Berg, and Dhruv Batra. Multi-target embodied question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6309–6318, 2019.

[175] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 471–487, 2018.

[176] Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. End-to-end concept word detection for video captioning, retrieval, and question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3165–3173, 2017.

[177] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. In *Advances in Neural Information Processing Systems*, pages 534–544, 2019.

[178] Yitian Yuan, Tao Mei, and Wenwu Zhu. To find where you talk: Temporal sentence localization in video with attention based location regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9159–9166, 2019.

[179] Vinicius Zambaldi, David Raposo, Adam Santoro, Victor Bapst, Yujia Li, Igor Babuschkin, Karl Tuyls, David Reichert, Timothy Lillicrap, Edward Lockhart, et al. Deep reinforcement learning with relational inductive biases. In *International Conference on Learning Representations*, 2018.

[180] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10287–10296, 2020.

[181] Bowen Zhang, Hexiang Hu, and Fei Sha. Cross-modal and hierarchical modeling of video and text. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 374–390, 2018.

[182] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1247–1257, 2019.

[183] Hao Zhang, Aixin Sun, Wei Jing, Liangli Zhen, Joey Tianyi Zhou, and Rick Siow Mong Goh. Natural language video localization: A revisit in span-based question answering framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[184] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. *arXiv preprint arXiv:2004.13931*, 2020.

[185] Haofeng Zhang, Yang Long, Yu Guan, and Ling Shao. Triple verification network for generalized zero-shot learning. *IEEE Transactions on Image Processing*, 28(1):506–517, 2018.

[186] Lingling Zhang, Xiaojun Chang, Jun Liu, Minnan Luo, Sen Wang, Zongyuan Ge, and Alexander Hauptmann. Zstad: Zero-shot temporal activity detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[187] Mingxing Zhang, Yang Yang, Xinghan Chen, Yanli Ji, Xing Xu, Jingjing Li, and Heng Tao Shen. Multi-stage aggregated transformer network for temporal language localization in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12669–12678, 2021.

[188] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. *arXiv preprint arXiv:1912.03590*, 2019.

[189] Songyang Zhang, Jinsong Su, and Jiebo Luo. Exploiting temporal relationships in video moment localization with natural language. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1230–1238, 2019.

[190] Zhu Zhang, Zhijie Lin, Zhou Zhao, and Zhenxin Xiao. Cross-modal interaction networks for query-based moment retrieval in videos. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 655–664, 2019.

[191] Wentian Zhao, Xinxiao Wu, and Jiebo Luo. Multi-modal dependency tree for video captioning. *Advances in Neural Information Processing Systems*, 34:6634–6645, 2021.

[192] Yang Zhao, Zhou Zhao, Zhu Zhang, and Zhijie Lin. Cascaded prediction network via segment tree for temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4197–4206, 2021.

[193] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(2):1–23, 2020.

[194] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 803–818, 2018.

[195] Hao Zhou, Chongyang Zhang, Yan Luo, Yanjun Chen, and Chuanping Hu. Embracing uncertainty: Decoupling and de-bias for robust temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8445–8454, 2021.

[196] Wang Zhu, Hexiang Hu, Jiacheng Chen, Zhiwei Deng, Vihan Jain, Eugene Ie, and Fei Sha. Babywalk: Going farther in vision-and-language navigation by taking baby steps. *arXiv preprint arXiv:2005.04625*, 2020.

[197] Yi Zhu, Yang Long, Yu Guan, Shawn Newsam, and Ling Shao. Towards universal representation for unseen action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9436–9445, 2018.

[198] Yi Zhu, Yue Weng, Fengda Zhu, Xiaodan Liang, Qixiang Ye, Yutong Lu, and Jianbin Jiao. Self-motivated communication agent for real-world vision-dialog navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1594–1603, 2021.