

UC Davis

UC Davis Electronic Theses and Dissertations

Title

Representational Learning for Grounded Reasoning Across Modalities

Permalink

<https://escholarship.org/uc/item/6bd2b2n5>

ISBN

9798189233449

Author

Li, Bangzheng

Publication Date

2026-06-25

Peer reviewed|Thesis/dissertation

Representational Learning for Grounded Reasoning Across Modalities

By

BANGZHENG LI
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Computer Science

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Muhao Chen, Chair

Zhe Zhao

Junshan Zhang

Committee in Charge

2026

Copyright 2026
by
BANGZHENG LI

Abstract

Representational Learning for Grounded Reasoning Across Modalities

by

BANGZHENG LI

Doctor of Philosophy in Computer Science

University of California, DAVIS

Associate Professor Muhao Chen, Chair

The quest for artificial general intelligence is fundamentally a quest for robust machine reasoning—the ability of a model to transcend simple pattern matching and perform multi-step logical inference. This dissertation argues that the bottleneck of current machine reasoning lies not just in model scale, but in the nature of the underlying representations the models utilize. We explore the evolution of these representations from static linguistic indices to dynamic, latent multimodal spaces, and finally to learned internal attention policies.

The first part of this work addresses the limitations of categorical representations in Natural Language Processing. We introduce LITE, a framework that reformulates entity typing as a Natural Language Inference (NLI) problem. By mapping diverse entity types into a semantic hypothesis space rather than fixed indices, we demonstrate that models can achieve superior generalization to unseen concepts, establishing that reasoning is more effective when grounded in rich semantic representations.

In the second part, we diagnose the fragility of these reasoning chains. Through the EUREQA benchmark, we uncover a systemic vulnerability in Large Language Models (LLMs): the tendency to follow "deceptive semantic shortcuts." We prove that models often rely on

superficial entity associations rather than faithful multi-hop logic. This finding motivates the need for reasoning mechanisms that are more deeply grounded in the input data rather than biased linguistic priors.

The third part of the dissertation extends these insights into the multimodal domain. We first propose Semantic-Clipping (SEMCLIP), a framework that uses textual semantics to dynamically select relevant visual representations, enhancing efficiency and fine-grained reasoning without retraining. We then advance this concept with Latent Visual Reasoning (LVR), a novel paradigm where reasoning occurs directly within the visual embedding space. By training Multimodal LLMs to generate latent states that reconstruct key visual features, we move beyond surface-level text rationales, enabling models to "think" in visual tokens.

Finally, we propose Reinforced Attention Learning (RAL) to optimize the fundamental mechanics of these representations. By applying policy-gradient reinforcement learning directly to internal attention distributions, we shift the optimization objective from what the model generates to where it attends. This approach, coupled with On-Policy Attention Distillation, provides a principled method for cross-modal alignment that outperforms standard token-based distillation.

Collectively, the contributions of this dissertation demonstrate that by advancing how models represent, select, and attend to information, we can move toward machine reasoning systems that are more efficient, less prone to hallucination, and capable of sophisticated latent inference across modalities.

Dedicated to those whose labor creates the world; may artificial intelligence belong to the many, not be wielded by the few.

Practice is the sole criterion for testing truth.

Contents

Contents	v
List of Figures	vii
List of Tables	ix
1 Introduction	1
1.1 From Indices to Semantics: Rethinking Entity Representations	2
1.2 Diagnosing Reasoning Fragility: Deceptive Semantic Shortcuts	4
1.3 Toward Grounded Multimodal Reasoning	5
1.4 Optimizing the Mechanics of Representation: Reinforced Attention Learning	7
1.5 Thesis Statement and Contributions	8
1.6 Dissertation Organization	10
2 Background	11
2.1 Reasoning as Representation and Control in Foundation Models	11
2.2 Transformers, Attention, and the Interface to Multimodal Reasoning	12
2.3 Entity Typing and Natural Language Inference	13
2.4 Multi-Hop Reasoning and Shortcut Vulnerabilities	15
2.5 Vision-Language Models and Cross-Modal Grounding	17
2.6 Latent Reasoning in Textual and Visual Domains	18
2.7 Reinforcement Learning for Post-Training and Knowledge Distillation	20
2.8 Summary	22
3 LITE: Ultra-Fine Entity Typing with Indirect Supervision from Natural Language Inference	24
3.1 Introduction	24
3.2 Method	25
3.3 Experiments	30
3.4 Conclusion and Transition	38
4 EUREQA: Deceptive Semantic Shortcuts on Reasoning Chains	39
4.1 Introduction	39

4.2	The EUREQA Benchmark	43
4.3	Experimental Setup	47
4.4	Results	49
4.5	Analysis and Discussion	50
4.6	Conclusion and Transition	55
5	SEMCLIP: Semantic-Guided Visual Selection for Vision-Language Models	57
5.1	Introduction	57
5.2	Method	59
5.3	Experiments	64
5.4	Conclusion and Transition	72
6	Latent Visual Reasoning	73
6.1	Introduction	73
6.2	Method	77
6.3	Experiments	82
6.4	Conclusion and Transition	87
7	Reinforced Attention Learning	89
7.1	Introduction	89
7.2	Reinforced Attention Learning	92
7.3	Experiments	96
7.4	Conclusion and Transition	103
8	Conclusion and Future Work	105
8.1	Summary of Contributions	105
8.2	Future Work	107
8.3	Closing Remarks	113
	Bibliography	115

List of Figures

3.1	Entity typing by LITE with indirect supervision from NLI.	27
3.2	Performance comparison of LITE and MLMET on the filtered UFET benchmark for zero-shot and few-shot typing. Zero-shot labels correspond to the 40% of test types unseen in training.	35
4.1	LLMs make errors when correct surface-level semantic cues—entities—are recursively replaced with descriptions, and the errors are likely related to token similarity. GPT-3.5-turbo is used for this example.	41
4.2	Even given all the required information needed for the question (selected information shown), GPT-4 still makes mistakes starting at early layers (highlighted in grey). Partial output is shown. Few-shot prompts are provided.	42
4.3	The data generation process of EUREQA.	45
4.4	Categorical distribution of seed entities in EUREQA questions.	48
4.5	Correlation between GPT-4 performance on EUREQA hard set and entity similarities.	52
4.6	Distribution of entity similarity scores in EUREQA.	52
4.7	GPT-4 performance with different entity similarity scores between the easy and hard sets.	53
5.1	Paying attention to task-relevant regions within a scene is an intuitive approach to answering visual questions. Our objective is to identify the optimal task-relevance measurement ψ that selects the most pertinent sub-region of an image, enhancing visual understanding.	60
5.2	SEMCLIP is a plug-and-play method that enhances a VLM through semantic-guided visual selection. Task-relevant sub-areas of the image are encoded and appended to the visual tokens of the overview image (colored in blue and purple, respectively). These visual tokens, together with text embeddings of the question (colored in orange), are processed by the backbone VLM to generate a response.	62

6.1	Conceptual illustration of LATENT VISUAL REASONING (LVR). We compare LVR with two paradigms: “Think about images,” which performs multimodal reasoning entirely in text space, and “Think with images,” which leverages external visual tools to highlight regions of interest (ROIs). In contrast, LVR leverages the LLM’s latent space to reconstruct the semantics of ROIs, enabling seamless cross-modal reasoning.	76
6.2	Training and inference pipeline of LVR. Images are encoded into tokens by a visual encoder and mapped into a joint semantic space with text embeddings. During the SFT stage, bounding boxes identify query-relevant visual tokens, which supervise the last hidden states in the LVR process via MSE loss. The LVR process is wrapped with special tokens indicating reasoning mode. Once all query-relevant visual tokens are consumed, the model exits LVR and resumes standard text generation with cross-entropy loss. During RL training, the model self-evolves the LVR process learned in SFT using GRPO _{latent} . At inference, the model triggers LVR upon generating the special token, propagates hidden states to reconstruct visual semantics, and resumes text generation when a stopping criterion is met.	78
7.1	Reinforced Attention Learning formulates internal attention distributions as a policy. Unlike traditional methods that optimize next-token probabilities (“what to generate”), our approach prioritizes the selective allocation of information focus (“where to look”). Optimizing for the advantage, the model explores a high-rewarding attention policy that effectively isolates salient information from dense contexts.	91
7.2	Sample data of the SFT and RL training stages. The SFT stage adapts the model to a “think-and-answer” paradigm, while the RL stage employs a reward function to verify the format and correctness of rollout responses.	96
7.3	RAL improves GRPO along increasing video frames or image resolution.	101

List of Tables

3.1	Type description instances of three templates. Entity mentions are boldfaced and underlined while label words are only boldfaced.	28
3.2	Results on the ultra-fine entity typing task. LITE uses the taxonomic statement template. “w/o label dependency” is applied to the NLI+L setting. The F1 result by LITE _{NLI+L} is statistically significant ($p < 0.01$ in t -test) compared to MLMET.	32
3.3	Results on fine-grained entity typing. All LITE results are statistically significant ($p < 0.05$ in t -test) compared to the best baseline (MLMET on OntoNotes, SEPREM on FINGER).	33
3.4	Examples from two sources of distant supervision data. Correct types are boldfaced; <i>incorrect ones</i> are in grey.	34
3.5	Behavior of different type description templates under three training settings.	34
3.6	Case study of labels on which LITE improves MLMET or vice versa. Correct predictions are in blue; * marks the representative label for each discussed pattern.	36
3.7	Performance comparison of LITE and MLMET on named entity, pronoun, and nominal entity mentions.	37
4.1	Statistics of EUREQA.	47
4.2	Accuracy of ChatGPT, Gemini-Pro, and GPT-4 across different reasoning depths d (number of layers) and difficulty levels. Two prompting strategies are evaluated: <i>direct</i> (zero-shot) and <i>icl</i> (two-shot with examples).	50
4.3	Accuracy of Llama-2 models across different reasoning depths d on hard questions. The 7b and 13b versions are chat variants; the 70b version is the instruct variant.	54
5.1	SEMCLIP achieves the highest performance improvements across most tasks, surpassing all baselines in average accuracy. While baseline methods require model-specific fine-tuning, SEMCLIP operates as a plug-and-play approach with minimal training on the top-performing relevance scorer, ψ_{clip}	67
5.2	Theoretical upper bound experiments reveal that modern VLMs have significant potential for improvement in visual understanding when provided with appropriate sub-image guidance. Ablation studies confirm that SEMCLIP’s effectiveness is not merely due to extra visual tokens, and that sub-images alone are insufficient for comprehensive understanding.	69

5.3	SEMCLIP delivers performance on par with the best baseline while requiring significantly less training and 60% fewer visual tokens, achieving the optimal balance between training cost, inference efficiency, and accuracy. SEMCLIP outperforms a VLM with a substantially larger language model, highlighting the superiority of task-guided visual selection over simply increasing language model size.	71
6.1	Experimental results on vision-centric tasks. LVR outperforms both “Think about Images” (PAPO, Vision-R1) and “Think with Images” (PixelReasoner) paradigms, highlighting the effectiveness and scalability of latent visual reasoning.	86
6.2	RL results with 3B models. GRPO _{latent} further boosts performance, demonstrating the effectiveness of adapting RL for latent reasoning and enabling self-evolution.	86
6.3	Ablation studies on the 7B model. The standard LVR approach performs best, indicating the LLM natively aligns visual and textual semantics without an extra head. The unstable Latent End Token highlights a need for future work on variable-length reasoning.	87
7.1	Hyperparameters for RAL experiments.	98
7.2	Performance comparison across image VQA benchmarks.	100
7.3	Performance comparison across long-video benchmarks. RAL and Attention Distillation consistently enhance multimodal understanding. RAL-zero denotes the variant where the explicit thinking process is removed, isolating the impact of the attention-based policy gradient.	101

Acknowledgments

I am deeply grateful to my advisor, Prof. Muhao Chen, for his unwavering support and mentorship throughout my doctoral journey and for fostering in me a genuine passion for academic inquiry. He generously afforded me the autonomy to pursue research directions aligned with my own curiosities, and through his example he has embodied the qualities of both an exceptional researcher and a dedicated teacher. Together, from the University of Southern California to UC Davis, we have navigated the ever-evolving landscape of artificial intelligence.

My sincere thanks also go to Professors Zhe Zhao and Junshan Zhang for their continued guidance during my doctoral studies at Davis. Their expertise and thoughtful suggestions have been instrumental to my progress.

I am equally grateful to my colleagues at USC for their support during the formative years of my doctoral work. In particular, I would like to thank Fei Wang, Nancy Xu, Ehsan Qasemi, Tenghao Huang, and Wenxuan Zhou, whose contributions were vital to the research presented in this thesis. Many of the core ideas explored here were forged during our collaborative brainstorming sessions, and this body of work is significantly richer for their intellectual partnership.

My heartfelt thanks go to my close research collaborators over the years, with special gratitude to Xuanyu Ben Zhou, Kyle Robert Chickering, and Xiaofei Wen. Ben’s influence was pivotal in directing my research toward embedding reasoning behaviors into AI systems. My work with Kyle has deeply strengthened my conviction in the future of AI—a unified model capable of accommodating all modalities. Xiaofei was my very first intern; our discussions taught me the true value of effective collaboration and of scaling research efforts.

I owe a tremendous debt of gratitude to my colleagues at UPenn, especially those in the CogComp Lab. I am deeply thankful to Professor Dan Roth, whose guidance inspired much of my research and helped shape my vision of AGI. Dan’s expertise in learning theory and his instinct for asking the right questions profoundly influenced my academic trajectory. I also want to thank Wenpeng Yin, Xuanyu Ben Zhou, and Xiaodong Yu for their camaraderie. My three years in Philadelphia were unforgettable—the city and the people I met there have given me memories I will treasure forever.

A special thanks goes to Heling Zhang, a lifelong friend from Beijing who shared my journey from high school deskmate to UIUC roommate. His support was indispensable during my transition to life in America; I truly could not have navigated the challenges of that first year without his friendship.

Beyond academia, I am grateful for the mentorship and support I received at AMD, particularly from Ximeng Sun, Xiaodong Yu, and Zicheng Liu. Ximeng's patient guidance was instrumental in helping me adapt to the industry environment, and I deeply appreciate the technical support and encouragement that Xiaodong and Zicheng provided throughout my internship.

I would also like to thank my colleagues at Google, including Jianmo Ni, Chen Qu, Ian Miao, Liu Yang, Nancy Xu, and Derek Zhiyuan Cheng. Jianmo has been far more than a technical mentor; he has offered invaluable guidance for both my career and my life. I am deeply grateful to Chen and Ian, who went above and beyond to ensure my internship ran smoothly. Nancy spared no effort in supporting my job search. Finally, my conversations with Derek were truly enlightening and helped me crystallize my long-term career goals.

I am thankful to the friends who have brought joy to my life beyond the lab—Yu Tang, Yujie Lu, Weijia Shi, Yining Lu, Yuxuan Jing, Chen Shen, Pan Lu, and Siyao Chen—for the laughter, companionship, and perspective they have offered over the years.

Lastly, I am eternally grateful to my family for their boundless love. To my parents, Chunqing Li and Huiqiu Gong: your courage in supporting my move across the world to begin college was a gift that changed my life. To my grandmothers, Shuying Zhang and Rongfen Wang: thank you for your unconditional love, despite the many miles between us. And to my cats, Lucky Fu and Targa Li: thank you for the quiet comfort you provided when I needed it most.

To Xingyu Fu, my love: thank you for being the joy and light of my life. From our days in Champaign and Philadelphia to our time in Los Angeles and Davis, you have made this journey beautiful. I cherish every step we have taken together and look forward to all the adventures that lie ahead.

Chapter 1

Introduction

The pursuit of artificial general intelligence is, at its core, a pursuit of robust machine reasoning—the capacity for a computational system to move beyond surface-level pattern matching and perform multi-step logical inference over complex, heterogeneous inputs. While the past decade has witnessed extraordinary progress in the scale and capability of neural language models, a growing body of evidence suggests that scale alone is insufficient to guarantee faithful reasoning. Models that appear to reason correctly on standard benchmarks often rely on spurious statistical correlations, fail to generalize to novel compositions of known facts, and hallucinate plausible but incorrect conclusions when the distribution of test inputs shifts even modestly from that of training data. These failures point to a deeper issue: the bottleneck in machine reasoning is not merely computational, but representational.

This dissertation argues that advancing machine reasoning requires a fundamental rethinking of how models represent, select, and attend to information. Across five interconnected studies, we trace an arc from the limitations of categorical representations in natural language processing, through the diagnosis of reasoning fragility in large language models, to the development of novel multimodal reasoning paradigms that operate in learned latent spaces. The unifying thesis is that reasoning quality is tightly coupled with representation quality, and that by engineering richer, more dynamic, and more semantically grounded representations, we can build systems that are more efficient, more robust, and less prone to hallucination.

The motivation for this research agenda emerges from a confluence of trends. First, the rapid adoption of large language models (LLMs) in high-stakes applications—from medical

question answering to legal document analysis—has made the reliability of reasoning chains a practical concern, not merely an academic one. Second, the extension of language models into the multimodal domain, where they must jointly process text and images, has exposed new failure modes that purely textual diagnostics cannot capture. Third, the community’s increasing interest in test-time scaling and chain-of-thought prompting has raised important questions about whether generating more tokens at inference time truly constitutes deeper reasoning, or whether it merely amplifies existing biases in the model’s internal representations.

In what follows, we provide the context necessary to understand each contribution of this dissertation. We begin by reviewing the evolution of entity representations in natural language understanding, which motivates our first contribution. We then discuss the problem of reasoning faithfulness in LLMs, which motivates our diagnostic benchmark. Next, we turn to the multimodal setting, reviewing the challenges of visual representation and reasoning in vision-language models. Finally, we describe the role of attention mechanisms as a representational bottleneck, motivating our work on reinforcement learning over attention distributions. A comprehensive review of related work and the technical preliminaries underlying each contribution is deferred to Chapter 2.

1.1 From Indices to Semantics: Rethinking Entity Representations

Entity typing—the task of inferring the semantic types of entity mentions in text—is a foundational problem in natural language understanding. The resulting type information grounds linguistic expressions to real-world concepts and provides valuable prior knowledge for downstream tasks such as entity linking, question answering, and information extraction. For much of its history, the entity typing community has formulated the problem as multi-way classification, in which each candidate type is assigned a fixed numerical index and a model is trained to select the correct indices given an input sentence and entity mention.

This classification-based paradigm works reasonably well when the type inventory is small and closed. However, real-world entity types are extremely diverse, and the community has increasingly recognized the need for more fine-grained modeling. The Ultra-Fine Entity

Typing (UFET) benchmark, for instance, defines over ten thousand free-form type candidates, ranging from coarse categories like “person” and “location” to highly specific types like “political activist” or “renaissance painter.” The dramatic expansion of the type space exposes two critical weaknesses of the classification paradigm. First, because types are reduced to indices, the model has no mechanism for understanding the semantics of the types themselves. Without access to what a type means, the model struggles to make correct predictions when annotated data for a given type is scarce or nonexistent. Second, classification-based systems are inherently closed-world: they cannot recognize entity types that were not present in the training label space unless the entire classifier is retrained.

These limitations motivated the first contribution of this dissertation, LITE (Language Inference based Typing of Entities). Rather than treating entity typing as classification over indices, LITE reformulates it as a natural language inference (NLI) problem. Each entity-mentioning sentence is treated as a premise, and each candidate type is transformed via templates into a natural-language hypothesis. An NLI model then scores the entailment relationship between premise and hypothesis, and a learning-to-rank objective optimizes these scores so that hypotheses corresponding to correct types are ranked above those for incorrect types. This reformulation preserves the semantics of both the input context and the candidate types, enabling the model to generalize to previously unseen types through compositional understanding of the hypothesis text. The approach yields state-of-the-art results on both ultra-fine and traditional fine-grained entity typing benchmarks and demonstrates strong transferability across datasets with different type inventories.

More broadly, the LITE framework establishes a principle that recurs throughout this dissertation: reasoning is more effective when grounded in rich semantic representations rather than arbitrary categorical indices. This insight carries implications well beyond entity typing. It suggests that whenever a model must make decisions over a large or open-ended space of possibilities, reformulating the problem so that each candidate is represented in a semantically meaningful way—rather than as a mere label index—can yield substantial gains in generalization and robustness.

1.2 Diagnosing Reasoning Fragility: Deceptive Semantic Shortcuts

The success of large language models on multi-hop reasoning benchmarks has been widely celebrated, yet it remains an open question whether these models genuinely perform faithful multi-step inference or instead exploit superficial statistical shortcuts to arrive at plausible answers. This question has profound implications: if LLMs are primarily relying on co-occurrence patterns absorbed during pre-training rather than executing the logical steps implied by a chain-of-thought prompt, then their apparent reasoning ability is far more brittle than benchmark scores suggest.

Consider a multi-hop question that asks a model to identify an individual who is simultaneously a producer of one film and a cast member of another. A truly reasoning system would decompose this into sub-queries, retrieve the relevant sets of individuals, and compute their intersection. In practice, however, a language model may bypass this process entirely if the correct answer—say, a famous actor—co-occurs frequently with both films in the pre-training corpus. The model can emit the correct answer through a shortcut of semantic association, never performing the intended multi-hop reasoning at all. Such shortcuts are effective on common test cases that align with pre-training distributions, but they fail catastrophically when the associations are removed or made misleading.

To rigorously investigate this phenomenon, the second contribution of this dissertation introduces EUREQA (Extended Understanding and REasoning QA), a multi-hop question-answering benchmark specifically designed to diminish semantic associations and introduce deceptive semantic cues. The dataset construction method replaces entities that are directly relevant to the correct answer with placeholder variables, supplemented by descriptive sentences that the model must process through genuine multi-hop reasoning to resolve. The reasoning depth can be systematically increased by recursively replacing entities in these descriptions with further sub-queries, creating chains of arbitrary length while maintaining a simple, consistent reasoning path.

Evaluation on EUREQA reveals a striking gap between human and model performance. State-of-the-art LLMs, including GPT-4, achieve only 62–64% accuracy on the benchmark,

while humans solve the same problems near-perfectly with minimal effort. Furthermore, model performance correlates strongly with the semantic similarity between the gold answer and other entities mentioned in the question, providing direct evidence that the models are exploiting surface-level entity associations rather than following faithful reasoning chains. These findings establish that the problem of reasoning in LLMs is not merely one of scale or prompting strategy—it is a representational problem, rooted in the model’s tendency to rely on biased linguistic priors rather than grounded logical inference.

The diagnostic lessons of EUREQA motivate the subsequent chapters of this dissertation. If reasoning fails because models latch onto superficial textual cues rather than engaging deeply with the input data, then one promising direction is to develop reasoning mechanisms that are more deeply grounded—either in richer textual representations or, as we explore next, in visual representations that complement and constrain the textual reasoning process.

1.3 Toward Grounded Multimodal Reasoning

Efficient Visual Selection with Semantic Guidance

Vision-language models (VLMs) have achieved remarkable results on tasks that require joint understanding of images and text. The dominant architectural paradigm processes an input image through a visual encoder, which produces a fixed set of patch-level tokens, and then feeds these tokens as a prefix to a large language model. While effective, this approach faces a fundamental tension: preserving fine-grained visual detail requires higher-resolution encoding, which produces more visual tokens, which in turn strains the context limits of the language model, increases computational cost, and can paradoxically degrade reasoning performance by flooding the model with irrelevant visual information.

The human visual system offers an instructive contrast. When searching for a specific object in a cluttered scene, a person does not process every visual region with equal attention; instead, they leverage task-relevant cues to focus on the most pertinent area. Motivated by this observation, the third contribution of this dissertation introduces Semantic-Clipping (SEMCLIP), a plug-and-play framework for semantic-guided visual selection. During inference, SEMCLIP segments an input image into multiple candidate sub-images and scores the

relevance of each sub-image to the textual query. The most relevant sub-image is then appended to the original image before being passed to the VLM, providing enhanced visual detail precisely where it is needed without incurring the cost of encoding the entire image at high resolution.

SEMCLIP explores multiple strategies for computing relevance scores, including approaches based on features extracted from VLM modules, pre-trained vision-language alignment models, and efficiently fine-tuned scoring models. Experimental results demonstrate that SEMCLIP significantly improves visual question-answering performance across several widely used VLMs while using fewer visual tokens than baseline high-resolution methods. Critically, SEMCLIP requires no model retraining, making it immediately applicable to existing deployed systems. The framework also reveals a substantial gap between the current visual understanding capabilities of VLMs and their theoretical upper bound under oracle sub-image selection, suggesting that better visual selection is a high-leverage avenue for future improvement.

Latent Visual Reasoning in Embedding Space

While SEMCLIP addresses the visual selection problem at the input level, a more fundamental question remains: can multimodal models reason directly in the visual embedding space, rather than translating all reasoning into text? Existing multimodal reasoning paradigms can be broadly categorized into two families. The first, often termed “thinking about images,” performs chain-of-thought reasoning entirely in text space, generating lengthy textual descriptions of visual content as an intermediate step. The second, “thinking with images,” uses external tools to manipulate images during reasoning—for example, by drawing auxiliary lines, zooming into regions, or highlighting sub-areas—and re-encoding the results. Both paradigms ultimately funnel visual information through textual channels, and both suffer from characteristic limitations: text-only reasoning can be overwhelmed by the very tokens it generates, causing the textual context to overshadow the original visual inputs, while tool-based approaches are constrained by the predefined operations of the external tools and may bypass newly injected visual information due to training biases.

The fourth contribution of this dissertation, Latent Visual Reasoning (LVR), proposes

an alternative paradigm that addresses these limitations. Observing that continuous visual tokens and discrete text tokens are already embedded in a shared latent semantic space within multimodal LLMs, LVR enables the language model to perform hybrid reasoning that alternates between standard text generation and a novel latent visual reasoning phase. In the LVR phase, the model’s last hidden states are trained to approximate question-relevant visual tokens—that is, the model learns to “think” in visual representations, reconstructing the semantics of regions of interest directly in the embedding space rather than describing them in text.

LVR is trained through a two-stage pipeline. The first stage uses supervised fine-tuning with a reconstruction loss alongside standard next-token prediction, teaching the model to produce latent states that faithfully encode relevant visual features. The second stage applies reinforcement learning, adapting the GRPO algorithm to replay latent reasoning steps during policy gradient computation, which allows the model to self-improve its latent reasoning process using rewards derived from text-generation quality. Experimental results show that LVR achieves substantial improvements over state-of-the-art multimodal LLMs, particularly on perception-intensive tasks and benchmarks that demand fine-grained visual detail understanding. By moving reasoning into the embedding space, LVR represents a significant step toward models that integrate visual and textual information at the level of internal representation rather than surface-level translation.

1.4 Optimizing the Mechanics of Representation: Reinforced Attention Learning

The preceding contributions advance how models represent inputs and structure reasoning. The final contribution of this dissertation turns to a still deeper question: can we optimize the fundamental mechanism by which models allocate attention across representations? In the Transformer architecture, the attention mechanism governs which tokens in the input context influence the generation of each output token. For multimodal models, this mechanism is the primary gateway through which visual information reaches the language model’s reasoning process. Standard post-training methods, including reinforcement learning from human

feedback (RLHF) and group relative policy optimization (GRPO), optimize the model’s next-token distribution—that is, they optimize *what* the model generates. However, they do so only indirectly with respect to *where* the model attends, leaving the internal information-routing process as an implicit byproduct of token-level optimization.

This indirection is particularly costly in the multimodal setting. Accurately answering fine-grained visual questions requires the model to attend precisely to task-relevant visual tokens among a potentially large set of input patches. Standard RL post-training can encourage correct final answers, but it provides no direct gradient signal to the attention layers regarding which visual regions should be prioritized. The result is that attention patterns are only loosely aligned with the visual grounding demands of the task.

To address this gap, the fifth and final contribution of this dissertation proposes Reinforced Attention Learning (RAL), an algorithm that reformulates the post-training policy to operate directly on attention distributions rather than token distributions. In RAL, the attention pattern itself is treated as the policy: when a generated response receives a high reward, the algorithm reinforces the underlying attention distribution by minimizing its divergence from a reference policy, and when a response receives a low reward, the algorithm penalizes the corresponding attention pattern. By shifting the optimization target from token likelihood to attention allocation, RAL provides a more direct training signal for cross-modal alignment.

The RAL framework naturally extends to a distillation setting. Traditional knowledge distillation focuses on aligning the token-level probability distributions of a student model with those of a teacher. RAL introduces on-policy attention distillation, which transfers knowledge through both token-level and attention-level alignment. Experimental results demonstrate that RAL consistently outperforms GRPO across diverse video and image question-answering benchmarks, with particularly strong gains on perception-intensive tasks, and that the inclusion of attention distillation provides significant additional improvements for student models inheriting fine-grained perceptual behaviors from a teacher.

1.5 Thesis Statement and Contributions

This dissertation advances the following thesis: *the quality of machine reasoning is fundamentally constrained by the quality of the underlying representations, and systematic*

improvements to how models represent, select, and attend to information yield reasoning systems that are more efficient, more robust, and less prone to hallucination across both unimodal and multimodal settings.

In support of this thesis, the dissertation makes five principal contributions:

1. **LITE: Language Inference based Typing of Entities.** We introduce a framework that reformulates entity typing as a natural language inference problem, demonstrating that grounding type representations in semantics rather than indices enables superior generalization, state-of-the-art performance, and zero-shot transfer across type inventories (Chapter 3).
2. **EUREQA: Extended Understanding and Reasoning QA.** We propose a diagnostic benchmark that exposes the tendency of large language models to follow deceptive semantic shortcuts rather than faithful multi-hop reasoning chains, establishing that the reasoning fragility of LLMs is a representational problem rooted in biased linguistic priors (Chapter 4).
3. **SEMCLIP: Semantic-Guided Visual Selection.** We develop a plug-and-play framework for semantically guided sub-image selection that enhances fine-grained visual understanding in vision-language models without retraining, demonstrating that dynamically selecting task-relevant visual representations improves both performance and efficiency (Chapter 5).
4. **LVR: Latent Visual Reasoning.** We introduce a novel multimodal reasoning paradigm in which the language model reasons directly in the visual embedding space by generating latent states that reconstruct question-relevant visual features, achieving substantial improvements on perception-intensive benchmarks (Chapter 6).
5. **RAL: Reinforced Attention Learning.** We propose a post-training algorithm that optimizes attention distributions via policy-gradient reinforcement learning, shifting the training signal from what the model generates to where it attends, and extend this framework to on-policy attention distillation for effective cross-modal knowledge transfer (Chapter 7).

1.6 Dissertation Organization

The remainder of this dissertation is organized as follows. Chapter 2 provides the background and technical preliminaries that underlie the contributions of this work, including foundations in natural language inference, multi-hop reasoning, vision-language modeling, and reinforcement learning for language models. Chapter 3 presents LITE and the reformulation of entity typing as natural language inference. Chapter 4 introduces the EUREQA benchmark and the analysis of deceptive semantic shortcuts in LLM reasoning. Chapter 5 describes the SEMCLIP framework for semantic-guided visual selection. Chapter 6 develops the Latent Visual Reasoning paradigm for multimodal reasoning in embedding space. Chapter 7 presents Reinforced Attention Learning and on-policy attention distillation. Finally, Chapter 8 concludes the dissertation with a synthesis of findings, a discussion of limitations, and directions for future research.

Chapter 2

Background

2.1 Reasoning as Representation and Control in Foundation Models

This dissertation studies machine reasoning through a representation-centric lens: reasoning performance depends not only on model scale and training data, but on (i) *what representations* the model reasons over (label semantics, latent multimodal states), (ii) *how evidence is selected* (query-conditioned visual selection), and (iii) *how internal computation is allocated* (attention policies). These three dimensions—representation, selection, and control—cut across several active research areas: entity typing and structured prediction, natural language inference, reasoning evaluation and multi-hop question answering, vision-language modeling, multimodal reasoning paradigms, and post-training via reinforcement learning and distillation. This chapter reviews each of these areas in turn and clarifies how they connect to the contributions of this dissertation.

The rapid scaling of pre-trained language models over the past several years has produced systems capable of fluent text generation, in-context learning, and increasingly complex forms of structured reasoning. Yet scaling alone has not resolved fundamental challenges: models still hallucinate, exploit distributional shortcuts, and struggle with fine-grained cross-modal grounding. A central argument of this dissertation is that many of these failures are best understood as representational failures—arising not from insufficient model capacity, but

from the mismatch between the representations a model operates over and the structure of the reasoning task it faces. The background presented in this chapter is organized to make this argument concrete, tracing the representational bottleneck from discrete label spaces in entity typing (Section 2.3), through the shortcut vulnerabilities exposed in multi-hop reasoning (Section 2.4), to the modality gap in vision-language models (Section 2.5), the emergence of latent-space reasoning as an alternative to text-mediated inference (Section 2.6), and finally the optimization of attention as the mechanism that governs information flow across all of these settings (Section 2.7).

2.2 Transformers, Attention, and the Interface to Multimodal Reasoning

Transformers (Vaswani et al., 2017) provide the dominant computational substrate for modern large language models (LLMs) and multimodal large language models (MLLMs). Their defining operation is *attention*, which dynamically allocates weight across tokens to compute contextualized representations. Given a sequence of token embeddings, each layer computes query, key, and value projections and produces an output that is a weighted combination of the values, where the weights are determined by the compatibility between queries and keys. This mechanism allows every token in the sequence to attend to every other token, enabling the model to capture long-range dependencies and perform implicit information retrieval within its context window.

In multimodal architectures, visual content is typically encoded into a sequence of continuous visual tokens—for example, patch features produced by a ViT (Dosovitskiy et al., 2020) or CLIP-style encoder (Radford et al., 2021)—and then aligned into the language model’s embedding space through a cross-modal connector such as an MLP projector (Liu et al., 2023), Q-Former (Dai et al., 2023), or resampler (Alayrac et al., 2022). The resulting unified sequence of visual and textual tokens is processed by the backbone LLM in the same autoregressive framework used for text-only generation. Crucially, this means that the same attention mechanism that mediates token-to-token dependencies in language also mediates the flow of information from visual tokens to the language model’s generation process.

This design yields a critical implication for multimodal reasoning: although outputs are generated as discrete text tokens, producing correct answers often requires precise internal selection and aggregation of multimodal evidence, governed largely by cross-modal attention patterns. A model that attends to the wrong visual region, or that allows irrelevant textual context to overshadow salient visual information, will fail regardless of how well its output distribution has been calibrated. This observation motivates the three broad strategies explored in this dissertation: (i) improving *which* visual tokens enter the context in the first place (Chapter 5), (ii) enabling reasoning *directly in latent multimodal representations* rather than translating everything into text (Chapter 6), and (iii) *optimizing attention behavior explicitly* through reinforcement learning rather than relying on indirect supervision through token-level objectives (Chapter 7). The earlier chapters on entity typing (Chapter 3) and reasoning diagnostics (Chapter 4) lay the groundwork by establishing that representational choices matter profoundly even in the text-only setting.

The remainder of this chapter reviews the specific technical areas that inform each contribution.

2.3 Entity Typing and Natural Language Inference

Entity Typing

Entity typing—the task of assigning semantic type labels to entity mentions in context—has been a long-standing problem in natural language understanding. The foundational formulation by Ling and Weld (2012) cast the problem as multi-label classification over a predefined type inventory. A central challenge in early work was the scarcity of annotated training data, which motivated a series of efforts based on automatic annotation using knowledge base links and distant supervision (Gillick et al., 2014; Ling and Weld, 2012; Yogatama et al., 2015). Subsequent research addressed the noise inherent in distant supervision and the structural dependencies among type labels. Hierarchy-aware loss functions were introduced to model parent–child relationships in type ontologies (Ren et al., 2016; Xu and Barbosa, 2018), and external knowledge from knowledge bases was incorporated to capture semantic relations among types (Jin et al., 2019; Dai et al., 2019; Obeidat et al., 2019).

The introduction of the Ultra-Fine Entity Typing (UFET) benchmark by Choi et al. (2018) marked a significant shift in the field. UFET expanded the type space to over ten thousand free-form candidates, moving far beyond the closed inventories of earlier benchmarks and exposing fundamental limitations of classification-based approaches. In response, several methods sought to model the structure of this large label space more effectively. Xiong et al. (2019) employed graph propagation layers to capture implicit type dependencies. Onoe and Durrett (2019) trained filtering and relabeling models on human-annotated data to denoise automatically generated training sets. Onoe et al. (2021) introduced box embeddings (Vilnis et al., 2018) to represent hierarchical type dependencies as geometric containment relationships among axis-aligned hyper-rectangles. To address data insufficiency, Dai et al. (2021) leveraged pre-trained language models for augmenting training data through masked entity generation.

Despite these advances, a fundamental limitation persisted: all of these systems encoded types as numerical indices, discarding the semantic content of the type labels themselves. This representational choice prevents models from generalizing to types unseen during training and limits their effectiveness in low-resource settings where individual types have few annotated examples. The work presented in Chapter 3 addresses this limitation by reformulating entity typing as a natural language inference problem, thereby preserving the semantics of type labels and enabling zero-shot generalization.

Natural Language Inference as a Tool for Generalization

Natural language inference (NLI)—the task of determining whether a hypothesis is entailed by, contradicted by, or neutral with respect to a given premise—has evolved from early approaches grounded in lexical semantics and syntactic analysis (Dagan et al., 2006) to modern neural methods that leverage pre-trained language models. Parikh et al. (2016) introduced attention-based decomposable models using bidirectional LSTMs (Hochreiter and Schmidhuber, 1997), while more recent systems build on pre-trained encoders such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), trained on large-scale NLI corpora including SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018).

The generalization capabilities of pre-trained NLI models have led to their adoption as indirect supervision signals for a range of prediction tasks with limited labeled data. Yin et al.

(2019) demonstrated the viability of NLI-based zero-shot and few-shot text classification by casting class labels as hypotheses. Shen et al. (2021) applied NLI-based document similarity for multi-class classification, and Sainz et al. (2021) leveraged NLI for few-shot and zero-shot relation extraction. In the question answering domain, Chen et al. (2021) proposed verifying QA system outputs by converting question–answer pairs into hypotheses and extracting textual evidence as premises. Yin et al. (2020) and White et al. (2017) explored NLI as a unified solver for multiple text understanding tasks, including coreference resolution and multiple-choice QA.

Prompt-based approaches have also emerged as a related strategy for leveraging the implicit knowledge in pre-trained models. Ding et al. (2021) adopted prompts to model the relationship between entities and type labels, using masked language models for label generation. While superficially similar to template-based hypothesis construction, prompt-based methods differ in that they query the generative distribution of a language model rather than the entailment judgment of an NLI model. The LITE framework introduced in Chapter 3 takes the latter approach, treating entity typing as a ranking problem over NLI-scored hypotheses, which enables principled handling of large, open-ended label spaces and cross-dataset transfer.

2.4 Multi-Hop Reasoning and Shortcut Vulnerabilities

Reasoning in Large Language Models

Large language models have demonstrated increasingly sophisticated reasoning abilities across diverse domains, including program synthesis (Kuznia et al., 2022), mathematical reasoning (Mishra et al., 2022), symbolic logic (Gaur and Saunshi, 2023), and commonsense inference (Feng et al., 2023). A key enabler of these capabilities has been the development of inference-time strategies that elicit structured reasoning. Chain-of-thought (CoT) prompting (Wei et al., 2022) encourages models to generate intermediate reasoning steps, while decomposition-based methods break complex queries into manageable sub-problems (Zhou et al., 2022; Yao et al., 2023). When combined with these techniques, LLMs have achieved strong results on multi-hop and complex question-answering benchmarks (Geva et al., 2021).

However, strong benchmark performance does not necessarily imply faithful reasoning. A growing body of work has questioned whether LLMs genuinely execute the multi-step inference implied by their chain-of-thought outputs, or whether they primarily exploit statistical regularities in the pre-training data to produce plausible answers. This concern is closely related to the broader problem of shortcut learning in neural networks, where models learn decision rules that perform well on in-distribution test sets but fail to capture the intended task structure (Geirhos et al., 2020). In the context of multi-hop reasoning, shortcuts may take the form of entity co-occurrence patterns: a model may “know” the answer to a multi-hop question not because it has performed the requisite chain of inference, but because the answer entity frequently co-occurs with other entities mentioned in the question within the training corpus.

Benchmarking Reasoning Faithfulness

Existing multi-hop QA benchmarks such as HotpotQA (Yang et al., 2018) and MuSiQue (Trivedi et al., 2022) have provided valuable testbeds for evaluating reasoning but were not specifically designed to control for shortcut exploitation. As a result, strong performance on these benchmarks may reflect a mixture of genuine reasoning and statistical shortcuts, making it difficult to diagnose the true reasoning capabilities of a model in isolation.

The EUREQA benchmark introduced in Chapter 4 addresses this gap by constructing questions where semantic shortcuts are systematically removed and replaced with deceptive cues. By controlling for entity co-occurrence and gradually increasing reasoning depth, EUREQA isolates the model’s capacity for faithful multi-hop inference from its ability to exploit distributional patterns. The benchmark draws on knowledge base facts that predate the training data of the evaluated models, ensuring that the required information should be accessible to the models while the surface-level shortcuts are not.

2.5 Vision-Language Models and Cross-Modal Grounding

Architecture and Training of Vision-Language Models

Vision-language models (VLMs) extend the capabilities of large language models to the visual domain by incorporating visual encoders and cross-modal connectors. The dominant architecture consists of three components: a pre-trained vision encoder, typically based on the Vision Transformer (ViT) (Dosovitskiy et al., 2020) or CLIP (Radford et al., 2021), which produces patch-level visual tokens; a cross-modal connector that maps these tokens into the language model’s embedding space; and a pre-trained LLM that processes the resulting multimodal token sequence (Alayrac et al., 2022; Dai et al., 2023; Liu et al., 2023).

Training generally proceeds in two stages. The first stage performs vision-language pre-training on image-caption pairs (Schuhmann et al., 2022; Singh et al., 2022) and interleaved image-text data (Laurençon et al., 2024; Zhu et al., 2024), aligning the visual and textual embedding spaces. The second stage applies instruction tuning with vision-grounded prompts (Liu et al., 2023; Zhang et al., 2023a), adapting the model to follow complex multimodal instructions.

The design of the cross-modal connector has been an active area of research. Early approaches used resampler modules (Alayrac et al., 2022) or Q-Former architectures (Bai et al., 2023; Dai et al., 2023) to compress visual information into a fixed number of tokens. More recent work has favored simpler MLP projectors, which map visual encoder outputs directly into the LLM token space and have been shown to outperform more complex alternatives in many settings (Liu et al., 2023).

High-Resolution Visual Understanding

A persistent challenge in VLMs is the tension between visual fidelity and computational efficiency. Standard visual encoders downsample input images to a fixed resolution, which can cause shape distortions and loss of fine-grained details critical for tasks such as optical character recognition (OCR), chart analysis, and detailed scene understanding (Liu et al.,

2024b; Thapa et al., 2024). The number of visual tokens produced by a patch-based encoder grows quadratically with image resolution (Tong et al., 2024b), creating a direct conflict between preserving visual detail and respecting the context length limits of the LLM backbone.

Several strategies have been proposed to address this challenge. High-resolution encoders (Li et al., 2024e; Lv et al., 2023) reduce excessive downsampling but produce proportionally more tokens. Multi-crop or patchification methods (Li et al., 2023a; Dong et al., 2024; Li et al., 2024f) segment images into overlapping sub-images that are individually encoded at native resolution and then concatenated. Training-based approaches such as Matryoshka Multimodal Models (Cai et al., 2024) and Scaling on Scales (Shi et al., 2025) fine-tune VLMs on instruction data specifically curated for high-resolution understanding. Other methods guide the model’s attention toward question-relevant regions through bounding box supervision (Shao et al., 2024b).

The SEMCLIP framework presented in Chapter 5 takes a complementary approach. Rather than encoding the entire image at high resolution, SEMCLIP uses textual signals to dynamically select the most task-relevant sub-image, appending only this selected region to the original encoding. This strategy achieves comparable or superior performance to full high-resolution methods while requiring significantly fewer visual tokens and no additional model training.

2.6 Latent Reasoning in Textual and Visual Domains

Chain-of-Thought Reasoning for Multimodal Models

The extension of chain-of-thought reasoning to multimodal settings has followed two broad paradigms. The first, which Su et al. (2025c) term “thinking about images,” performs reasoning entirely in text space. Under this paradigm, the LLM generates textual descriptions or analyses of visual content as intermediate reasoning steps before producing a final answer. Early approaches focused on constructing supervised fine-tuning datasets that demonstrate such reasoning patterns (Xu et al., 2024; Shao et al., 2024a; Wei et al., 2025b). More recently, the field has shifted toward reinforcement learning methods that allow models to discover effective reasoning strategies through trial and reward (Peng et al., 2025; Tan et al., 2025;

Meng et al., 2025; Yang et al., 2025a), with active investigation into reward design (Hong et al., 2025), training curricula (Wei et al., 2025a; Deng et al., 2025c; Liu et al., 2025b; Chen et al., 2025b), and verification mechanisms for quality control (Wang et al., 2025c; Shen et al., 2025a; Zhang et al., 2025b; Chen et al., 2025a). Complementary efforts have sought to reduce visual hallucination through auxiliary caption generation (Xia et al., 2025) or image masking (Wang et al., 2025g), and to ground answers in specific visual evidence by predicting points, bounding boxes, or region descriptions (Jiang et al., 2025; Ni et al., 2025; Yu et al., 2025; Liu et al., 2025c).

The second paradigm, “thinking with images,” augments the reasoning process with active visual manipulation. Methods in this family invoke external tools during chain-of-thought reasoning to zoom into image regions (Su et al., 2025a; Zhang et al., 2025d), draw auxiliary annotations (Hu et al., 2024), shift image styles (Liu et al., 2025a), highlight sub-regions (Fu et al., 2025c), or apply specialized processing such as OCR (Huang et al., 2025b). Early implementations relied on supervised fine-tuning to teach tool-use behaviors (Wang et al., 2025b; Zhang et al., 2025a; Chung et al., 2025), while more recent work employs reinforcement learning to learn when and how to invoke tools (Zhang et al., 2025d; Su et al., 2025b; Geng et al., 2025; Wu et al., 2025a), enabling interleaved reasoning and tool execution (Wu et al., 2025b; Zheng et al., 2025).

Both paradigms address a core limitation: despite sophisticated visual encoders, backbone LLMs often fail to capture the visual details most relevant to a query. This failure stems from modality projection bias (Zhang et al., 2024; Liu et al., 2023), modality interference (Cai et al., 2025; Wang et al., 2025e; Pezeshkpour et al., 2025), and cross-modality attention bias (Zhang et al., 2025c). However, both paradigms also share a fundamental limitation: they ultimately route visual understanding through textual channels, either by generating descriptive text tokens or by relying on the predefined operations of external tools.

Latent-Space Reasoning

A distinct line of research explores the possibility of reasoning directly in continuous representation spaces rather than discrete token spaces. In natural language processing, Hao et al. (2024) proposed passing last hidden states forward as continuous “thoughts” rather

than decoding them into text, demonstrating more efficient expression of complex reasoning. Subsequent work has investigated both fixed-length and variable-length latent reasoning strategies (Cheng and Van Durme, 2024), and Shen et al. (2025b) explored using latent representations to approximate token-level reasoning.

A key challenge for latent reasoning in the purely textual domain is interpretability: latent states lack the grounding that would make supervision straightforward. The visual domain offers a potential advantage in this regard, since latent visual tokens correspond to spatially localized, semantically meaningful regions of the input image. Recent efforts have begun to explore this direction, using latent spaces to capture visual content for reasoning (Yang et al., 2025b; Bigverdi et al., 2025). However, many of these approaches require auxiliary images to supervise the latent tokens, introducing additional labeling costs and limiting scalability.

A parallel line of work on omni-modality foundation models generates outputs in both text and visual modalities simultaneously (Team, 2024; Deng et al., 2025b; Xie et al., 2025). Some subsequent studies attempt to exploit image generation capabilities for reasoning, but effectiveness has been demonstrated only in narrow domains such as navigation and maze solving (Li et al., 2025b; Xu et al., 2025), and it remains unclear whether visual inputs can be faithfully preserved through decode–re-encode cycles.

The Latent Visual Reasoning paradigm introduced in Chapter 6 addresses these limitations by training the LLM to generate latent states that reconstruct question-relevant visual features directly in the shared embedding space, without requiring auxiliary images or external tools, and without decoding visual information into text.

2.7 Reinforcement Learning for Post-Training and Knowledge Distillation

Post-Training with Reinforcement Learning

Post-training through reinforcement learning has become the standard approach for aligning large language models with human intent (Ouyang et al., 2022). The canonical RLHF pipeline consists of three stages: supervised fine-tuning on demonstration data, training a

reward model to approximate human preferences (Christiano et al., 2017), and optimizing the language model policy via RL to maximize expected reward. Early implementations relied on Proximal Policy Optimization (PPO) (Schulman et al., 2017), which significantly improved model safety and helpfulness (Bai et al., 2022) but imposed substantial memory overhead due to its actor–critic architecture.

Group Relative Policy Optimization (GRPO) (Shao et al., 2024d) addressed this computational burden by replacing the learned critic with group-averaged reward estimates computed from multiple sampled responses. This simplification reduces overhead while maintaining strong performance, particularly in domains with verifiable rewards such as mathematical reasoning and code generation (Guo et al., 2025), giving rise to the paradigm of RL with Verifiable Rewards (RLVR).

Extending post-training to multimodal LLMs introduces challenges beyond text-only alignment, including visual hallucination and robust cross-modal grounding (Liu et al., 2023; Li et al., 2023c; Zhu et al., 2023). Recent methods have adapted RLHF, RLVR, and Direct Preference Optimization (DPO) to improve visual grounding and reduce hallucinations (Bai et al., 2025b; Zhu et al., 2025; Huang et al., 2025a). A persistent issue is modality bias, where models over-rely on linguistic priors or overfit to superficial visual cues (Cai et al., 2025). Recent work has designed reward functions that discourage text-only shortcuts, penalize spurious visual correlations, and promote faithful evidence-based responses (Xia et al., 2025; Wang et al., 2025f).

Critically, all of these methods operate at the level of token-level output distributions. They optimize *what* the model generates, but provide only indirect supervision for *where* the model attends internally. The Reinforced Attention Learning framework introduced in Chapter 7 addresses this gap by reformulating the RL policy to operate directly on attention distributions, providing a more direct training signal for cross-modal information selection.

Knowledge Distillation

Knowledge distillation transfers the learned representations of a high-capacity teacher model to a smaller student model by training the student to match softened output distributions rather than hard labels (Hinton et al., 2015). This approach provides richer supervisory

signals than standard training and has been widely adopted for model compression, domain adaptation, and efficient deployment (Romero et al., 2015; Zagoruyko and Komodakis, 2016). In the context of large language models, distillation has been extended beyond output logits to intermediate representations, attention maps, and hidden states, enabling improved preservation of structural and reasoning behaviors (Sun et al., 2019; Jiao et al., 2020; Wang et al., 2020).

A particularly relevant development is on-policy distillation (Agarwal et al., 2024), in which the student generates responses under its own policy and receives teacher supervision along these self-generated trajectories. Compared to offline distillation on static datasets, on-policy distillation mitigates exposure bias and better aligns the student’s learned distribution with deployment-time behavior. This paradigm bridges the gap between supervised distillation and RL-based post-training, retaining the stability of supervised objectives while benefiting from the distributional alignment properties of on-policy learning.

The on-policy attention distillation framework introduced in Chapter 7 builds on these foundations by incorporating attention-level supervision into the distillation process. Rather than transferring knowledge solely through token-level distribution matching, we additionally align the student’s internal attention patterns with those of the teacher, providing a principled mechanism for transferring fine-grained perceptual and grounding behaviors.

2.8 Summary

The background reviewed in this chapter highlights several recurrent themes that motivate the contributions of this dissertation. First, in entity typing and text classification, the representational choice of encoding labels as indices versus preserving their semantics has a profound impact on generalization, particularly in low-resource and open-world settings. Second, the impressive benchmark performance of large language models on reasoning tasks masks a fragility rooted in shortcut exploitation, which existing evaluation protocols are often unable to detect. Third, vision-language models face a fundamental tension between visual fidelity and computational efficiency, and current approaches to multimodal reasoning remain bottlenecked by the need to translate visual understanding into text. Fourth, latent-space reasoning offers a promising alternative to text-mediated reasoning, but existing approaches

in the visual domain are limited by their dependence on auxiliary supervision or external tools. Fifth, post-training through reinforcement learning has been transformative for language model alignment, but current methods optimize only the output token distribution, leaving the internal attention mechanism—arguably the most critical component for cross-modal information selection—as an unoptimized byproduct.

The following chapters present our contributions to addressing each of these challenges in turn.

Chapter 3

LITE: Ultra-Fine Entity Typing with Indirect Supervision from Natural Language Inference

3.1 Introduction

Entity typing—inferring the semantic types of entity mentions in text—is a fundamental problem in natural language understanding that supports downstream tasks such as entity linking (Ling et al., 2015; Onoe and Durrett, 2020), question answering (Yavuz et al., 2016), and information extraction (Koch et al., 2014). As discussed in Chapter 2, prior studies have predominantly formulated entity typing as a multi-way classification problem (Wang et al., 2021; Zhang et al., 2019; Chen et al., 2020; Hu et al., 2020), encoding candidate types as numerical indices. While this paradigm is adequate when the type inventory is small and closed, it proves insufficient for real-world scenarios where entity types are extremely diverse.

The Ultra-Fine Entity Typing (UFET) benchmark (Choi et al., 2018) exemplifies this challenge, requiring models to search among over ten thousand free-form type candidates. The dramatic increase in types raises two critical issues. First, classification-based systems that convert types into indices have no mechanism for understanding the semantics of the types themselves, making it difficult to match entity mentions to correct types when annotated data per type is scarce. Second, such classifiers cannot recognize types unseen during training

without complete retraining to accommodate new labels.

To address these challenges, this chapter presents **LITE** (Language InfERENCE based Typing of Entities), a learning framework that reformulates ultra-fine entity typing as a natural language inference (NLI) problem (Dagan et al., 2006). **LITE** treats each entity-mentioning sentence as a premise in NLI and transforms candidate types into textual hypotheses via template-based generation. The entailment score from an NLI model serves as the confidence for each type, and a learning-to-rank objective optimizes these scores so that hypotheses of correct types rank above those of incorrect types.

This formulation benefits entity typing from three perspectives. First, the inference ability of a pre-trained NLI model provides effective indirect supervision. Second, the hypothesis, as a type description, offers a semantically rich representation that benefits few-shot learning. Third, label dependency across granularities is captured by learning entailment relations between fine-grained and coarse-grained type descriptions. Experimental results on the UFET benchmark show that **LITE** drastically outperforms prior state-of-the-art systems (Dai et al., 2021; Onoe et al., 2021; Liu et al., 2021) without requiring distantly supervised data. **LITE** also yields the best performance on traditional fine-grained entity typing benchmarks and demonstrates strong transferability to datasets with unseen types.

The contributions of this chapter are three-fold. First, to our knowledge, this is the first work that uses NLI formulation and NLI supervision for entity typing, preserving label semantics and encoding label dependencies effectively. Second, **LITE** achieves state-of-the-art performance on both ultra-fine and regular fine-grained typing tasks, with particularly strong results on zero-shot and few-shot cases. Third, we demonstrate that **LITE**, once trained, can transfer to different test sets containing unseen types. The work presented in this chapter was previously published in Transactions of the Association for Computational Linguistics as "Ultra-fine entity typing with indirect supervision from natural language inference" (Li et al., 2022).

3.2 Method

In this section, we introduce the proposed method for ultra-fine entity typing with NLI. We begin with the problem definition and an overview of the framework (Section 3.2), followed

by technical details of type description generation (Section 3.2), label dependency modeling (Section 3.2), the learning objective (Section 3.2), and inference (Section 3.2).

Preliminaries

Problem Definition. The input to an entity typing task is a sentence s and an entity mention of interest $e \in s$. The task aims to assign one or more type labels from the label space L to e . For instance, in “*Jay is currently working on his Spring 09 collection, which is being sponsored by the YKK Group,*” the entity “Jay” should be labeled as *person*, *designer*, or *creator* rather than *organization* or *location*.

The structure of the label space L varies across benchmarks. In OntoNotes (Gillick et al., 2014), labels follow a strict hierarchy (e.g., */location/transit/bridge*). In FINGER (Ling and Weld, 2012), some labels have hierarchical dependencies while others are free-form (e.g., *currency*). For UFET (Choi et al., 2018), which is our primary benchmark, no ontology is provided and the label vocabulary consists of free-form words only.

Overview of LITE. Given a sentence with an entity mention, LITE treats the sentence as the premise in NLI, then learns to type the entity in three consecutive steps (illustrated in Figure 3.1). First, LITE employs a template-based technique to generate a natural language description for each type candidate, which serves as the hypothesis. Second, to capture label dependency, LITE generates type descriptions for any ancestor types and learns their logical dependencies. Third, LITE fine-tunes a pre-trained NLI model with a learning-to-rank objective that ranks positive type descriptions above negative-sampled ones according to entailment scores. During inference, the model predicts types whose hypothetical descriptions receive the highest entailment scores.

Type Description Generation

Given each sentence s with an annotated entity mention e , LITE generates a natural language type description $T(a)$ for each type label annotation a , which acts as a hypothesis in NLI. We consider three generation techniques:

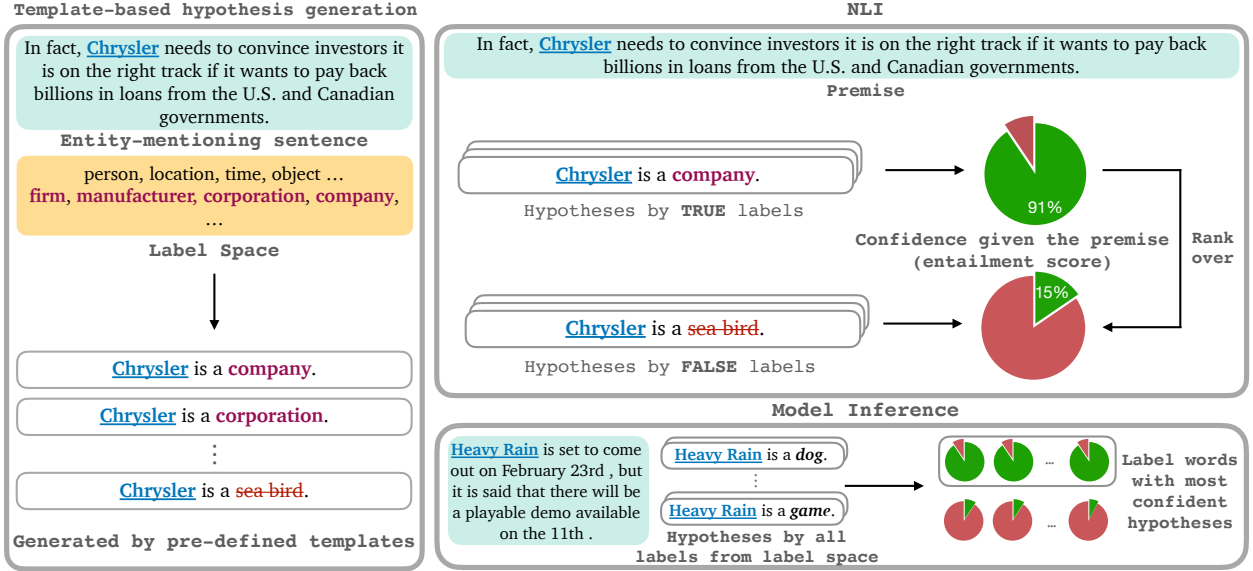


Figure 3.1: Entity typing by LITE with indirect supervision from NLI.

- *Taxonomic statement.* This template directly connects the entity mention and the type label with an “is-a” statement: “[ENTITY] is a [LABEL]”.
- *Contextual explanation.* This template generates a declarative sentence with a context-related connective: “In this context, [ENTITY] is referring to [LABEL]”.
- *Label substitution.* Inspired by the coreference-to-NLI transformation of Yin et al. (2020), this technique directly replaces the [ENTITY] in the original sentence with [LABEL], so the NLI model treats the modified sentence as the hypothesis.

Table 3.1 shows examples of each template. Each provides a semantically meaningful way to connect the entity with a candidate label, enabling the NLI model to leverage its inference ability to assess the entity-label relationship.

We also explored automatic template generation using the method of Gao et al. (2021), which employs the pre-trained T5 model (Raffel et al., 2020) to fill template placeholders. The majority of T5-generated templates followed the pattern of a prepositional phrase (e.g.,

Templates	Type Descriptions	Premise-Hypothesis Pairs for NLI
Taxonomic Statement	<u>Jay</u> is a producer .	Premise: " <u>Jay</u> is currently working on his Spring 09 collection, ..." Hypothesis: " <u>Jay</u> is a producer ."
Contextual Explanation	In this context, <u>career at a company</u> is referring to duration .	Premise: "No one expects a <u>career at a company</u> any more, ..." Hypothesis: "In this context, <u>career at a company</u> is referring to duration ."
Label Substitution	Musician knows how to make a hip-hop record sound good.	Premise: " <u>He</u> knows how to make a hip-hop record sound good." Hypothesis: " Musician knows how to make a hip-hop record sound good."

Table 3.1: Type description instances of three templates. Entity mentions are boldfaced and underlined while **label words** are only boldfaced.

in fact, in this context) followed by a statement such as "[ENTITY] is a [LABEL]." This observation guided our selection of the contextual explanation template.

Based on preliminary results on the development set, the taxonomic statement template generally yields the best performance across most settings (see Section 3.3). The main experiments therefore report results using this template.

Modeling Label Dependency

Entity type vocabularies often form hierarchies that enforce logical dependency among labels of different specificity. LITE extends the type description generation process to capture such dependencies. For a specific type label whose description has been generated, if ancestor types exist, LITE generates descriptions for each ancestor and learns entailment relations among them. The descendant type description serves as the premise and the ancestor description serves as the hypothesis. For instance, in OntoNotes (Gillick et al., 2014), if the entity *London* is labeled as */location/city* and the taxonomic statement template is used, LITE generates both "London is a city" and "London is a location." The fine-grained description "London is a city" then acts as the premise for the coarse-grained hypothesis "London is a location," helping the model capture the dependency between these labels.

This technique adapts naturally to tasks where no type ontology is available but instances are annotated with multiple labels of different specificity. For UFET (Choi et al., 2018), while

no ontology is provided, the task separates labels into *general*, *fine*, and *ultra-fine* categories. For an entity *Mike Tyson* simultaneously labeled as *person* (general), *sportsman* (fine), and *boxer* (ultra-fine), each pair of descendant and ancestor descriptions is added to training.

Learning Objective

Let L denote the type vocabulary. Given a sentence s with a mentioned entity e , we use P to denote all true type labels of e , including both the original labels and any induced ancestor labels (Section 3.2). For each label $p \in P$, the NLI model computes an entailment score $\varepsilon(s, H(p)) \in [0, 1]$ for the premise s and the hypothesis $H(p)$ generated by one of the templates in Section 3.2. A negative sample $p' \in L \setminus P$ is randomly selected and scored as $\varepsilon(s, H(p'))$. The margin ranking loss for an annotated training case is:

$$\mathcal{L}_t = [\varepsilon(s, H(p')) - \varepsilon(s, H(p)) + \gamma]_+, \quad (3.1)$$

where $[x]_+ = \max(x, 0)$ and γ is a non-negative margin constant.

A similar ranking loss captures label dependency. For any existing pair of ancestor type p_{an} and descendant type p_{de} from P , the training process regards $H(p_{de})$ as the premise and $H(p_{an})$ as the hypothesis, yielding score $\varepsilon(H(p_{de}), H(p_{an}))$. The label dependency loss is:

$$\mathcal{L}_d = [\varepsilon(H(p_{de}), H(p'_{an})) - \varepsilon(H(p_{de}), H(p_{an})) + \gamma]_+, \quad (3.2)$$

where p'_{an} is a negative-sampled type label.

The overall learning objective optimizes the following joint loss:

$$\mathcal{L} = \frac{1}{|S|} \sum_{s \in S} \frac{1}{|P_s|} \sum_{p \in P_s} \mathcal{L}_t + \lambda \mathcal{L}_d, \quad (3.3)$$

where S denotes the dataset, P_s denotes the set of true labels for an entity in sentence s , and λ is a non-negative hyperparameter controlling the influence of dependency modeling.

Inference

During inference, LITE ranks descriptions for all type labels in the vocabulary. For any given sentence s mentioning an entity e , the model generates a type description for each

candidate label. Taking s as the premise, the fine-tuned NLI model ranks the hypothetical type descriptions by their entailment scores. LITE then selects labels whose descriptions receive scores above a threshold, or selects the top-ranked label when single-label prediction is required.

3.3 Experiments

We evaluate the LITE framework on the UFET benchmark (Section 3.3) and on traditional fine-grained entity typing tasks (Section 3.3), followed by comprehensive analyses (Section 3.3).

Ultra-Fine Entity Typing

We use the UFET benchmark created by Choi et al. (2018). The dataset consists of two parts: (i) human-labeled data (L) with 5,994 instances split equally into train/dev/test (1,998 each); and (ii) distant supervision data (D) comprising 5.2M instances labeled by linking entities to a knowledge base, plus 20M instances generated by headword extraction. We follow the original benchmark design and evaluate using loose macro-averaged precision (P), recall (R), and F1.

Training Data. In our approach, supervision can come from the MNLI data (NLI) (Williams et al., 2018), distant supervision data (D), and human-labeled data (L). We investigate the best combination through the following training pipelines:

- LITE_{NLI} : Pre-train on MNLI, then predict directly without tuning on D or L.
- LITE_{L} : Fine-tune only on L.
- $\text{LITE}_{\text{NLI}+\text{L}}$: Pre-train on MNLI, then fine-tune on L.
- $\text{LITE}_{\text{D}+\text{L}}$: Pre-train on D, then fine-tune on L.
- $\text{LITE}_{\text{NLI}+\text{D}+\text{L}}$: Pre-train on MNLI, then on D, finally fine-tune on L.

Model Configurations. The system is initialized from RoBERTa-large (Liu et al., 2019) and optimized with AdamW (Loshchilov and Hutter, 2018). Hyperparameters are tuned on

the development set: batch size 16, pre-training learning rate 1e-6, fine-tuning learning rate 5e-6, margin $\gamma = 0.1$, and $\lambda = 0.05$. Pre-training epochs are limited to 5; fine-tuning epochs are limited to 2,000 with evaluation every 30 epochs on the development set.

Baselines. We compare LITE with the following systems: **UFET-biLSTM** (Choi et al., 2018), which uses GloVe embeddings and a bi-LSTM encoder with a type label embedding matrix; **LabelGCN** (Xiong et al., 2019), which adds a GCN layer to capture latent label dependency; **LDET** (Onoe and Durrett, 2019), which uses ELMo embeddings and trains a filter and relabeler to clean distant supervision data; **Box4Types** (Onoe et al., 2021), which uses BERT-large with box embeddings for type dependency; **LRN** (Liu et al., 2021), which uses BERT-base with LSTM-based auto-regression for label relation reasoning; and **MLMET** (Dai et al., 2021), the prior state-of-the-art, which augments distant supervision data using BERT masked language modeling and trains a linear classifier.

Results. Table 3.2 compares LITE with all baselines, using the taxonomic statement template. Overall, $\text{LITE}_{\text{NLI}+\text{L}}$ achieves state-of-the-art performance, outperforming MLMET by 1.5% absolute F1. This comparison verifies the effectiveness of the entailment scheme and indirect supervision from NLI—while MLMET builds a multi-way classifier on newly collected distant supervision data and human-labeled data, LITE optimizes a textual entailment scheme on MNLI and the human-labeled entity typing data alone.

The bottom block of Table 3.2 explores different training data combinations. Training on MNLI alone (LITE_{NLI}) yields poor results, indicating that MNLI does not generalize directly to UFET. Removing NLI supervision (LITE_{L} vs. $\text{LITE}_{\text{NLI}+\text{L}}$) causes a noticeable performance drop. The comparison between $\text{LITE}_{\text{NLI}+\text{L}}$ and $\text{LITE}_{\text{D}+\text{L}}$ demonstrates that MNLI, as an out-of-domain resource, provides more beneficial supervision than distant annotations—to our knowledge, this is the first work showing that indirect supervision from entailment data can be more effective than domain-specific distant supervision for entity typing. However, incorporating distant supervision into $\text{LITE}_{\text{NLI}+\text{L}}$ ($\text{LITE}_{\text{NLI}+\text{D}+\text{L}}$) degrades performance, which we analyze further in Section 3.3.

Label dependency modeling contributes a notable improvement: removing it from $\text{LITE}_{\text{NLI}+\text{L}}$ drops F1 from 50.6 to 49.7, with the main effect being a large increase in recall

Model		P	R	F1
UFET-biLSTM (Choi et al., 2018)		48.1	23.2	31.3
LabelGCN (Xiong et al., 2019)		50.3	29.2	36.9
LDET (Onoe and Durrett, 2019)		51.5	33.0	40.1
Box4Types (Onoe et al., 2021)		52.8	38.8	44.8
LRN (Liu et al., 2021)		54.5	38.9	45.4
MLMET (Dai et al., 2021)		53.6	45.3	49.1
LITE	NLI	1.5	7.1	2.5
	L	48.7	45.8	47.2
	D+L	27.5	56.4	37.0
	NLI+D+L	45.4	49.9	47.4
	NLI+L	52.4	48.9	50.6
	– w/o label dependency	53.3	46.6	49.7

Table 3.2: Results on the ultra-fine entity typing task. LITE uses the taxonomic statement template. “w/o label dependency” is applied to the NLI+L setting. The F1 result by LITE_{NLI+L} is statistically significant ($p < 0.01$ in t -test) compared to MLMET.

(from 46.6 to 48.9) at the cost of minor precision loss.

Fine-Grained Entity Typing

We further evaluate LITE on two traditional fine-grained benchmarks to assess (i) its effectiveness with smaller type sets, and (ii) whether a model trained on UFET can transfer to other tasks containing unseen types.

OntoNotes (Gillick et al., 2014) contains 3.4M automatically labeled entity mentions for training and 11k manually annotated instances (8k dev, 2k test), with a label space of 89 types. **FIGER** (Ling and Weld, 2012) contains 2M training samples labeled with 113 types, with 1,000 dev and 562 test instances. Within FIGER, 82 types have hierarchical dependencies while 30 are free-form.

Table 3.3 reports results for two variants of LITE: one pre-trained on UFET and transferred directly (zero-shot), and one with task-specific training after NLI pre-training. The task-specific variant outperforms the respective prior state-of-the-art on both benchmarks (OntoNotes: 86.4 vs. 85.4 macro-F1; FIGER: 86.7 vs. 84.9 macro-F1).

Model		OntoNotes		FIGER	
		macro-F1	micro-F1	macro-F1	micro-F1
Hierarchy-Typing (Chen et al., 2020)		73.0	68.1	83.0	79.8
Box4Types (Onoe and Durrett, 2020)		77.3	70.9	79.4	75.0
DSAM (Hu et al., 2020)		83.1	78.2	83.3	81.5
SEPREM (Xu et al., 2021)		–	–	86.1	82.1
MLMET (Dai et al., 2021)		85.4	80.4	–	–
LITE	Pre-trained on NLI+UFET	86.6	81.4	80.1	74.7
	NLI+task-specific training	86.4	80.9	86.7	83.3

Table 3.3: Results on fine-grained entity typing. All **LITE** results are statistically significant ($p < 0.05$ in t -test) compared to the best baseline (MLMET on OntoNotes, SEPREM on FIGER).

A notable advantage of **LITE** is its **transferability** across benchmarks. **LITE** pre-trained on UFET achieves competitive performance on both OntoNotes and FIGER with zero-shot transfer, even exceeding the task-specific variant on OntoNotes (86.6 macro-F1). This is particularly remarkable given that 12.5% of OntoNotes types and 16.8% of FIGER types are unseen in the UFET type space. Traditional multi-way classifiers, once trained on one benchmark, must be retrained to handle new benchmarks. In contrast, **LITE**’s reliance on a ranking objective and understanding of type semantics enables prediction of types never encountered during training.

Analysis

We provide analyses addressing several key questions: why distant supervision data did not help, the effectiveness of each type description template, **LITE**’s generalization to zero-shot and few-shot types, qualitative patterns in predictions, and time efficiency.

Effect of Distant Supervision Data. As Table 3.2 indicates, adding distant supervision data in $\text{LITE}_{\text{NLI}+\text{D}+\text{L}}$ leads to a 3.2% absolute drop in F1 compared to $\text{LITE}_{\text{NLI}+\text{L}}$. This is attributable to pervasive noise in the distant supervision data (Onoe and Durrett, 2019). Table 3.4 lists representative problems. Entity linking introduces a large number of false

Data Source	Sentence	Labels
Entity Linking	(a) From 1928–1929, he enrolled in graduate coursework at Yale University in New Haven, Connecticut .	location , <i>author</i> , province , <i>cemetery</i> , <i>person</i>
	(b) Once Upon Andalusia is a video game based on the film of the same name.	art , film
Head Word	(c) You can also use them in casseroles and they can be grated and fried if you want to make hash browns .	<i>brown</i>
	(d) He has written a number of short stories in different fictional worlds, including Dragonlance.	<i>number</i>
	(e) Despite obvious parallels and relationships, video art is not film .	film

Table 3.4: Examples from two sources of distant supervision data. **Correct types** are boldfaced; *incorrect ones* are in grey.

Templates	LITE _{NLI+L}			LITE _{NLI+D+L}			LITE _{D+L}		
	P	R	F1	P	R	F1	P	R	F1
Taxonomic Statement	52.4	48.9	50.6	45.4	49.9	47.4	27.5	56.4	37.0
Contextual Explanation	50.8	49.2	50.2	45.3	48.5	46.8	26.9	55.4	36.2
Label Substitution	47.4	49.3	48.3	42.5	50.7	46.2	24.8	59.3	35.0

Table 3.5: Behavior of different type description templates under three training settings.

positive types—for instance, the state *Connecticut* may be mislabeled as *author*, *cemetery*, and *person*. Headword extraction struggles with compositional expressions: *hash brown* is labeled as *brown*, and *a number of short stories* yields *number* as the type. Moreover, distant supervision often fails to recall all positive types, meaning that correct but un-annotated types are treated as negative samples during training, actively harming LITE’s learning-to-rank objective.

Type Description Templates. Table 3.5 reveals how template choices affect performance across three training settings. The taxonomic statement consistently outperforms the other two templates. Contextual explanation yields close but slightly lower results. Label substitution leads to a more noticeable F1 drop, likely because the absence of the entity mention in the hypothesis makes it harder for the NLI model to ground the type assessment.

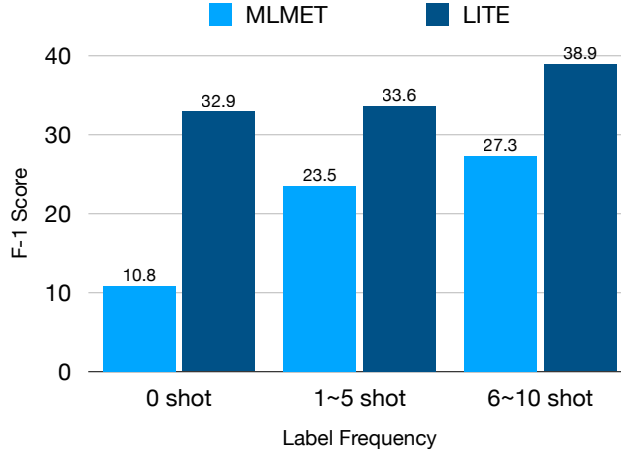


Figure 3.2: Performance comparison of LITE and MLMET on the filtered UFET benchmark for zero-shot and few-shot typing. Zero-shot labels correspond to the 40% of test types unseen in training.

Zero-Shot and Few-Shot Prediction. To rigorously evaluate generalization, we conducted an experiment in which portions of UFET training labels are randomly filtered out so that 40% of the test labels are unseen during training. Figure 3.2 shows the results of $\text{LITE}_{\text{NLI+L}}$ and the strongest baseline, MLMET. Note that while the held-out labels are completely unseen to LITE, the full type vocabulary is provided to MLMET during its language-model-based data augmentation.

As expected, performance on more frequent labels exceeds that on rare labels for both systems. However, LITE outperforms MLMET across all zero-shot and few-shot frequency bins. The advantage is most dramatic on zero-shot labels: LITE achieves 32.9% F1 compared to MLMET’s 10.8%. This demonstrates that the NLI-based formulation, by preserving the semantics of type labels, enables far more reliable inference on types unseen during training.

Case Study. We randomly sampled 100 labels on which LITE improves over MLMET by at least 50% in F1 and identified three characteristic patterns, as illustrated in Table 3.6:

- **Contextual inference (28%):** LITE leverages broader contextual cues. For example, from “winning the 50-meter backstroke gold medal,” LITE correctly types a pronoun as *swimmer* in addition to *athlete*.

	Input	True Labels	Prediction
LITE exceeds MLMET	(a) The University of California communications major gave <u>her</u> mother a fitting present, surprising herself by winning the 50-meter backstroke gold medal.	athlete, person, swimmer*, contestant, scholar, child	LITE: <u>athlete</u> , <u>person</u> , <u>swimmer*</u> , female, student, winner MLMET: <u>athlete</u> , <u>person</u> , <u>child</u> , adult, female, mother, woman
	(b) The apology is being viewed as a watershed in Australia , with major television networks airing <u>it</u> live and crowd gathering around huge screens in the city.	event, apology*, plea, regret	LITE: <u>event</u> , <u>apology*</u> , ceremony, happening, concept MLMET: <u>event</u> , message
	(c) A drawing table is also sometimes called a mechanical desk because , for several centuries , most <u>mechanical desks</u> were drawing tables.	object, desk, furniture*, board, table	LITE: <u>object</u> , <u>desk</u> , <u>furniture*</u> MLMET: <u>object</u> , <u>desk</u> , computer
MLMET exceeds LITE	(d) He attended the University of Virginia , where he played <u>basketball and baseball</u> ; his brother Bill also played baseball for the University.	basketball*, baseball, fun, action, activity, contact sport, game, sport, athletics, ball game, ball, event	LITE: <u>activity</u> , <u>game</u> , <u>sport</u> , <u>event</u> , <u>ball game</u> , <u>ball</u> , <u>athletics</u> MLMET: <u>activity</u> , <u>game</u> , <u>sport</u> , <u>event</u> , <u>basketball*</u>
	(e) The <u>manner in which it was confirmed</u> however smacked of an acrimonious end to the relationship between club and player with Chelsea.	manner*, way, concept, style, method	LITE: event MLMET: <u>manner*</u> , event

Table 3.6: Case study of labels on which LITE improves MLMET or vice versa. **Correct predictions** are in blue; * marks the representative label for each discussed pattern.

- **Coreference (20%)**: LITE correctly resolves pronominal references. For the pronoun *it* in a sentence about an apology, LITE predicts *apology* while MLMET captures only surface-level cues.
- **Hypernym inference (19%)**: Even without explicit mention of *furniture* in text, LITE assigns high confidence to this type for *mechanical desks*, recognizing the hypernym relationship.

LITE falls short in two scenarios: **multi-word nominal entities (30%)**, where ambiguous type descriptions hinder performance (e.g., “basketball and baseball is a basketball” is semantically confusing); and **clausal entity mentions (28%)**, where clausal information distracts the model from the key entity.

Prediction on Different Categories of Entity Mentions. We also investigated performance across three categories of entity mentions from the UFET test data: named entities,

	Named Entity			Pronoun			Nominal		
	P	R	F1	P	R	F1	P	R	F1
LITE _{NLI+L}	58.6	55.5	57.0	51.2	57.5	54.2	45.3	47.1	46.2
MLMET	58.3	54.4	56.3	57.2	50.0	53.4	49.5	38.9	43.5

Table 3.7: Performance comparison of LITE and MLMET on named entity, pronoun, and nominal entity mentions.

pronouns, and nominals. For each category, we randomly sampled 100 instances. As shown in Table 3.7, LITE consistently outperforms MLMET across all three categories, with the most significant improvement on nominal phrases (46.2% vs. 43.5% F1). This aligns with LITE’s strength in hypernym inference. Typing on nominals is generally more challenging because nominal phrases with multiple words are harder for the encoder to represent, and nominals tend to be more abstract, leading LITE to generate overly abstract types (e.g., labeling *the drink* as *substance*).

Time Efficiency. LITE requires approximately 40 hours of training, compared to over 180 hours for the previous strongest model MLMET, on the UFET task.¹ During inference, LITE requires approximately 35 seconds per sentence to score the full type vocabulary of over 10,000 labels, compared to approximately 0.2 seconds for a standard multi-way classifier. This increased inference cost stems from encoding longer texts and computing multiple entailment scores per input. However, LITE is substantially more efficient for *dynamic type vocabularies*: when new types are added, LITE requires virtually no recomputation, whereas multi-way classifiers must be retrained with an extended classification layer (e.g., over 180 hours for MLMET).

¹All time estimates are based on a commodity server with a TITAN RTX GPU. Training and evaluation batch sizes are maximized to 16 and 128 for LITE and MLMET, respectively.

3.4 Conclusion and Transition

This chapter presented LITE, a framework that reformulates entity typing as a natural language inference problem. Through template-based type hypothesis generation, LITE leverages indirect supervision from NLI to produce semantically rich representations of entity types, alleviating the data scarcity problem that plagues ultra-fine entity typing. The learning-to-rank objective enables generalized prediction across benchmarks with disjoint type sets. Experimental results demonstrate that LITE achieves state-of-the-art performance on UFET, OntoNotes, and FIGER, with particularly strong results on zero-shot and few-shot types. A LITE model pre-trained on UFET also exhibits strong transferability, outperforming baselines when applied directly to OntoNotes and FIGER without further training.

Looking ahead. The results of this chapter establish a representational principle that will recur throughout this dissertation: reasoning over rich semantic descriptions of labels, rather than opaque numerical indices, enables models to generalize to novel concepts and transfer across tasks. However, a natural question arises: if the way we *represent* the output space matters this much, what happens when the model’s internal reasoning process itself relies on shallow representational shortcuts? The LITE framework assumes that the NLI backbone performs genuine semantic inference over premise–hypothesis pairs, yet large language models—including those that underpin modern NLI systems—have been shown to exploit spurious statistical patterns rather than faithful reasoning chains. In the next chapter, we turn from improving representations of *what models predict* to diagnosing representations of *how models reason*. We introduce EUREQA, a benchmark designed to expose the deceptive semantic shortcuts that large language models take when faced with multi-hop reasoning tasks, revealing that the representational bottleneck extends from the output space into the reasoning process itself.

Chapter 4

EUREQA: Deceptive Semantic Shortcuts on Reasoning Chains

4.1 Introduction

The previous chapter demonstrated that reasoning is more effective when grounded in rich semantic representations rather than arbitrary categorical indices. A natural follow-up question is whether the reasoning processes of large language models themselves are similarly grounded, or whether models exploit shallow representational shortcuts to produce plausible but unfaithful answers. This chapter investigates this question directly. The research described in this chapter was originally published in the Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies under the title "Deceptive semantic shortcuts on reasoning chains: How far can models go without hallucination?" (Li et al., 2024a).

Recent progress in large language models (LLMs) has showcased their emergent abilities across a wide range of reasoning tasks, including program synthesis (Kuznia et al., 2022), mathematical reasoning (Mishra et al., 2022), symbolic logic (Gaur and Saunshi, 2023), and commonsense reasoning (Feng et al., 2023). Innovations in inference methods—such as chain-of-thought prompting (Wei et al., 2022) and problem decomposition (Zhou et al., 2022; Yao et al., 2023)—have further improved these capabilities, and LLMs have achieved exceptional results on multi-hop and complex QA benchmarks (Geva et al., 2021).

Nonetheless, it remains under-investigated whether the effectiveness of LLMs’ reasoning pattern OpenAI (2023); Team (2023) is genuinely based on a sensible reasoning path or if these models predominantly depend on semantic associations (i.e., word distributions from the pre-training data) to generate a plausible answer. Figure 4.1 shows how ChatGPT¹ fails to find the missing entity as we gradually remove relevant semantic cues as the reasoning depth grows. The original question (labeled as “1 Layer” in Figure 4.1) asks models to find an individual that satisfies two distinct conditions. It can be resolved by finding the set of people who are producers of the film “Gardener of Eden” and the set of cast members of “J. Edgar”, and checking for the overlapping entity. However, language models such as ChatGPT may not necessarily perform this sensible entity-seeking process. This is because the correct answer can simply (but incorrectly) be inferred by the high occurrence of “Leonardo DiCaprio” in the model’s training data with “Gardener of Eden” and “J. Edgar”.² In such a scenario, ChatGPT may generate the correct response through this shortcut of semantic association. Using semantic shortcuts such as co-occurring entities works well on common test cases. Still, they limit systems’ generalizability and robustness on cases not aligning with pre-training distributions. This is evident if we remove entities directly relevant to the answer, such as “J. Edgar”, and replace them with recursive entity-seeking problems where models must first identify them, shown as “3 Layers” in Figure 4.1. We observe that ChatGPT no longer produces the right answer without enough semantic hints and hallucinates “Adam Sandler”, which may be because “Eden” is related “Adam”, and at the same time, both the writer and main character of “Gardener of Eden” is named “Adam.” In addition, ChatGPT almost always comes up with “Adam Sandler” if asked for actors named “Adam.”

To investigate and quantify whether LLMs can follow a correct reasoning path and resist the attempt to take “greedy shortcuts” during inference, we introduce EUREQA (Extending Underlying reasoning Chains in QA), a sophisticated multi-hop question answering dataset, meticulously designed for diminishing semantic associations and gauging the capability of LLMs in undertaking extensively chained reasoning processes. To create EUREQA, we design a method that removes semantic shortcuts that will lead to the correct answer

¹In this paper, we consistently utilize the gpt-3.5-turbo-0301 checkpoint for ChatGPT.

²In fact, if we prompt ChatGPT with only three keywords “Gardner of Eden,” “J. Edgar” and “actor” and let it generate anything, the outcome is almost always “Leonardo DiCaprio.”

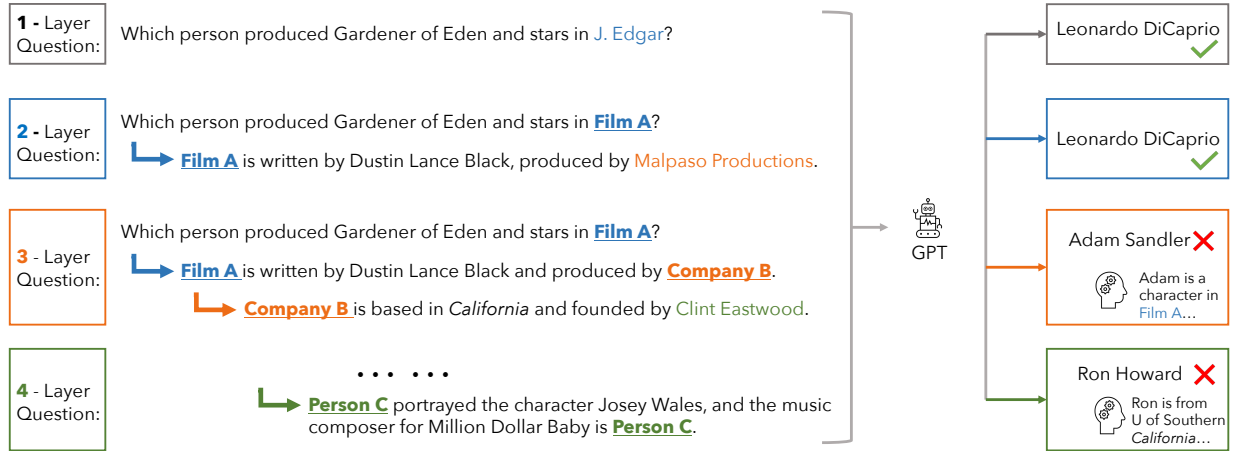


Figure 4.1: LLMs make errors when correct surface-level semantic cues—entities—are recursively replaced with descriptions, and the errors are likely related to token similarity. GPT-3.5-turbo is used for this example.

while adding “deceptive semantic cues” that are distracting and irrelevant to the correct answer. We also guarantee a simple and consistent reasoning path and only use common facts that should be memorized by the models. In this way, if an LLM fails on such tasks, it directly suggests that the LLM is taking incorrect shortcuts during inference. For instance, we modify the original question in Figure 4.1 by substituting “J. Edgar” with a placeholder dubbed “Film A” and supplementing a descriptive sentence about the film. Our intent with this strategy is to lessen the model’s reliance on direct information from the film name. We can automatically extend the depth of a reasoning chain by further replacing entities in the descriptive sentence about “Film A” with corresponding abstract names. In particular, this generation method is developed on a knowledge base, ensuring that the information employed predates the training data of the examined LLMs.

We then evaluate model performances on the EUREQA benchmark. We show that state-of-the-art LLMs,³ are incapable of proper reasoning for instances in EUREQA, with GPT-4 only achieving 62% to 64% accuracy in identifying common entities in Wikipedia and ChatGPT less than 40%. On the contrary, humans achieve near-perfect performances without

³In this context, ChatGPT and GPT-4.

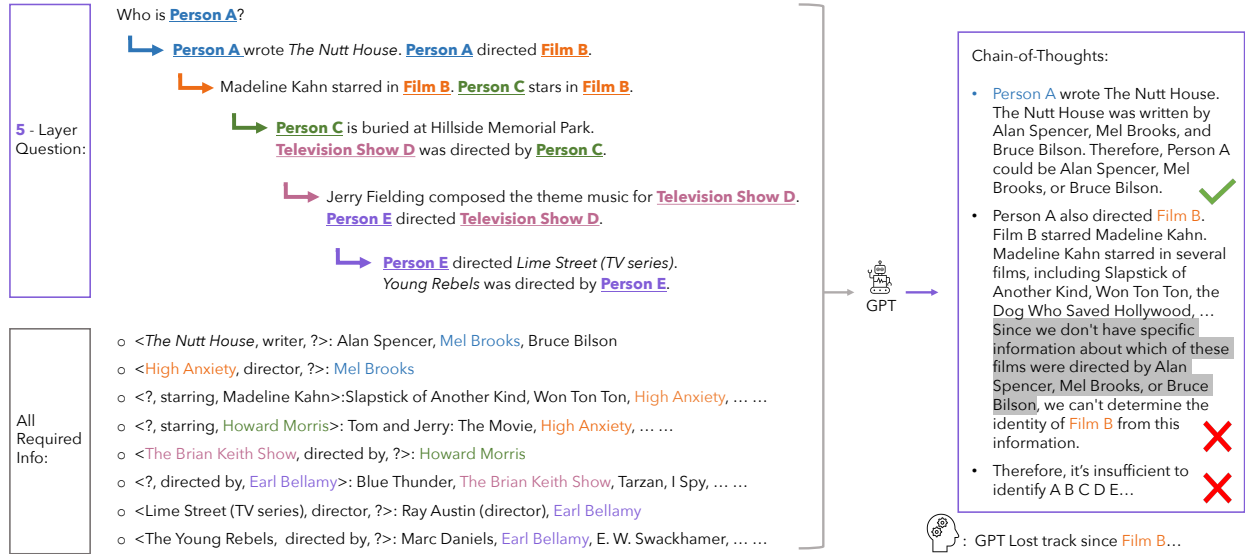


Figure 4.2: Even given all the required information needed for the question (selected information shown), GPT-4 still makes mistakes starting at early layers (highlighted in grey). Partial output is shown. Few-shot prompts are provided.

much effort because of the simplicity and consistency of the gold reasoning path. We also show that GPT-4 performance strongly correlates with the semantic similarities between the gold answer and other entities mentioned in the question. These findings combined suggest that even the best language models, with detailed in-context-learning (ICL) processes, still fall for deceptive semantic shortcuts and hence hallucinate and fail on EUREQA.

The contributions of this chapter are three-fold. First, we propose a novel method for generating QA data with extended reasoning chains that allow us to create deceptive semantic shortcuts. Second, we propose EUREQA, a dataset specifically designed for evaluating LLMs in scenarios with reduced or deceptive semantic associations. Third, our experimental findings establish that contemporary LLMs predominantly depend on semantic associations in question answering, failing to adhere to accurate reasoning processes as entities are replaced with alternate inquiries.

4.2 The EUREQA Benchmark

This section introduces the EUREQA dataset for evaluating LLMs on extended reasoning chains. We describe the reasoning chain construction (Section 4.2), question generation (Section 4.2), question refinement (Section 4.2), and dataset statistics (Section 4.2).

Reasoning Chain Construction

Every question in EUREQA is constructed through an implicit reasoning chain, as depicted in Figure 4.3. This chain is structured in layers, each comprising three components: an **entity** e_i , an associated **fact** f_i about e_i , and a **relation** r_i connecting e_i to the entity in the next layer, e_{i+1} . The identity of e_i is constrained by f_i , e_{i+1} , and r_i together—in most cases, neither f_i nor e_{i+1}/r_i alone can identify e_i , but we guarantee uniqueness when both constraints are applied. This design allows extending the reasoning chain by replacing e_i with f_i , r_i , and e_{i+1} , until the actual entity name of e_N is provided. A valid reasoning path starts from the provided e_N and resolves each e_{N-1} based on f_{N-1} and r_{N-1} until identifying e_1 , the answer.

To collect such data automatically, we use a structured knowledge base DBpedia Auer et al. (2007), which provides accurate facts and relations between entities, and at the same time, all these facts and relations should have been memorized by most LLMs since Wikipedia is included in pre-training corpora. We use the 2021-09 DBpedia snapshot to guarantee that the entity information was accessible to LLMs during their training phase. The criteria for choosing valid e_i , r_i , and f_i to form chains are the following:⁴

Entity. Each entity e_i must exist in the knowledge base as a recognized entity.

Relation. Each relation r_i should connect e_i with more than one potential candidate entities e_{i+1} . This ensures no direct semantic shortcut between e_i and e_{i+1} . For example, the relation “award” is deemed valid for the entity “Tiger Woods” because he has won multiple awards. In contrast, the relation “college” is invalid because DBpedia only has one corresponding entity, which is *Stanford University*.

⁴Being “valid” here refers to satisfying the following criteria for data extraction.

Fact. Facts are generated from DBpedia relations and destination entities for each e_i . A valid fact f_i consists of a relation r_i^f and a destination entity e_i^f (distinct from the chain relation r_i). We adopt two criteria for fact selection:

- **easy:** Select (r_i^f, e_i^f) pairs where e_i is the *unique* entity satisfying $(?entity, r_i^f, e_i^f)$. For example, *Jason Connery* is the only entity fitting $(?entity, \text{father}, \text{Sean Connery})$. In this case, f_i alone suffices to identify e_i .
- **hard:** Ensure that e_i is *not* the only entity satisfying the fact triplet. For instance, $(\text{parentCompany}, \text{Comcast})$ is a **hard** fact for *NBCUniversal* since *Comcast* is also the parent company of *Xfinity*.

These criteria yield two distinct datasets: EUREQA^{easy} and EUREQA^{hard}.

Chain Construction. Following the previously established criteria, we employ a random-walk algorithm to fabricate chains of reasoning. The procedure initiates with a designated seed entity e_1 in the knowledge base. Subsequently, a relation r_1 , an associated fact f_1 , and a subsequent entity e_2 are chosen randomly to construct a valid layer of the reasoning chain. This operation is continuously executed until no more valid layer is found or the cumulative number of layers equals N_{max} . We set $N_{max} = 5$ in this work. We perform at most 50 random walks starting from each seed entity to optimize efficiency and remove duplicated chains. Seed entities are sourced via two methods, either through prompting ChatGPT or by data extraction from .

This approach ensures that reasoning chain generation for EUREQA satisfies the following three properties of each chain:

- **Reasoning-dependent.** Every intermediate layer relies on information from the subsequent layer in the chain for resolution. Consequently, the model must commence from the terminal layer and engage in sequential reasoning throughout the chain.
- **Length-flexible.** The extent of the reasoning chain can be conveniently adjusted by adding or removing layers, thereby enabling an evaluation of the depth of reasoning.

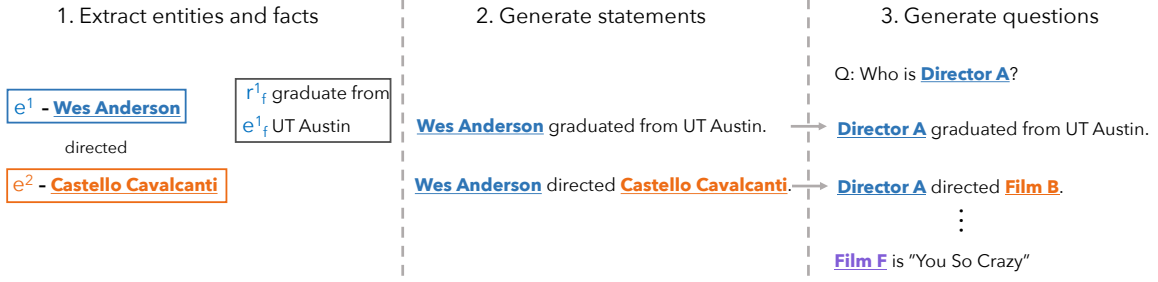


Figure 4.3: The data generation process of EUREQA.

- **Determinism-adjustable.** Determinism of the reasoning chain can be altered by omitting fact f_i in a layer, facilitating an evaluation of the model in addressing questions with multiple potential answers, which is not considered in this work.

Question Generation

In EUREQA, each question, denoted as q , is a natural language articulation of the above reasoning chains. A layer-by-layer procedure translates this structured chain into a human-readable text. Here we denote $q = q_0 + q_1 + \dots + q_n + m_{n+1}$ where $\{q_i | 1 \leq i \leq n\}$ indicates the sub-question for the i -th layer and “+” stands for concatenation. q_0 and m_{n+1} will be introduced later in this subsection. Figure 4.3 (2) provides a tangible example of this translation process. Every layer comprises an entity e_i , a relation r_i , a succeeding entity e_{i+1} , and an associated fact f_i , which includes the corresponding r_i^f and e_i^f . The initial step involves translating the triplets, (e_i, r_i, e_{i+1}) and (e_i, r_i^f, e_i^f) into two statements, $s_i^{relation}$ and s_i^{fact} , respectively.

The subsequent step involves substituting each entity e in every statement with a placeholder, as illustrated in Figure 4.3 (3). In particular, each layer now associates with a pair of statements, namely $s_i^{relation}$ and s_i^{fact} , which derive from the translation of the relational triplet (e_i, r_i, e_{i+1}) and the factual triplet (e_i, r_i^f, e_i^f) correspondingly. In this context, we proceed to solely obfuscate e_i and e_{i+1} in $s_i^{relation}$ and s_i^{fact} by substituting them with their respective hypernyms, h_i and h_{i+1} , whilst preserving the identity of e_i^f . Hypernym details

are extracted from DBpedia if available and sought from the LLM otherwise. ChatGPT is employed in our approach with two few-shot examples obtained from DBpedia, to ensure a consistent level of granularity of hypernym. Each hypernym is subsequently appended with a specific label that is assigned in alphabetical order. For example, the third layer of a reasoning chain could hold “Actor C stars in TV Series D” as $s_3^{relation}$ while s_3^{fact} could be “Actor C is the artist for *Along on Christmas Day*.” Here, e_3 is substituted by “Actor C” and e_4 , which also appears in the fourth layer, is replaced by “TV Series D.” The resulting statements with masked entities are denoted as $m_i^{relation}$ and m_i^{fact} respectively.

The final step involves generating an interrogative question, represented as q_0 , and an entity information statement, denoted as m_{n+1} , pertaining to q_n . The selection of the initial interrogative question is contingent upon the hypernym of the entity e_1 in the first layer. Specifically, q_0 starts with “Who is ...” if the hypernym h_1 corresponds to a person or “What is ...” in all other cases. In the example shown in Figure 4.3, “Wes Anderson” has the hypernym “director.” The generation process for m_{n+1} is essentially uncomplicated. Given that $m_i^{relation}$ in the terminal layer exhibits the relation between e_n and e_{n+1} , and no additional layers exist to identify e_{n+1} , we generate m_{n+1} utilizing the template “ h_{n+1} is e_{n+1} .”, which is exemplified as “Film F is *You So Crazy*.” in Figure 4.3.

Question Refinement

We perform **viability filtering** to check if a reasoning chain can be correctly followed to derive the expected answer. As mentioned above, each layer of a specific chain specifies e_i by both a relational triplet (e_i, r_i, e_{i+1}) and a factual triplet (e_i, r_i^f, e_i^f) . By excluding e_i from these triplets, we query our knowledge base with both $(?variable, r_i, e_{i+1})$ and $(?variable, r_i^f, e_i^f)$. The layer is deemed viable if e_i is a unique solution of $?variable$. We only retain those reasoning chains where each layer has passed the viability filtering.

This work aims to analyze models’ reasoning capabilities. As such, we employ a **knowledge filtering** procedure to ensure that most LLMs have sufficient world knowledge to answer our questions. To illustrate this, we consider a question, $q = q_0 + q_1 + \dots + q_n + m_{n+1}$, and verify both $s_i^{relation}$ and s_i^{fact} , where $1 \leq i \leq n$. In more explicable terms, we evaluate knowledge by presenting ChatGPT with the prompt “*Is this statement correct: [s_i] Yes or*

Difficulty	easy	hard	hard	hard	hard	hard
#Layers	5	5	4	3	2	1
Count	428	1363	300	300	300	300

Table 4.1: Statistics of EUREQA.

No?” A statement is considered as memorized if the LLM response includes a “Yes.”⁵ A q_i is deemed to have satisfied the examination criteria if and only if *both* of $s_i^{relation}$ and $s_i^{relation}$ are categorized as memorized.

Data Statistics

As indicated in 4.1, EUREQA encompasses a total of 2,991 questions. The dataset is split into two levels of difficulty: **hard** and **easy**. Specifically, the **easy** set comprises 428 five-layer questions while the **hard** set includes a larger set of 1,363 five-layer questions. This **easy** version enables manifesting LLM behavior towards questions with sufficient semantic shortcuts. From this **hard** set, we randomly select 300 five-layered questions. By removing layers sequentially, the **hard** set also yields 300 questions in each number of layers from four to one respectively. These **hard** questions with varying layers allow for a comprehensive assessment of LLMs in terms of their reasoning capabilities across various depths of reasoning. The distribution of categories of seed entities for the questions in EUREQA is shown in Figure 4.4, showcasing a broad spectrum of topics encompassed by these entities, effectively reducing any potential bias arising from specific entity categories.

4.3 Experimental Setup

This section presents the experiment configurations employed for assessing the long-chain reasoning capabilities of LLMs through EUREQA. This section begins with a discussion on model configurations (4.3), subsequently introducing the adopted prompting methodologies (4.3), and finally, it scrutinizes the evaluation configurations (4.3).

⁵we applied self-consistency strategy Wang et al. (2022) with three runs for a majority vote.

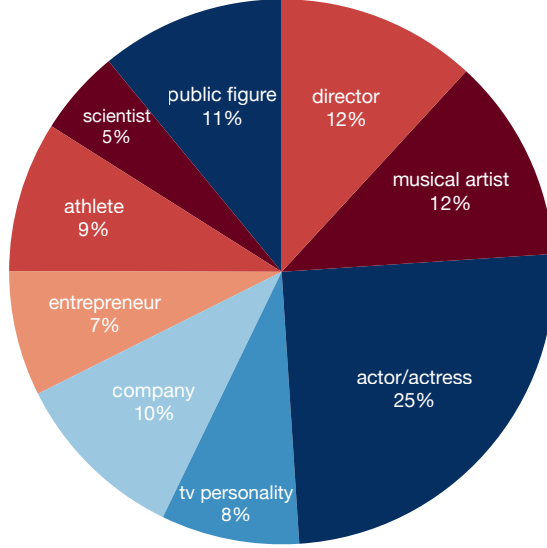


Figure 4.4: Categorical distribution of seed entities in EUREQA questions.

Model Configuration

We evaluate state-of-the-art LLMs on EUREQA: ChatGPT (gpt-3.5-turbo-0301), Gemini-Pro (Gemini 1.0 Pro), and GPT-4 (gpt-4-0314). All models are run with temperature $\tau = 0.8$.

Prompting Methods

We employ two prompting methods in our experiment: namely, *direct* and *icl*. The *direct* technique presents raw questions to the model without any supplementary context. On the other hand, *icl* method utilizes a few-shot approach, preceding each query with two context-specific examples. Each example is characterized by a sample problem, succeeded by the statement “Let’s solve this question step by step” and a step-by-step solution written by humans. In these solutions, the problem is analyzed from the final layer through the initial layer to give the correct answer, thereby obviating any necessity for backtracking during the problem-solving procedure.

Evaluation Protocol

We evaluate the **accuracy** of LLMs towards the correct answer. Considering that EUREQA functions as a free-form QA benchmark and our observation indicates that LLMs typically respond with comprehensive reasoning processes, following prior studies Agrawal et al. (2015); Ossowski and Hu (2023), we employ a string-match criterion: If the correct answer, identifiable as an entity, is present in the LLM response, we deem such a response as correct. We use a self-consistency of five runs with ChatGPT and Gemini-Pro, but we do not apply self-consistency to GPT-4 due to cost issues.

4.4 Results

The results of the state-of-the-art LLMs on EUREQA, as reported in Table 4.2, illustrate the variations in performance across different depths of reasoning and levels of difficulty. In general, with the entities recursively substituted by the descriptions of reasoning chaining layers, and therefore eliminating surface-level semantic cues, these models generate more incorrect answers. When the reasoning depth increases from one to five on **hard** questions, there is a notable decline in performance for all models, with an average accuracy decrease of 14.2% when using the *icl* prompt and 14.3% when using the *direct* prompt. The performance of LLMs is significantly higher on the **easy** set compared to the **hard** set. This is evidenced by a marked increase in accuracy, averaging 7.0% for ChatGPT, 15.4% for Gemini-Pro, and 19.7% for GPT-4. This finding underscores the significant impact that semantic shortcuts have on the accuracy of responses, and it also indicates that GPT-4 is considerably more capable of identifying and taking advantage of these shortcuts.

It can also be inferred from the results that the demonstrated human-written examples do help improve the performance. For both ChatGPT and GPT-4, using the *icl* prompt consistently leads to better performance than the *direct* prompt. This suggests that these models can benefit from examples that provide explicit and well-crafted human-written reasoning processes. These few-shot examples encourage the models to resolve the questions with the correct reason instead of guessing with semantic cues. Nonetheless, even with clear reasoning processes provided, models still tend to fail for incorrect shortcuts.

depth	hard										easy	
	d=1		d=2		d=3		d=4		d=5		d=5	
	<i>direct</i>	<i>icl</i>	<i>direct</i>	<i>icl</i>	<i>direct</i>	<i>icl</i>	<i>direct</i>	<i>icl</i>	<i>direct</i>	<i>icl</i>	<i>direct</i>	<i>icl</i>
ChatGPT	22.3	53.3	7.0	40.0	5.0	39.2	3.7	39.3	7.2	39.0	13.1	47.0
Gemini-Pro	45.0	49.3	29.5	23.5	27.3	28.6	25.7	24.3	17.2	21.5	30.6	38.9
GPT-4	60.3	76.0	50.0	63.7	51.3	61.7	52.7	63.7	46.9	61.9	66.4	81.8

Table 4.2: Accuracy of ChatGPT, Gemini-Pro, and GPT-4 across different reasoning depths d (number of layers) and difficulty levels. Two prompting strategies are evaluated: *direct* (zero-shot) and *icl* (two-shot with examples).

Another observation is that LLMs do not always perform worse on questions with more depth of reasoning. For instance, ChatGPT with *direct* prompt is 3.5% more accurate when the depth of reasoning d increases from four to five. We believe this also indicates that LLMs do not always follow the reasoning path and may instead answer the question based on surface-level semantic shortcuts. Further analyses can be found at Section 4.5

To identify the errors in these LLM responses, we randomly sampled 20 data points where GPT-4 with *icl*, the best performing method in this benchmark, has generated incorrect answers. The most frequent error pattern lies in hallucination during the reasoning process, encompassing 80% of the cases. In such cases, an intermediate reasoning stage erroneously conceived an inaccurate, albeit plausible answer, culminating in accumulated errors in the final answer. For instance, the “founder of a rocket company” was misconstrued as “*Elon Musk*, the founder of SpaceX”, rather than the correct answer of “*Jeff Bezos*, the founder of Blue Origin.” This phenomenon proves our claim that LLM relies on semantic cues for problem-solving instead of cognitive reasoning.

4.5 Analysis and Discussion

In this section, we perform further analyses and provide a more in-depth discussion of our findings.

Human Performance

Towards a more comprehensive insight into human proficiency on EUREQA, a human analysis was carried out focusing on the **hard** set characterized by a reasoning depth of five. This evaluation involves two computer science PhD students as the annotators to ensure their expertise. A selection of 50 questions, randomly extracted from the set, was subjected to annotation. Each question received annotations from both annotators independently, ensuring a dual perspective on every query. During the annotation process, annotators are presented with an input question to which they are required to provide an answer. To facilitate this, they are granted access to conduct information searches through the *dbpedia-snapshot-2021-09* database. This could be achieved either through visiting the specific entity page or by querying via a SPARQL portal. The averaged accuracy of annotators achieves 95% with an Inter-annotator Agreement of Cohen $\kappa = 0.79$.

Do LLMs Take Shortcuts?

One of our main motivations is to test if language models can follow a simple yet effective reasoning chain instead of taking semantic shortcuts based on entity associations. We design an analytical experiment based on entity similarity to investigate whether LLMs take such semantic shortcuts. Our intuition is to model the correlation between the performances and the averaged semantic similarities between the gold answer and other entities mentioned in the question. In an ideal situation where a model takes an optimal reasoning path for each instance, we will see a uniform distribution of performances: the accuracy for various degrees of entity similarities will be relatively the same. If the model relies much on entity biases and takes semantic shortcuts, we will see an increasing performance when the mentioned entities are more closely related to the target answer.

To infer entity similarities, we employ an off-the-shelf sentence Transformer model⁶ and encode the entity strings into embeddings. With such embeddings, we calculate the dot-product similarity between the target answer and all other Wikipedia entities mentioned in the question and compute an averaged similarity for each instance. We then draw the

⁶<https://huggingface.co/sentence-transformers/msmarco-distilbert-cos-v5>.

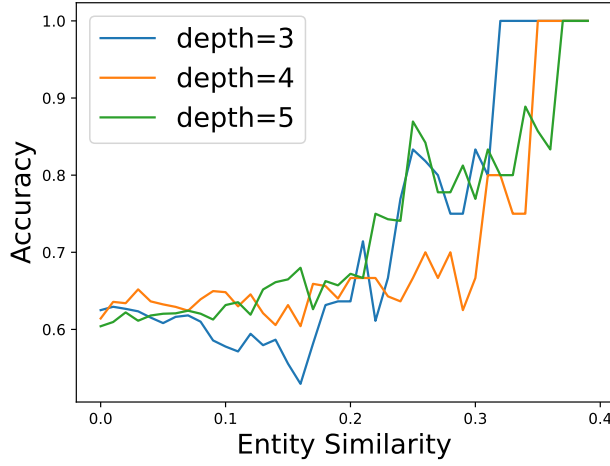


Figure 4.5: Correlation between GPT-4 performance on EUREQA hard set and entity similarities.

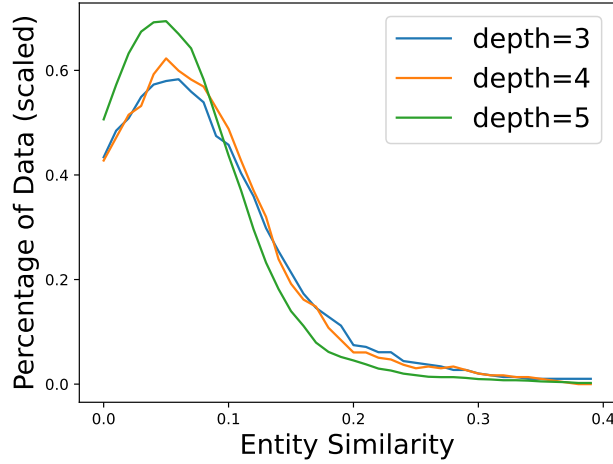


Figure 4.6: Distribution of entity similarity scores in EUREQA.

correlation curve of model performances on instances with certain similarities.⁷ This method derives a relatively continuous curve based on our limited evaluation data.

Observation 1. Fig. 4.5 shows that GPT-4’s performance positively correlates with the

⁷Specifically, we start with a similarity value $x_{start} = 0.0$ and increment it by a step of 0.01. For each x_{start} , we find instances that have an averaged similarity score between $[x_{start}, x_{start} + 0.1]$, and compute the accuracy on these instances.

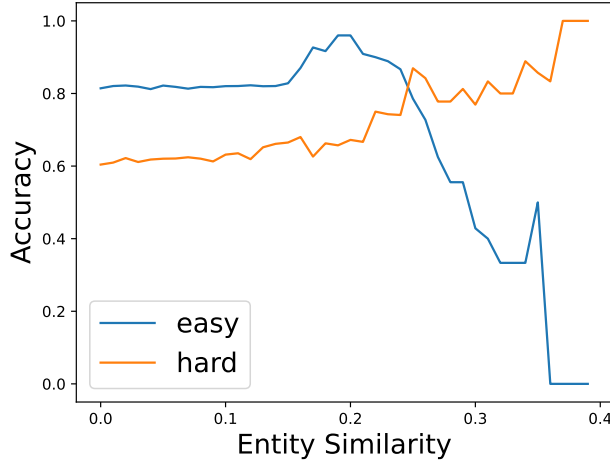


Figure 4.7: GPT-4 performance with different entity similarity scores between the **easy** and **hard** sets.

entity similarities in the instances. This shows that the model relies on semantic shortcuts instead of following the correct reasoning path.

Observation 2. GPT-4 performs similarly on different depths if the entity similarity scores are the same, especially on instances with lower similarities (< 0.25). This again suggests that the model relies on spurious entity biases to solve the question, because it does not do better on instances with shorter reasoning paths. That being said, the performance differences between lower and higher depths mostly come from the different data distributions, as shown in Fig. 4.6.

Observation 3. Fig. 4.7 shows another interesting finding, where GPT-4 does better on easy instances with lower similarity scores (< 0.25) and performs worse on those with higher similarity scores. Our explanation is that GPT-4 tends to take more obvious shortcuts by relying on only one or two entities in the easy instance, and “early exits” without even considering the entities that are closely relevant to the gold answer.

Open-Source Model Performance

To expand the scope of our findings, we experimented with open-sourced LLAMA-2 models across different sizes on hard questions of EUREQA and their results can be found at Table 4.3.

depth	d=1		d=2		d=3		d=4		d=5	
	<i>direct</i>	<i>icl</i>	<i>direct</i>	<i>icl</i>	<i>direct</i>	<i>icl</i>	<i>direct</i>	<i>icl</i>	<i>direct</i>	<i>icl</i>
Llama-2-7b	29.5	22.0	21.4	11.9	15.8	13.0	10.6	11.3	10.9	14.5
Llama-2-13b	40.0	28.5	27.0	16.4	20.5	15.8	11.3	15.9	13.9	19.4
Llama-2-70b	45.0	57.0	24.5	35.2	17.8	39.0	18.5	32.5	12.1	34.5

Table 4.3: Accuracy of Llama-2 models across different reasoning depths d on **hard** questions. The 7b and 13b versions are chat variants; the 70b version is the instruct variant.

Similar to our observations on GPT-series models, there’s a notable decline in the accuracy of Llama models as the reasoning depth increases from one to five on the hard set. We’ve also conducted the entity similarity analysis which led to the same observation as Observation 1 in Section 4.5: LLMs’ performance positively correlates with the entity similarities in the instances. These conclusions strengthened our claim that models rely on semantic shortcuts instead of following the correct reasoning path.

Will prompting solve EUREQA?

Although the effectiveness of prompting techniques is out of the scope of this paper, we still additionally tested the Tree of Thought(TOT)Yao et al. (2023) method on ChatGPT with a “propose” strategy, which tries to decompose the questions layer by layer and solve them sequentially. Our results show that such a prompting method fails completely even if we provide human-written examples for question decomposition. In no experiment did the TOT method generate a valid answer in the final response.

That being said, it is unlikely that there exists a prompting technique that will significantly improve models’ performance on EUREQA over CoT, as long as we do not enforce the correct reasoning chain (which defeats the purpose of testing generalizability) during inference.

Can Prompting Solve EUREQA?

We additionally tested the Tree of Thought (ToT) method (Yao et al., 2023) on ChatGPT with a “propose” strategy that attempts to decompose questions layer by layer. The ToT method

failed completely: in no experiment did it generate a valid answer. This is unsurprising, as no prompting technique is likely to significantly improve performance over chain-of-thought prompting on EUREQA unless it enforces the correct reasoning chain—which would defeat the purpose of testing generalizability.

Can Retrieval Augmentation Solve EUREQA?

Although our knowledge filtering process has already removed the knowledge barrier, it can still be pointed out that whether Retrieval Augmented Generation(RAG) method can solve this task. To address such concerns, we tested GPT-4 on 300 randomly sampled 5-layer hard questions. To minimize the impact of the retrieval method, we investigate a “performance upper bound” setting, which directly injects the retrieval result of the visible entities and relations in the input question from DBpedia. To be specific, statements like “**Storm Warning (1951 film) stars Actor D**” will have “<Storm Warning (1951 film), starring, ?>: Ginger Rogers, Steve Cochran, Ronald Reagan, Doris Day” prepended to the input question. The retrieved knowledge will obtain at most 20 candidate entities and the correct entity is guaranteed to be included. A complete example of a question and retrieved knowledge can be found in Figure 4.2. We believe this represents an upper bound of the retrieval result. Additionally, we provide each input with a 1-shot demonstration with the retrieved knowledge, input questions, and human-written reasoning thoughts.

The accuracy of GPT-4 through the above process is 62.0%, which is close to our reported performance of 61.9% in the original setting. We can then hypothesize that simply injecting knowledge into the model can hardly solve the problem and the bottleneck remains at the reasoning/question decomposition ability of LLMs. Moreover, it can be concluded that our knowledge-filtering process can effectively estimate parametric knowledge, and it is also reliable after self-consistency.

4.6 Conclusion and Transition

This chapter introduced EUREQA, a benchmark that probes LLM hallucinations induced by semantic shortcuts on reasoning chains. We presented a systematic method for generating

QA data with extended reasoning chains and deceptive semantic cues. Experimental results demonstrate that state-of-the-art LLMs predominantly depend on semantic shortcuts, with performance strongly correlated to entity similarity rather than reasoning depth. Human annotators, by contrast, solve these problems near-perfectly, confirming that the reasoning paths are simple and well-defined—the difficulty lies entirely in the model’s tendency to bypass genuine inference.

Looking ahead. Chapters 3 and 4 together establish that representational choices shape reasoning quality at two levels: the *output* level (how candidate answers are represented) and the *process* level (whether inference follows a grounded path or exploits distributional shortcuts). Both chapters operate in the text-only domain. A natural question is whether these representational challenges intensify or take new forms when models must reason jointly over text and images. The next two chapters explore this multimodal frontier. Chapter 5 addresses the first challenge—*which* visual information enters the model’s reasoning context—by introducing a framework that uses textual semantics to dynamically select relevant visual representations. Chapter 6 then tackles the deeper question of whether multimodal models can reason *directly* in the visual embedding space rather than funneling all reasoning through text. Together, they extend the representational lens of this dissertation from language into vision.

Chapter 5

SEMCLIP: Semantic-Guided Visual Selection for Vision-Language Models

5.1 Introduction

The previous two chapters established that representational choices have a profound impact on reasoning quality in the text-only domain—both at the output level (Chapter 3) and within the reasoning process itself (Chapter 4). This chapter extends the representational lens into the multimodal setting, where models must jointly reason over text and images. As discussed in Chapter 2, vision-language models (VLMs) face a fundamental tension between preserving fine-grained visual detail and maintaining computational efficiency. This chapter addresses a specific instance of this tension: *which* visual information should enter the model’s reasoning context? This chapter is based on the paper “Semantic-Clipping: Efficient Vision-Language Modeling with Semantic-Guided Visual Selection,” which is in submission and available as a preprint (Li et al., 2025a).

VLMs have demonstrated remarkable proficiency in vision-language understanding and reasoning. Most leverage aligned visual encoders to map input images into a fixed set of visual tokens, which are processed as prefix tokens by a large language model (LLM) alongside natural language tokens (Alayrac et al., 2022; Dai et al., 2023; Liu et al., 2023). When handling varied image resolutions, the visual encoding process typically involves downsizing the image to a fixed resolution and encoding patches of the low-resolution image into patch-

level tokens (Radford et al., 2021; Dosovitskiy et al., 2020). However, this approach often results in shape distortions, loss of fine-grained details, and a reduction in visual signals, hindering task performance for tasks requiring intricate visual comprehension (Liu et al., 2024b; Thapa et al., 2024).

Recent research has highlighted the benefits of higher-resolution encoders and multi-crop techniques that segment an image into multiple parts for processing at or near native resolution. While these methods preserve more visual details, they also generate significantly larger numbers of tokens, straining the context limits of LLMs, increasing computational costs, and potentially hindering reasoning performance.

Humans inherently retain task-related instructions while interacting with their environment. As illustrated in Figure 5.1, when searching for a chocolate cookie, one would first identify the jar containing chocolate cookies, focus on the relevant area, and then make a decision. Similarly, VLMs must prioritize image regions that are directly relevant to the given query in order to generate accurate responses.

To achieve this, we propose SEMCLIP, a novel sub-image selection method designed to enhance the fine-grained visual understanding of existing VLMs without requiring additional fine-tuning. SEMCLIP functions as a plug-and-play approach that improves high-resolution visual comprehension while minimizing the number of extra visual tokens needed. During inference, SEMCLIP first segments the original image into multiple candidate sub-images. It then evaluates the **relevance** of each sub-image to the given textual query, selecting the most relevant one to append to the original image before feeding it into the VLM. We explore various relevance-scoring strategies, which can leverage extracted features from VLM modules, pre-trained language-image models, or efficiently fine-tuned models that require minimal training on distantly supervised data.

Our proposed approach offers several merits. First, experimental results demonstrate that SEMCLIP significantly enhances the visual question answering (VQA) capabilities of widely used VLMs. Second, SEMCLIP achieves this improvement with minimal additional computational overhead, ensuring both performance gains and token efficiency. Analysis reveals that SEMCLIP attains comparable or superior performance on VQA benchmarks while requiring fewer computation resources compared to baseline methods. Third, SEMCLIP is a plug-in solution compatible with common VLMs, enabling high-resolution visual under-

standing without the need for model retraining. Finally, we provide experimental evidence showing that text-guided visual selection has a high theoretical upper bound, highlighting the substantial gap between the current visual understanding capabilities of VLMs and their full potential.

In summary, the key contributions of this chapter are: (1) We propose SEMCLIP, a plug-and-play approach that integrates textual signals into the visual encoding process, significantly improving the visual understanding capabilities of modern VLMs. (2) We demonstrate the effectiveness and computational efficiency of SEMCLIP through extensive experiments. (3) We train and publicly release a CLIP-based model optimized to enhance the performance of SEMCLIP. (4) We provide experimental evidence showing that text-guided visual selection has a remarkably high theoretical upper bound, revealing the substantial gap between current VLM capabilities and their full potential.

5.2 Method

We begin by outlining the fundamentals of vision-language models (Section 5.2), then describe the SEMCLIP pipeline (Section 5.2), with a focus on its core component—relevance measurement (Section 5.2)—and the training procedure for the distantly supervised scorer (Section 5.2).

Preliminaries

Modern VLMs consist of three key components: a vision encoder $g(\cdot)$, an LLM backbone $f_\phi(\cdot)$, and a multimodal projection module $w(\cdot)$ that bridges the two modalities. Given an input consisting of an image-question pair $(\mathbf{X}_v, \mathbf{X}_q)$, the vision encoder processes the image into visual features $g(\mathbf{X}_v) = \mathbf{Z}_v$, where $\mathbf{Z}_v$ is a fixed-length sequence of visual feature vectors. The textual question $\mathbf{X}_q$ is encoded into language features $\mathbf{H}_q$ through the embedding layers of the language model. Since the visual and textual features reside in different latent spaces, the multimodal projection module transforms the visual features into a representation compatible with the language model’s latent space: $w(\mathbf{Z}_v) = \mathbf{H}_v$. This projection module can take various forms, such as an MLP (Liu et al., 2024a; He et al., 2024) or a cross-attention



Figure 5.1: Paying attention to task-relevant regions within a scene is an intuitive approach to answering visual questions. Our objective is to identify the optimal task-relevance measurement ψ that selects the most pertinent sub-region of an image, enhancing visual understanding.

module (Alayrac et al., 2022). Finally, the aligned visual and textual features, $\mathbf{H}_v$ and $\mathbf{H}_q$, are fed into the language model to generate a response: $f_\phi(\mathbf{H}_v, \mathbf{H}_q)$.

Problem Definition

Current VLMs typically process images at a fixed resolution Liu et al. (2023); Chen et al. (2023). For high-resolution images, this mechanism requires either cropping or downscaling to

fit the model’s input restriction, often leading to the loss of fine-grained details or distortions of images. To mitigate this issue, recent methods Liu et al. (2024b) propose to segment high-resolution images into multiple patches, encoding each independently, and using all as inputs. Specifically, given an image-question pair $(\mathbf{X}_v, \mathbf{X}_q)$, the image $\mathbf{X}_v$ is divided into an $n \times n$ grid of smaller sub-images, denoted as $\{\mathbf{X}_{v_1}, \dots, \mathbf{X}_{v_{n \times n}}\}$. Each sub-image is then processed separately by the vision encoder $g(\cdot)$ and the projection module $w(\cdot)$, producing a set of visual token sequences $\{\mathbf{H}_{v_1}^q, \dots, \mathbf{H}_{v_{n \times n}}^q\}$. Finally, all sub-images, along with the original image and the question, are fed into the language model to generate a response:

$$f_\phi(\mathbf{H}_v, \mathbf{H}_{v_1}^q, \dots, \mathbf{H}_{v_{n \times n}}^q, \mathbf{H}_q).$$

While this approach preserves high-resolution details, it significantly increases the number of visual tokens, leading to higher computational costs.

To address these challenges, we formulate a visual token selection problem. We seek a subset of $k \ll n$ high-resolution sub-images that maximizes model performance as measured by evaluator $\mathcal{P}$:

$$S^* = \arg \max_{S \subseteq \{\mathbf{H}_{v_1}^q, \dots, \mathbf{H}_{v_{n \times n}}^q\}} \mathcal{P}(f_\phi(\mathbf{H}_v, S, \mathbf{H}_q)) \quad \text{s.t.} \quad |S| = k.$$

The parameter k serves as a hyperparameter depending on the resource limit. Since $\mathcal{P}$ is typically not accessible during inference, directly optimizing for S^* is infeasible, motivating the need for an efficient approximation.

Semantic Clipping

We propose a semantic-aware image clipping mechanism. In vision-language tasks such as VQA, not all image regions are pertinent to the given question. Including irrelevant visual tokens not only increases computational overhead but also risks diverting the language model’s attention from informative sub-images. To this end, we introduce a relevance scoring function ψ that evaluates the relevance of each sub-image $\mathbf{X}_{v_i}$ with respect to the question $\mathbf{X}_q$. Rather than processing visual features from all sub-images, we retain only the top- k most relevant:

$$S = \{\mathbf{X}_{v_i} \mid i \in \arg \text{top-k}(\psi(\mathbf{X}_q, \mathbf{X}_{v_1}), \dots, \psi(\mathbf{X}_q, \mathbf{X}_{v_{n \times n}}))\}.$$

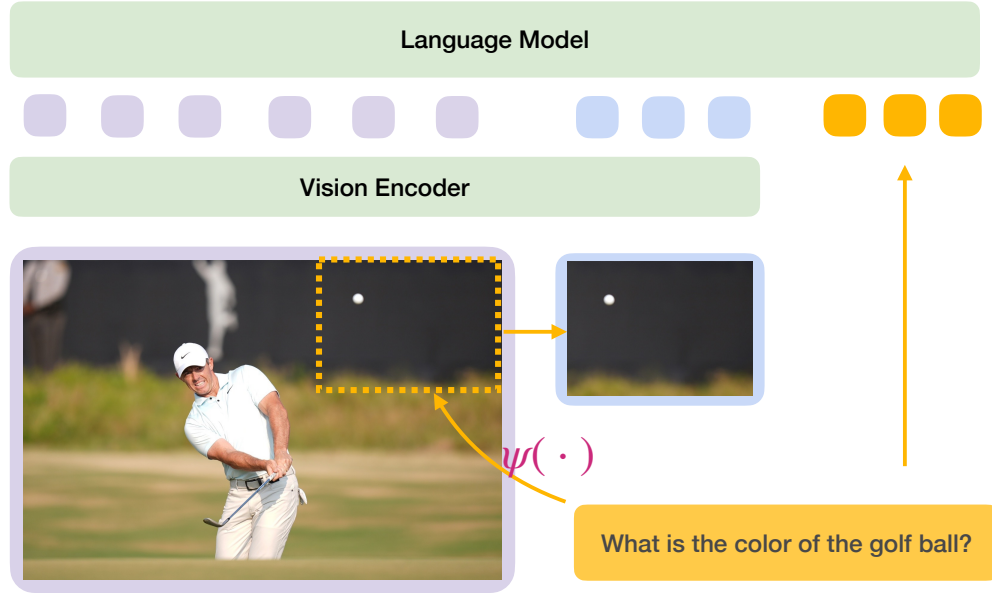


Figure 5.2: SEMCLIP is a plug-and-play method that enhances a VLM through semantic-guided visual selection. Task-relevant sub-areas of the image are encoded and appended to the visual tokens of the overview image (colored in blue and purple, respectively). These visual tokens, together with text embeddings of the question (colored in orange), are processed by the backbone VLM to generate a response.

This approach effectively reduces the number of visual tokens while ensuring that the retained visual features contribute meaningfully to the final response. The overall pipeline is illustrated in Figure 5.2.

Relevance Measurement

We explore multiple approaches to assess the relevance of a sub-image with respect to a given question. We first consider **semantic similarity-based methods**, which evaluate how closely a sub-image aligns with the question. The underlying assumption is that if a sub-image exhibits high similarity to the question, it is likely to contain relevant information that helps answer it. Within this category, we examine two alternatives: using the VLM itself as a semantic similarity measure, or utilizing an external pretrained similarity model. However, relevance may extend beyond semantic similarity alone. For instance, given the

question “*What should be used to cut a watermelon?*”, a sub-image containing a watermelon may exhibit a high semantic similarity, yet it does not provide the necessary information to answer the question. Instead, the model should prioritize attending to a sub-image of a knife, as it directly pertains to the action described in the query. To capture possibly complex relationships between a question and its corresponding sub-images, we propose a **distant supervision approach** that directly trains a relevance model using distantly labeled data. We provide a detailed discussion of each method in the remaining of the section.

VLM-based similarity (ψ_{lm}). This approach utilizes the VLM itself to measure similarity between a sub-image and the question. Each sub-image is processed through the vision encoder and projection module, and the projected classification visual token embedding is extracted as the sub-image representation. The LLM backbone encodes the question, with the hidden state of the last token in the final layer serving as its textual representation. The relevance score ψ_{lm} is computed as the cosine similarity between these features. This method requires no additional training and leverages the modality adaptation capabilities of the projection module $w(\cdot)$.

Pre-trained similarity model (ψ_{siglip}). This approach leverages SigLIP (Zhai et al., 2023), a state-of-the-art pre-trained bi-encoder model designed to measure text-image similarity. Trained via contrastive learning, SigLIP is applied directly, without modifications, to compute similarity scores used as relevance scores to rank sub-images.

Distant supervision (ψ_{clip}). We also introduce ψ_{clip} , which requires minimal training on distant supervision data while achieving the best performance among the proposed ψ methods for SEMCLIP. This approach fine-tunes a CLIP model to evaluate the relevance between an image and the given question. For our primary study, we use LLaVA-1.5-7B as the base model. Since LLaVA’s supervised fine-tuning process incorporates the ScienceQA dataset—a multiple-choice VQA benchmark — we select it as the target dataset for training.

Training ψ with Distant Supervision

We create synthetic training data for ψ based on VQA data. For each question-image pair $(\mathbf{X}_v, \mathbf{X}_q)$, we partition the high-resolution image into several non-overlapping sub-images $\mathbf{X}_{v_i}$ with corresponding visual tokens $\mathbf{H}_{v_i}$. Each sub-image that leads to the correct answer is labeled as a positive instance, and each that does not as a negative instance. Since the original VQA training data are annotated with ground-truth answers, we have the evaluator $\mathcal{P}$, and $\mathcal{P}(f_\phi(\mathbf{H}_v, \mathbf{H}_{v_i}, \mathbf{H}_q))$ returns the answer correctness given the low-resolution image, the selected high-resolution sub-image, and the question. The positive sub-images are defined as:

$$\mathbf{X}_{v_+} \in \{\mathbf{X}_{v_i} \mid \mathcal{P}(f_\phi(\mathbf{H}_v, \mathbf{H}_{v_i}, \mathbf{H}_q)) = 1\},$$

while the negative sub-images are:

$$\mathbf{X}_{v_-} \in \{\mathbf{X}_{v_i} \mid \mathcal{P}(f_\phi(\mathbf{H}_v, \mathbf{H}_{v_i}, \mathbf{H}_q)) = 0\}.$$

This automatic labeling strategy yields approximately 10k question-image pairs from the ScienceQA training split, containing both positive and negative sub-images. Using this dataset, we fine-tune a CLIP model (`openai/clip-vit-large-patch14`) with a contrastive learning approach based on margin ranking loss:

$$\mathcal{L} = \sum_{\mathbf{X}_q, \mathbf{X}_{v_+}, \mathbf{X}_{v_-}} \max(0, m + \psi(\mathbf{X}_q, \mathbf{X}_{v_-}) - \psi(\mathbf{X}_q, \mathbf{X}_{v_+})),$$

where m is the margin. This training process ensures that text-image pairs with positive sub-images receive higher relevance scores than those with negative ones. Notably, the training data is sourced from the model’s original dataset, mitigating potential data leakage concerns.

5.3 Experiments

We present model configurations, evaluation methods, and baseline approaches (Section 5.3), then analyze the effectiveness of different relevance measurements (Section 5.3), the performance improvements over the original model (Section 5.3), the theoretical upper bound (Section 5.3), and additional ablation studies (Sections 5.3–5.3).

Experiment Settings

We use LLaVA-1.5 (Liu et al., 2024a) as the base VLM, with Vicuna 7B as the LLM backbone. Since SEMCLIP functions as an inference-time method, no additional training is performed on the VLM. For inference, we set temperature $\tau = 0$ for reproducibility. To train ψ_{clip} , we use ViT-L-14 as the backbone, fine-tuning for 32 epochs on the constructed training set with learning rate 5e-6 and batch size 64. The model with the lowest training loss is selected.

Evaluation. We evaluate on the following benchmarks: V^* (Wu and Xie, 2023) for detailed understanding, and POPE (Li et al., 2023c), GQA (Hudson and Manning, 2019), ScienceQA (Lu et al., 2022), MMBench (Liu et al., 2024c), MMStar (Chen et al., 2024a), and SeedBench (Li et al., 2023b) for visual question answering. All results are reported as accuracy.

Baselines. We compare against two state-of-the-art baselines built on the LLaVA architecture. Matryoshka Multimodal Model (M^3 , Cai et al. (2024)) dynamically compresses visual token length during training, allowing control over visual granularity. Scaling on Scales (S^2 , Shi et al. (2025)) processes multiple cropped versions, stacks their visual features into a higher-dimensional representation, and projects them back into the textual feature space. Both baselines require retraining LLaVA on visual instruction data, whereas SEMCLIP operates without additional VLM training. We use LLaVA-1.5-Vicuna-7B as the base VLM for all methods and employ full-length visual tokens for M^3 .

Distant Supervision Yields the Best Performance

The evaluation results of SEMCLIP with three different ψ methods are presented in Table 5.1. Both ψ_{siglip} and ψ_{clip} outperform the base VLM by 1.0% and 3.3%, respectively, with ψ_{clip} achieving the highest average accuracy of 63.2%. This suggests that ψ_{clip} effectively captures task-specific signals through distant supervision, allowing it to extract the most relevant visual information from the image. At inference, the selected query-related sub-image is encoded as additional visual input tokens, providing more precise guidance for the language model. However, SEMCLIP with ψ_{lm} performs comparably to or worse than the original

model on most benchmarks. We attribute this to the fact that the next-token embedding falls short of capturing contextual cues within the textual question, limiting its effectiveness in selecting the correct sub-image.

SEMCLIP Improves Visual Understanding

As shown in Table 5.1, SEMCLIP shows better performance compared to other baselines. Among all methods, SEMCLIP with ψ_{clip} achieves the highest scores on four benchmarks, particularly excelling in MMBench with a 7.9% accuracy boost. Additionally, ψ_{siglip} achieves the best performance on V^* , the detailed understanding benchmark. As demonstrated by Table 5.2, ψ_{siglip} even surpasses LLaVA-NeXT-7B—a model that requires additional instruction-tuning data and uses 2.5 times more visual tokens—by 0.6%. Furthermore, SEMCLIP with ψ_{clip} improves the model’s average accuracy by 3.3%, outperforming both M^3 (0.2%) and S^2 (1.4%), demonstrating its superior effectiveness in improving visual comprehension.

SEMCLIP also offers significant efficiency advantages. Unlike M^3 and S^2 , both of which require extensive retraining, SEMCLIP is an inference-time method that operates with minimal or no additional training. Specifically, ψ_{clip} is trained on approximately 3,000 data instances in under an hour using a single NVIDIA RTX 6000 ADA (48GB) GPU. In contrast, S^2 requires at least four hours to retrain the multimodal projection layer, while M^3 requires a minimum of ten hours for full instruction tuning on an 8×NVIDIA A100 (80GB) setup.¹

These disparities in training efficiency and benchmark performance suggest that valuable signals from VQA training data are still underutilized. The textual signal from the questions can offer significant guidance to the visual encoding process, ultimately enhancing end-task performance. This also prompts an important question that we seek to discuss next: What is the theoretical performance potential of SEMCLIP?

Where Is the Upper Bound of SEMCLIP?

To explore the potential performance ceiling of SEMCLIP, we conduct an *Optimal Selection* experiment. Specifically, for a given question-image pair, the image is divided into an $n \times n$

¹Time estimates are based on the training specifications reported by each baseline.

Model	Avg.	V^*	GQA	SQA	POPE	MMBench	MMStar	SeedBench
<i>Baseline Methods</i>								
LLaVA-1.5-7B	59.9	47.6	62.0	65.8	86.2	64.7	33.3	60.5
M^3	60.1	44.0	61.9	65.5	87.4	65.9	35.6	60.6
S^2	61.3	49.7	59.9	66.2	88.3	63.7	33.5	68.0
<i>SEMCLIP</i>								
ψ_{lm}	59.2	49.2	59.9	65.6	84.6	60.5	31.8	63.0
ψ_{siglip}	60.9	57.6	60.3	64.5	86.3	61.6	32.1	63.7
ψ_{clip}	63.2	52.9	62.5	68.3	89.0	72.6	33.5	63.7

Table 5.1: SEMCLIP achieves the highest performance improvements across most tasks, surpassing all baselines in average accuracy. While baseline methods require model-specific fine-tuning, SEMCLIP operates as a plug-and-play approach with minimal training on the top-performing relevance scorer, ψ_{clip} .

grid² of sub-images. Each sub-image is then paired with the original image and input into the VLM along with the textual question to generate a response. This process yields n^2 responses per instance. If any of these responses match the correct answer, the question is considered “answerable” by SEMCLIP using optimal sub-images. We hypothesize the existence of an optimal relevance measurement, $\psi_{optimal}$, capable of selecting the best sub-image for the VLM to correctly answer all “answerable” questions. The performance of $\psi_{optimal}$ on each benchmark serves as the theoretical upper bound for SEMCLIP.

As results showed in Table 5.2, $\psi_{optimal}$ improves the original LLaVA-1.5-7B by 15.4% on average, which is much higher than LLaVA-NeXT-7B—the same-sized model using more training data and more visual token inputs, and LLaVA-1.5-13B that produced with the same training data but a much larger language model backbone. These observations suggest the following conclusions:

(1) Expanding the language model size may not be an efficient way to enhance visual understanding. LLaVA-1.5-13B follows the same training approach and dataset as LLaVA-1.5-7B but achieves only a modest 3% average accuracy improvement—falling short of both the 3.3% performance gain achieved by SEMCLIP and the theoretical 15.4% improvement predicted by $\psi_{optimal}$.

²For the reported results, $n = 3$.

This advantage of SEMCLIP can be attributed to its strong performance on benchmarks like V^* and MMBench. In these datasets, images can be as large as 1700×1500 pixels, and downscaling them to 336×336 for vision encoding results in a significant loss of detail. In such cases, the vision encoder of LLaVA-1.5-13B becomes a bottleneck, whereas SEMCLIP enhances performance by selecting the most task-relevant sub-image and providing additional detailed explanations to the VLM.

(2) Greedily encoding all sub-areas of an image provides limited benefits. The primary difference between SEMCLIP and LLaVA-NeXT-7B lies in their vision encoding strategies. SEMCLIP selectively focuses on the most query-relevant sub-areas. On the contrary, LLaVA-NeXT takes a more aggressive approach by dividing the image into a 2×2 grid, encoding each section separately, and appending the additional visual tokens to the encoded overview tokens. The technique of SEMCLIP is shown to be particularly effective for tasks requiring detailed comprehension, such as V^* , where it achieves a 9.5% higher accuracy than LLaVA-1.5-7B with the same model size.

By generating extra visual tokens, the crop-and-encode method enables higher-resolution analysis. However, its effectiveness diminishes for tasks that do not demand fine-grained details, especially with smaller images. For example, on POPE, an object detection benchmark, LLaVA-NeXT achieves only a 1.2% accuracy improvement, as its images (approximately 333×600 pixels on average) already contain sufficient information for object presence detection, rendering the additional visual tokens redundant. A similar trend is observed in SQA, a scientific question-answering dataset, where only an overview image is necessary to answer the questions—resulting in degraded performance due to the extra visual tokens.

(3) The upper bound grows with more pre-training or larger model size. We also conduct $\psi_{optimal}$ experiments on both of the other LLaVA variants for comparison. In this setup, the cropped sub-images encoded by LLaVA-NeXT are replaced with a single sub-image. As shown in the bottom section of Table 5.2, there remains a significant gap between the models’ actual performance and their theoretical upper bound.

Our observations indicate that, with expanded training data and an extended visual token length during training, the upper bound of the 7B-parameter LLaVA model increases by an

Model	Avg.	V^*	GQA	SQA	POPE	MMBench	MMStar	SeedBench
<i>Ablation Study</i>								
Sub-img only	53.3	42.9	51.9	68.0	70.6	52.5	31.2	56.1
Majority Vote	63.3	49.2	61.4	69.5	89.2	74.2	35.4	64.5
Random	60.3	49.5	60.6	65.1	86.1	64.6	32.7	63.7
$\psi_{optimal}$	76.0	79.6	72.4	76.9	95.7	80.0	49.7	77.9
LLaVA-NeXT-7B	63.8	57.1	64.1	65.7	87.4	66.7	35.5	70.0
+ $\psi_{optimal}$	80.9	72.8	72.4	89.3	94.8	88.7	65.9	82.1
LLaVA-1.5-13B	62.9	48.7	63.4	68.8	88.4	68.6	34.5	68.0
+ $\psi_{optimal}$	78.6	82.2	75.6	78.7	96.5	83.8	54.3	79.3

Table 5.2: Theoretical upper bound experiments reveal that modern VLMs have significant potential for improvement in visual understanding when provided with appropriate sub-image guidance. Ablation studies confirm that SEMCLIP’s effectiveness is not merely due to extra visual tokens, and that sub-images alone are insufficient for comprehensive understanding.

average of 4.9%. This suggests that the upper bound of SEMCLIP is not static—it **grows as the base VLM undergoes more training with additional data**, even when the model size remains unchanged.

Similarly, LLaVA-1.5-13B sees an average improvement of 2.6%, with a substantial 33.5% performance boost on V^* , a detail-oriented understanding task. As discussed earlier, a larger backbone LLM benefits from a higher-dimensional latent space, which enhances modality projection. $\psi_{optimal}$ helps overcome the vision encoder’s resolution bottleneck, making the 13B model outperform its 7B counterpart across all tasks. However, its theoretical upper bound is still 2.3% lower than that of LLaVA-NeXT-7B.

The discrepancy of upper bounds highlights a key insight: while LLaVA-1.5-13B improves performance by increasing the language model size, LLaVA-NeXT-7B benefits from longer visual token sequences and more training data. From this, we hypothesize that **enhancing visual understanding in VLMs is more efficiently achieved by training with more data and refining visual encoding strategies, rather than merely increasing language model size.**

How Are Useful Sub-Images Distributed?

A key follow-up question regarding the theoretical upper bound is how accessible and densely distributed the sub-images contributing to a correct answer are. To illustrate this “density,” we conduct a “Majority Vote” experiment, where each of the n^2 sub-images is paired with the overview image to generate a response. The final answer is determined by selecting the most frequently occurring response. The results of this experiment are presented in Table 5.2.

Overall, “Majority Vote” improves performance by 3.4% on average compared to LLaVA-1.5-7B, with only a 0.1% difference from ψ_{clip} . Notably, it performs better on general VQA tasks, suggesting that in most cases, task-relevant sub-images dominate the image crops. This inherent signal enhances VLM performance but remains underutilized in current models. However, “Majority Vote” underperforms compared to the trained ψ_{clip} on the detailed understanding benchmark, V^* , highlighting the necessity and effectiveness of selection training for high-resolution detail comprehension.

A Trade-Off Between Image Token Numbers and Performance

We analyze the trade-off between efficiency and performance across different methods. Based on experiments varying the number k of top-relevant sub-images used by SEMCLIP, the best performance is consistently achieved at $k = 1$, effectively doubling the number of visual tokens fed into the VLM. We present a comparison of the number of processed image tokens, total VLM parameters, and overall performance on visual understanding tasks, as detailed in Table 5.3. Compared to the best-performing LLaVA-NeXT, SEMCLIP-equipped LLaVA-1.5 achieves similar results with the same model size while requiring 60% fewer visual tokens. Notably, SEMCLIP is a plug-and-play method that does not require any additional training for the VLM. Furthermore, compared to LLaVA-1.5-13B, which uses nearly twice the language model size, SEMCLIP outperforms it, particularly on fine-grained detail understanding tasks. We conclude that SEMCLIP effectively encodes fine-grained visual information by selectively pruning visual tokens based on the semantic relevance of the task query.

Method	#img tokens	#Params.	Avg. Acc
LLaVA-1.5	576	7B	59.9
M^3	576	7B	60.1
S^2	576	7B	61.3
LLaVA-1.5	576	13B	62.9
SEMCLIP	576×2	7B	63.2
LLaVA-NeXT	576×5	7B	63.8

Table 5.3: SEMCLIP delivers performance on par with the best baseline while requiring significantly less training and 60% fewer visual tokens, achieving the optimal balance between training cost, inference efficiency, and accuracy. SEMCLIP outperforms a VLM with a substantially larger language model, highlighting the superiority of task-guided visual selection over simply increasing language model size.

Is the Sub-Image Itself Enough?

To highlight the importance of pairing the overview image, we conduct a *Sub-img only* experiment, removing visual tokens from the original image and providing only the encoded sub-image to the language model. On average, performance drops by 6.6% (Table 5.2), indicating that the overview image remains essential to the effectiveness of SEMCLIP.

Extra Visual Tokens Do Not Guarantee Improvement

A potential concern is whether SEMCLIP’s performance improvement stems from the additional computation required by extra visual tokens rather than from the selection mechanism itself. To address this, we conduct an experiment where a randomly selected sub-image is paired with the original image (“Random” in Table 5.2). We repeat this process 32 times and report the average.

The Random approach slightly outperforms the original LLaVA-1.5-7B, with clear improvements of 1.9% on V^* and 3.2% on SeedBench. The V^* improvement is likely due to the vision encoder’s resolution bottleneck—a randomly chosen sub-image may carry task-relevant zoomed-in details. The SeedBench gain primarily comes from the “instance attribution” sub-task. However, the Random approach is substantially worse than ψ_{clip} (60.3% vs. 63.2% average), confirming that SEMCLIP’s improvement is driven by *semantic selection*, not merely

extra tokens.

5.4 Conclusion and Transition

This chapter introduced SEMCLIP, a plug-and-play approach that enhances VLMs’ visual understanding through semantic-guided visual selection. Integrating SEMCLIP significantly improves VLMs’ ability to comprehend fine-grained visual details while requiring minimal training for the relevance scorer. The theoretical upper bound analysis reveals remarkably high potential, with performance improving further as the base VLM is trained on more data. These results demonstrate that valuable signals from the textual query remain underutilized in current VLM architectures, and that intelligent visual selection is a high-leverage avenue for improvement.

Looking ahead. SEMCLIP addresses the question of *which* visual information should enter the model’s reasoning context by selecting task-relevant sub-images at the input level. However, even after the most relevant visual tokens have been selected and projected into the shared embedding space, a fundamental question remains: can the language model reason *directly* over these visual representations, or must it first translate them into text? Current multimodal reasoning paradigms—whether “thinking about images” (text-only chain-of-thought) or “thinking with images” (external visual tools)—ultimately route visual understanding through textual channels. The next chapter proposes a radically different approach: Latent Visual Reasoning (LVR), in which the language model generates latent states that reconstruct question-relevant visual features directly in the embedding space, enabling seamless cross-modal reasoning without the bottleneck of text-mediated inference.

Chapter 6

Latent Visual Reasoning

6.1 Introduction

The previous chapter introduced SEMCLIP, a framework that addresses the question of *which* visual information should enter the model’s reasoning context by selecting task-relevant sub-images at the input level. This chapter takes the next step: rather than improving visual selection before reasoning begins, we ask whether multimodal models can reason *directly* in the visual embedding space, bypassing the bottleneck of text-mediated inference altogether. The content of this chapter is adapted from "Latent Visual Reasoning," published in the Proceedings of the Fourteenth International Conference on Learning Representations (ICLR 2026) (Li et al., 2026b).

Multimodal large language models (MLLMs) (Li et al., 2024c; Bai et al., 2025c; Wang et al., 2025d) have shown remarkable capability in jointly understanding visual and textual content. By leveraging the generative capabilities of their backbone large language models (LLMs), MLLMs extend the expressiveness of visual encoders beyond simple perception tasks. This advancement has enabled the integration of chain-of-thought (CoT) reasoning into MLLMs, allowing them to perform structured textual reasoning in response to complex multimodal queries. In this paradigm, the LLM decomposes a query into intermediate steps and resolves each step while conditioning on static visual inputs. This approach, referred to as “Thinking about Images” by (Su et al., 2025c), has proven effective across diverse domains, including scientific visual question answering (Zhang et al., 2023b), mathematics (Huang

et al., 2025a), and visual grounding (Bai et al., 2025d).

Further research has expanded the multimodal reasoning workspace by enabling active editing of input images alongside textual reasoning trajectories. Such editing includes drawing auxiliary lines (Hu et al., 2024), zooming in (Shao et al., 2024a; Su et al., 2025a; Surís et al., 2023), shifting image styles (Liu et al., 2025a), highlighting sub-regions (Fu et al., 2025c), and more. During intermediate CoT steps, methods in this paradigm either call external tools or generate programs to manipulate images, re-encode the edited outputs, and inject the new image tokens as visual enhancements into subsequent textual reasoning. In this way, the salient visual information is actively incorporated throughout the reasoning process. These methods are commonly termed “Thinking with images.”

Both “Thinking about Images” and “Thinking with Images” aim to address a core limitation in current MLLMs: despite sophisticated visual encoders projecting visual information into text spaces, backbone LLMs often fail to capture the visual details most relevant to the text query. This shortcoming arises from factors such as modality projection bias (Zhang et al., 2024; Liu et al., 2023), modality interference (Cai et al., 2025; Wang et al., 2025e; Pezeshkpour et al., 2025; Deng et al., 2025a), and cross-modality attention bias (Zhang et al., 2025c). “Thinking about Images” addresses these issues by generating additional task-relevant text tokens in the context window, thereby increasing the likelihood of correct answers. However, excessive token generation may cause the textual context to dominate, overshadowing essential visual inputs (Huang et al., 2024). In contrast, “Thinking with Images” leverages external tools to inject visual information during text generation, calibrating the alignment between generated text and the original visual input. Yet these approaches often bypass the newly injected sub-images due to training data bias, or remain constrained by the predefined operations of the tools (Su et al., 2025a). In short, both categories primarily refine text generation to improve cross-modality understanding, yet a fundamental gap persists between visual inputs and text generation in producing the final answer.

A parallel line of research explores omni-modality foundation models that accept both text and visual inputs and generate outputs in both modalities simultaneously (Team, 2024; Deng et al., 2025b; Xie et al., 2025). Some subsequent works attempt to exploit image generation capabilities for multimodal reasoning, but effectiveness has so far been demonstrated only in specific downstream tasks such as navigation and maze solving (Li et al., 2025b; Xu et al.,

2025). Moreover, it remains unclear whether visual inputs, once decoded and re-encoded by such models, can still faithfully preserve the original information.

Rethinking the input of MLLMs, where continuous visual tokens and discrete text tokens are embedded in a shared latent semantic space, we ask the following question:

If visual and textual tokens are embedded in a joint semantic space within an MLLM, why not reason over both jointly as well?

It is both natural and efficient to extend reasoning beyond discrete text tokens to include visual tokens that directly encode visual information. However, conventional LLMs are limited to operating on discrete tokens due to their next-token prediction training objective. To address this, Hao et al. proposed passing last hidden states rather than text tokens, enabling more efficient expression of complex thoughts through latent reasoning. Building on this idea, we introduce LATENT VISUAL REASONING (LVR), a novel paradigm for multimodal reasoning (see Fig. 6.1). Our approach involves a simple yet fundamental modification to the conventional Vision-Projector-LLM structure: the LLM is able to perform hybrid reasoning that alternates between LVR and standard text generation. In the LVR phase, the LLM leverages the last hidden state to approximate the question-relevant visual tokens within the visual inputs. During the text generation phase, the model predicts the next text token in sequence. Both phases operate in an auto-regressive manner.

We propose a two-stage training pipeline for LVR. The first stage is supervised fine-tuning (SFT), which jointly optimizes LVR’s internal processes alongside next-token prediction for text generation. The second stage applies reinforcement learning (RL), allowing LVR to self-evolve the latent reasoning process while receiving policy rewards from generated text, thereby encouraging a more unified semantic space. Specifically, we adapt the GRPO algorithm (Shao et al., 2024c) to replay latent reasoning steps during policy gradient loss computation. In addition, GRPO leverages verifiable rewards to evaluate roll-out responses, where the policy gradient loss is computed solely from the token distribution of the text generation component. Experimental results demonstrate that LVR achieves substantial improvements over state-of-the-art MLLMs, particularly on perception-intensive and visual detail-dependent understanding tasks.

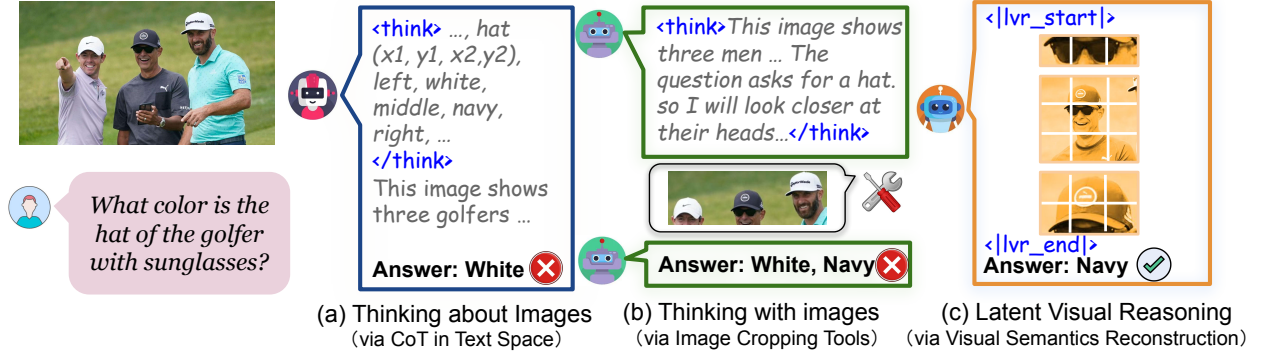


Figure 6.1: **Conceptual illustration of LATENT VISUAL REASONING (LVR).** We compare LVR with two paradigms: “Think about images,” which performs multimodal reasoning entirely in text space, and “Think with images,” which leverages external visual tools to highlight regions of interest (ROIs). In contrast, LVR leverages the LLM’s latent space to reconstruct the semantics of ROIs, enabling seamless cross-modal reasoning.

In summary, the contributions of this chapter are as follows:

- We propose LATENT VISUAL REASONING, a novel multimodal reasoning paradigm that unifies latent reasoning over visual inputs with text generation in the language space, enabling deeper integration of visual and textual signals throughout the model’s reasoning process.
- We introduce architectural innovations and training frameworks for stable and scalable training of MLLMs with LVR. Our approach combines a reconstruction loss with next-token prediction for SFT and extends the GRPO algorithm to latent reasoning for reinforcement learning.
- Through extensive evaluation, we demonstrate that LVR achieves strong performance across diverse visual question answering benchmarks requiring fine-grained visual understanding and perception. Our ablation studies explore alternative architectural designs and training objectives, providing insights for future research.

6.2 Method

This section introduces LATENT VISUAL REASONING, a new paradigm that enables MLLMs to reason jointly over text and visual tokens. The overall inference pipeline is illustrated in Figure 6.2. At a high level, LVR is trained to reconstruct visual semantics relevant to both the input image and the accompanying text query. These reconstructed semantics, termed *latent visual thoughts*, are combined with the original inputs to guide the generation of textual responses.

We begin with an overview of the LVR architecture, built upon the Qwen-2.5-VL series (Bai et al., 2025a) (Section 6.2). Next, we present a two-stage training pipeline—combining supervised fine-tuning and reinforcement learning—that jointly teaches the model to reconstruct visual semantics and generate text (Section 6.2). Finally, we introduce decoding strategies that allow the model to flexibly alternate between LATENT VISUAL REASONING and standard text generation during inference (Section 6.2).

Method Overview

Figure 6.2 illustrates the architecture of LVR, which largely follows the standard MLLM design. It consists of three key components: a vision encoder $vision(\cdot)$, an LLM backbone $\theta(\cdot)$, and a multimodal projector $proj(\cdot)$ that aligns the two modalities. Given an input image-question pair $(\mathbf{X}_v, \mathbf{X}_t)$, the vision encoder transforms the image into visual features $\mathbf{V} = vision(\mathbf{X}_v)$. In parallel, the textual question $\mathbf{X}_t$ is embedded into language features $\mathbf{T}$ by the LLM’s embedding layers. Since visual and textual features reside in distinct latent spaces, the projector $proj(\cdot)$ maps the visual features into a representation aligned with the language model’s latent space: $\mathbf{V}_T = proj(\mathbf{V})$.

In standard MLLMs, both $\mathbf{V}_T$ and $\mathbf{T}$ are passed into the LLM $\theta(\cdot)$. However, the decoding process remains strictly text-centric: hidden states are computed auto-regressively and mapped, through the language modeling head, to discrete tokens in the vocabulary. Thus, while visual information can guide the reasoning process, the output space is constrained to text tokens, limiting the model’s ability to directly reason over visual semantics.

To overcome this limitation, we propose LVR, which extends the conventional text-only

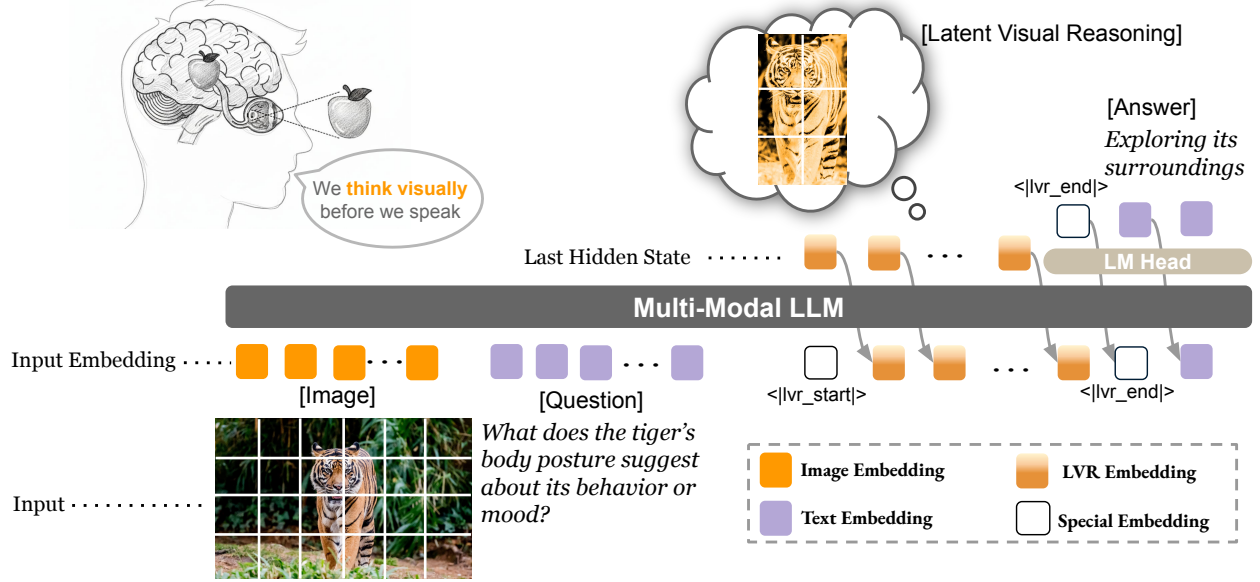


Figure 6.2: **Training and inference pipeline of LVR.** Images are encoded into tokens by a visual encoder and mapped into a joint semantic space with text embeddings. During the SFT stage, bounding boxes identify query-relevant visual tokens, which supervise the last hidden states in the LVR process via MSE loss. The LVR process is wrapped with special tokens indicating reasoning mode. Once all query-relevant visual tokens are consumed, the model exits LVR and resumes standard text generation with cross-entropy loss. During RL training, the model self-evolves the LVR process learned in SFT using $\text{GRPO}_{\text{latent}}$. At inference, the model triggers LVR upon generating the special token, propagates hidden states to reconstruct visual semantics, and resumes text generation when a stopping criterion is met.

generation paradigm by enabling interleaved LATENT VISUAL REASONING. When the special token $\langle \text{lvr_start} \rangle$ is generated, the model enters a latent reasoning mode, where it reconstructs visual semantics in the space of $\mathbf{V}_T$ to better support answering the textual query. The hidden states produced in this mode are directly propagated as input embeddings to subsequent positions until a stopping criterion is reached. At that point, the model generates the token $\langle \text{lvr_end} \rangle$ and resumes standard text generation. The hidden-state sequence produced during the LVR segment can be regarded as an analogue of a human’s “visual thinking,” where query-relevant visual semantics are mentally reconstructed to enhance the precision of reasoning and answering.

Two-Stage Training Pipeline

Supervised Fine-Tuning

We begin with supervised fine-tuning (SFT) to instill the basic pattern of latent visual reasoning. During this stage, we explicitly supervise the reasoning content by forcing the MLLM to use its latent space embeddings (i.e., last hidden states) to reconstruct the ground-truth regions of interest for each image–text pair. Since the reconstruction loss directly dictates what the model must reason about, the reasoning process is constrained rather than free-form. We therefore characterize this stage as a teacher-forcing paradigm for latent visual reasoning: although restrictive, it allows the model to quickly acquire the fundamental ability to reason in the latent space.

Each SFT instance consists of an image-question pair, accompanied by a pre-annotated bounding box of a region of interest (ROI) relevant to the query. For a given image, the model’s image processor first crops it into a grid of visual patches. Based on the ROI bounding box, LVR efficiently selects the corresponding patches and retrieves their indices $\mathbf{I} = \{I_1, I_2, I_3, \dots, I_{T_v}\}$ from the sequence of flattened visual patches in $O(1)$ time. During the forward pass, the image is encoded into a sequence of visual embeddings in the semantic space, followed by the embeddings of the text tokens. A subset of visual embeddings $\mathbf{v} = \{v_1, v_2, \dots, v_{T_v}\}$ is then gathered using the index list $\mathbf{I}$, which—by design of the ViT encoder—directly corresponds to the visual patches within the ROI. These selected tokens are enclosed by the special tokens $< |\text{lvr_start}| >$ and $< |\text{lvr_end}| >$, thereby defining the LATENT VISUAL REASONING process. Finally, the remaining textual response is appended to the sequence for generation.

We train the model with two joint learning objectives that explicitly couple latent visual reasoning with downstream text generation.

Visual Reconstruction Loss. During latent reasoning, the model predicts a sequence of final hidden states $\{\mathbf{h}_t\}_{t=1}^{T_v}$ that are expected to encode the underlying visual semantics. We enforce these hidden states to approximate the ground-truth visual embeddings $\{\mathbf{v}_t\}_{t=1}^{T_v}$ via a

mean squared error (MSE) objective:

$$\mathcal{L}_{\text{LVR}} = \frac{1}{T_v} \sum_{t=1}^{T_v} \|\mathbf{h}_t - \mathbf{v}_t\|_2^2. \quad (6.1)$$

Next-Token Prediction (NTP) Loss. For the subsequent language modeling phase, the model generates the final response tokens $\{y_t\}_{t=1}^{T_y}$ to answer the visual question. We adopt the standard cross-entropy loss:

$$\mathcal{L}_{\text{NTP}} = -\frac{1}{T_y} \sum_{t=1}^{T_y} \log p_{\theta}(y_t \mid y_{<t}, \mathbf{h}_{1:T_v}). \quad (6.2)$$

Joint Objective. The overall training loss is a weighted sum:

$$\mathcal{L} = \mathcal{L}_{\text{NTP}} + \lambda_{\text{LVR}} \cdot \mathcal{L}_{\text{LVR}}, \quad (6.3)$$

where λ_{LVR} is a hyperparameter balancing reconstruction and generation signals.

Reinforcement Learning

We employ GRPO to further refine the interaction between LVR and standard text generation. In this stage, rewards are computed solely based on output accuracy and adherence to the required response format. Unlike supervised fine-tuning, no constraints are imposed on the intermediate outputs of LVR, allowing the model to freely explore the latent visual reasoning space. This relaxation also removes the need for pre-annotated ROI bounding boxes, as GRPO training can be performed directly on image-text pairs. Moreover, GRPO implicitly encourages the generation of the $\langle \text{lv_start} \rangle$ token, thereby increasing the likelihood of activating the LVR process during response generation.

However, directly applying the standard GRPO algorithm to LVR is non-trivial. The key challenge lies in the mismatch between the policy-gradient loss, which is defined over token distributions, and the latent reasoning process, which lacks an explicit token distribution. **To address this issue, we propose $\text{GRPO}_{\text{latent}}$, a variant that can be seamlessly applied to any latent-reasoning model.**

The objective is:

$$J_{\text{GRPO}_{\text{latent}}}(\theta) = \mathbb{E}_{q, I, o \sim \pi_{\theta_{\text{old}}}} \left[\frac{1}{|y|} \sum_{t=1}^{|y|} \min \left(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1-\varepsilon, 1+\varepsilon) \hat{A}_t \right) - \beta D_{\text{KL}}(\pi_{\theta}(\cdot | q, I) \parallel \pi_{\text{ref}}(\cdot | q, I)) \right]. \quad (6.4)$$

Here, the token-wise importance ratio $r_{i,t}(\theta)$ for a text token $y_{i,t}$ is computed using a teacher-forcing log-probability pass, where **the latent reasoning hidden states are replayed from the original rollout**:

$$r_{i,t}(\theta) = \frac{\pi_{\theta}(y_{i,t} | q, I, \tilde{h}_i^{\text{latent}}, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t} | q, I, \tilde{h}_i^{\text{latent}}, y_{i,<t})}. \quad (6.5)$$

During rollout we record the last hidden states of LVR processes:

$$\tilde{h}_i^{\text{latent}} = \{h_{i,1}^{\text{latent}}, \dots, h_{i,L}^{\text{latent}}\}.$$

To evaluate importance ratios, we perform a teacher-forcing forward pass under both π_{θ} and $\pi_{\theta_{\text{old}}}$. The recorded hidden states $\tilde{h}_i^{\text{latent}}$ are patched into the latent-reasoning positions of the model, thereby restoring the exact context preceding text generation and ensuring consistent conditional log-probabilities for the output sequence.

Rewards are derived solely from the text output y . We incorporate two reward types: a **format reward**, which equals 1 if the response contains both `< |lvr_start| >` and `< |lvr_end| >` tokens and 0 otherwise, and an **accuracy reward**, which equals 1 if the answer is correct and 0 otherwise. The format reward encourages the LVR process while the accuracy reward indirectly supervises the latent reasoning through its impact on text generation.

The group-normalized reward is:

$$\tilde{R}_i = \frac{R(y_i) - \text{mean}(R(y_1), \dots, R(y_G))}{\text{std}(R(y_1), \dots, R(y_G))}, \quad \hat{A}_{i,t} = \tilde{R}_i \quad (\forall t \in \{1, \dots, |y_i|\}). \quad (6.6)$$

Decoding Strategies

Our initial results reveal that decoding in LVR can be challenging, as it is often unclear when the model should exit the LVR process. Specifically, when the next-token prediction

introduces a $\langle \text{lv_start} \rangle$ token, the decoder switches into the LVR mode, passing the last hidden states instead of the language model head’s predicted tokens. During this phase, the language model head continues producing token predictions, but it should ideally emit $\langle \text{lv_end} \rangle$ once an optimal length of latent reasoning has been reached, reconstructing all necessary visual semantics for the current task. In practice, however, the predicted tokens during latent reasoning are often unstable, motivating us to propose three decoding strategies:

- **Fixed Token:** assigns a constant budget of reasoning steps. Once the budget is reached, the model immediately exits latent reasoning mode.
- **Latent End Token:** introduces a trainable tensor in the hidden state space. When the last hidden state approaches this tensor, the decoder resumes text-token generation.
- **Mode Switching Loss:** adds an auxiliary loss during SFT that supervises the token distribution predicted by the language model head in the latent reasoning phase. A BCE loss encourages the distribution of the final latent reasoning token toward $\langle \text{lv_end} \rangle$ (close to 1), while all intermediate tokens are pushed away from $\langle \text{lv_end} \rangle$ (close to 0). At inference time, the model exits latent reasoning mode once $\langle \text{lv_end} \rangle$ is predicted.

Empirically, we find that **Fixed Token** achieves the best performance, while **Mode Switching Loss** fails to encode stopping conditions properly. A detailed analysis is provided in Section 6.3.

6.3 Experiments

We adopt Qwen-2.5-VL 3B and 7B as the backbone MLLMs. For the visual encoder, we set the maximum resolution to $5120 \times 28 \times 28$ pixels and the minimum to $128 \times 28 \times 28$ pixels. In both training stages, the visual encoder and multimodal projector are kept frozen, with only the LLM parameters updated. This reflects LVR’s learning objective of unifying the reasoning space under the hypothesis that an optimal modality projection can be achieved without additional tuning.

In the SFT stage, we use VISUAL COT (Shao et al., 2024b) as the primary training data. This large-scale VQA dataset contains 438k question-answer pairs, each annotated with bounding boxes that mark the critical regions needed to derive the answer. During training, the variable number of image tokens and LVR tokens results in imbalanced instance lengths. To address this, we adopt the adaptive multimodal data-packing strategy from Chen et al. (2024b), which supports dynamic batching: multiple shorter instances can be packed together, while longer ones are grouped in smaller numbers. On average, this yields an effective batch size of ~ 3.2 per device. The learning rate is set to 1×10^{-5} . For the 7B variant, supervised fine-tuning requires roughly 40 hours to complete 2,500 steps on a $4 \times$ AMD MI250 GPU cluster.

For reinforcement learning, we implement a latent GRPO framework customized from the HuggingFace TRL package, using ViRL (Wang et al., 2025a) as the training data. The policy generates 8 responses per input, with a learning rate of 1×10^{-5} , a fixed sampling temperature of $\tau = 0.9$, and a KL coefficient of $\beta = 0.04$. This stage is applied only to the 3B variant, requiring about 20 hours to complete 1,500 steps. We do not extend it to the 7B variant due to limited computational resources.

Evaluation Benchmarks

We evaluate LVR on the visual detail understanding task and a diverse set of vision-centric benchmarks. For visual detail understanding, we adopt V^* Bench, which assesses MLLMs’ ability to perform fine-grained visual detail search and relative spatial reasoning respectively on two subsets. We further employ MMVP (Tong et al., 2024c) to measure perception robustness under subtle image perturbations, providing a rigorous test of LVR’s fine-grained reasoning capabilities.

Beyond these, we evaluate on Counting (object enumeration), JigSaw (image reconstruction from fragments), Relative Reflectance (pixel-level albedo comparison), and Spatial Relation (object-relation understanding within a scene). These tasks are drawn from BLINK (Fu et al., 2024), a benchmark of expert-annotated, perception-heavy tasks designed for MLLMs.

To ensure consistency and fairness, all evaluations follow the standardized metrics provided by LMMs-Eval (Li et al., 2024b).

Baselines

We compare LVR against state-of-the-art MLLM baselines in two categories:

Thinking about Images. This includes Vision-R1 (Huang et al., 2025a) and PAPO (Wang et al., 2025f), which use RL with verifiable rewards to equip MLLMs with CoT reasoning. Vision-R1 follows a “think before answer” trajectory, while PAPO adds an implicit perception loss for image-grounded descriptions.

Thinking with Images. This category includes PixelReasoner (Su et al., 2025a) and Argus-X3 (Man et al., 2025). PixelReasoner employs image-editing tools to iteratively enhance the input image during reasoning. In contrast, Argus-X3 detects regions of interest via bounding boxes, extracts the corresponding visual tokens, and reinjects them into the MLLM input to strengthen perception. This makes Argus-X3 a strong reference point for comparison with LVR: whereas Argus-X3 depends on external tools for feature extraction, LVR directly learns to reconstruct visual semantics.

To isolate the effect of training data, we also include a baseline trained with standard supervised fine-tuning on the same dataset as LVR, denoted as SFT. For completeness, we include each baseline model’s system prompt when it is available in the documentation.

Main Results

The main evaluation results are summarized in Table 6.1. We report performance under the best decoding strategy, **Fixed Token**, with LVR steps set to [4, 8, 16]. Overall, LVR achieves state-of-the-art results across most benchmarks. Importantly, for a fair comparison, all open-source baselines are built on the same backbone MLLMs as LVR, highlighting the effectiveness of latent reasoning over existing “Think about images” or “Think with images” approaches.

The largest gains are observed on the V^* and MMVP benchmarks, where LVR consistently achieves top performance across all step settings. In particular, it improves the base model by 2.7% on the $V_{D.A.}^*$ subset, which measures visual detail search, and by 5.3% on the $V_{R.P.}^*$ subset, which evaluates relative spatial reasoning. On MMVP, which perturbs subtle image details to assess perception robustness, LVR demonstrates strong performance by reconstructing target visual semantics, accurately identifying differences, and thereby improving accuracy.

Moreover, it surpasses PixelReasoner in both categories, despite PixelReasoner relying on external tools to crop image sub-regions. This underscores that **reconstructing visual semantics can be more effective than depending on external visual-editing tools (as in “Think with Images”) for fine-grained visual understanding**. In addition, both PAPO and Vision-R1 exhibit degraded performance on V^* , suggesting that **textual-space CoT in MLLMs (i.e., “Think about Images”) may introduce cross-modal interference and weaken perception, whereas LVR avoids this issue by reasoning jointly across modalities**.

LVR also achieves leading results on Counting, IQ-Test, JigSaw, and Spatial Relation. In Counting, the model is required to enumerate specific objects in a scene; in IQ-Test, it solves geometry-based puzzles; In JigSaw, it reassembles image crops; and in Spatial Relation, it answers questions about relative object positions. These results demonstrate that LATENT VISUAL REASONING effectively unifies object detection, visual-dependent logical reasoning, visual reconstruction, and spatial relation understanding.

However, LVR does not achieve top performance on Relative Reflect. We attribute this gap to a distribution shift between training and evaluation data: such tasks require reasoning over multiple images, whereas LVR was trained exclusively on single-image data. We anticipate that incorporating cross-image data augmentation in future work will further strengthen LVR’s capability for multi-image reasoning.

RL Results

We present the results of GRPO_{latent} in Table 6.2. Reinforcement learning further enhances LVR performance beyond the SFT stage across multiple benchmarks. The superior results demonstrate that incorporating a format reward based on the `< |lvr_start| >` and `< |lvr_end| >` tokens effectively encourages the model to perform LATENT VISUAL REASONING. In contrast, removing these trigger tokens from the reward function destabilizes training, resulting in purely textual responses and degraded performance.

Ablation Studies

We examine variants in architectural design and decoding strategies.

Method	V^*	$V_{D.A.}^*$	$V_{R.P.}^*$	MMVP	Counting	IQ-Test	JigSaw	Rel. Reflect	Spatial Rel.
<i>Closed-Source Models</i>									
GPT-4o	62.8	–	–	–	51.7	30.0	58.0	38.8	76.9
Gemini2.5-Pro	79.2	–	–	–	–	–	–	–	–
ARGUS-X3	78.5	–	–	45.5	–	–	–	–	–
<i>Open Models based on Qwen2.5-VL-7B</i>									
Qwen2.5-VL	78.5	81.7	73.7	66.7	66.7	26.0	52.0	38.8	87.4
PAPO	36.1	25.2	52.6	54.3	66.7	29.3	52.0	39.6	88.8
Vision-R1	70.2	70.4	69.7	46.7	51.7	26.7	27.3	44.8	66.4
PixelReasoner	80.1	81.7	77.6	67.0	66.7	25.3	52.7	42.5	88.1
SFT	79.1	82.6	73.7	65.7	67.5	26.7	45.3	33.6	88.8
LVR (4 Steps)	81.2	84.4	76.3	72.0	69.2	28.7	52.7	42.5	89.5
LVR (8 Steps)	81.7	84.4	77.6	71.7	70.0	29.3	52.0	42.5	86.0
LVR (16 Steps)	80.6	81.7	79.0	71.7	70.8	27.3	52.7	41.8	87.4

Table 6.1: **Experimental results on vision-centric tasks.** LVR outperforms both “Think about Images” (PAPO, Vision-R1) and “Think with Images” (PixelReasoner) paradigms, highlighting the effectiveness and scalability of latent visual reasoning.

Benchmark	PAPO	LVR (4 8 16)				LVR _{RL} (4 8 16)		
V^*	31.9	64.9	65.5	66.5	65.5	67.0	66.5	
$V_{D.A.}^*$	22.6	69.6	71.3	71.3	69.6	72.2	71.3	
$V_{R.P.}^*$	46.1	60.5	60.5	56.6	59.2	59.2	59.2	
MMVP	50.0	54.7	56.0	56.0	55.3	55.3	58.0	
IQ-Test	31.3	29.3	30.7	30.0	30.7	32.0	30.0	
JigSaw	46.7	52.7	52.7	52.0	52.7	52.7	50.7	

Table 6.2: **RL results with 3B models.** GRPO_{latent} further boosts performance, demonstrating the effectiveness of adapting RL for latent reasoning and enabling self-evolution.

LVR with heads. Analogous to the LM head in LLMs, we add a LVR head on top of the latent reasoning positions to transfer LLM hidden states into visual semantics. We evaluate two designs: (i) a 2-layer MLP without intermediate up-casting, and (ii) a gated linear unit (GLU) with an intermediate dimension expanded to $3\times$ the LM hidden size. The 7B results (Table 6.3) show that the standard **LVR** consistently achieves the best performance across benchmarks. This is expected, as the LVR process is directly supervised in the joint semantic

Method	V^*	$V_{D.A.}^*$	$V_{R.P.}^*$	MMVP	IQ-Test	JigSaw
LVR	81.7	84.4	77.6	71.7	29.3	52.0
LVR _{LatentEnd}	39.8	32.2	51.3	19.0	6.7	13.3
LVR _{MLP Head}	74.4	76.5	71.1	69.7	23.3	50.0
LVR _{GLU Head}	79.6	82.6	75.0	69.0	25.3	44.0

Table 6.3: Ablation studies on the 7B model. The standard LVR approach performs best, indicating the LLM natively aligns visual and textual semantics without an extra head. The unstable Latent End Token highlights a need for future work on variable-length reasoning.

space of text and vision, ensuring no semantic gap between the LLM’s last hidden states and the LVR outputs.

Unrestricted LVR decoding. As discussed in Section 6.2, we explored **Mode Switching Loss** and **Latent End Token** for variable-length LVR processes. However, Mode Switching Loss failed to encode stopping conditions, collapsing to zero LVR steps. The performance of Latent End Token is reported in Table 6.3, where we observe significant instability. Human evaluation indicates that this instability stems from unreliable distance measurements between LVR hidden states and the latent end token. Despite testing cosine similarity, L1, and L2 distances under varying thresholds, the model often failed to terminate properly and maxed out generation steps. We expect future work on improving stability for fully free-form LATENT VISUAL REASONING.

6.4 Conclusion and Transition

This chapter presented LATENT VISUAL REASONING (LVR), a novel multimodal reasoning paradigm that unifies latent reasoning over visual tokens with standard text generation. By extending the Vision–Projector–LLM structure and training with supervised fine-tuning plus GRPO-based reinforcement learning, LVR achieves stable hybrid reasoning. Experiments show substantial gains on perception-intensive benchmarks, demonstrating that reasoning jointly over latent visual and textual spaces offers a promising direction for future multimodal reasoning.

Looking ahead. Chapters 5 and 6 together advance multimodal reasoning by improving both *what* the model sees (semantic-guided visual selection) and *how* the model thinks (latent reasoning in the embedding space). Both contributions improve the representations that the model reasons over, but neither directly addresses the *mechanism* by which the model allocates its computational resources across these representations. In the Transformer architecture, this allocation is governed by attention: which tokens in the multimodal context receive weight during each generation step. Standard post-training methods—including the GRPO algorithm employed in this chapter—optimize the model’s *output* token distribution but provide only indirect supervision for the *internal* attention patterns that determine cross-modal information flow. The next chapter proposes Reinforced Attention Learning (RAL), which closes this gap by reformulating the RL policy to operate directly on attention distributions, shifting the optimization target from *what* the model generates to *where* it attends.

Chapter 7

Reinforced Attention Learning

7.1 Introduction

The preceding chapters advanced how models represent inputs and structure reasoning—from semantic label representations (Chapter 3), through reasoning diagnostics (Chapter 4), to multimodal visual selection (Chapter 5) and latent visual reasoning (Chapter 6). This final contribution turns to a still deeper question: can we optimize the fundamental mechanism by which models allocate attention across representations? The work in this chapter is based on the paper "Reinforced Attention Learning," which is available as a preprint (Li et al., 2026a).

Large Language Models (LLMs) have achieved remarkable proficiency in complex domains such as mathematics and programming. Beyond massive unsupervised pre-training Brown et al. (2020); Team (2025), post-training has emerged as a critical technology for eliciting long-form Chain-of-Thought (CoT) reasoning. Current paradigms predominantly employ Reinforcement Learning (RL) to optimize the model’s policy, utilizing rewards derived from learned models or programmatic verifiers to favor high-utility token sequences. Fundamentally, these policy gradient methods refine the next-token distribution to maximize expected rewards. Recent research has further highlighted a robust correlation between reasoning length and task accuracy—a phenomenon central to the concept of test-time scaling.

To extend these gains to multimodal LLMs (MLLMs), recent studies have attempted to incorporate “thinking processes” into Visual Question Answering (VQA) tasks. Under this paradigm, models are incentivized to generate exhaustive textual descriptions of visual

inputs as a CoT prior to providing a final answer. However, our results reveal that for core perception tasks—such as image or video question answering—extensive textual reasoning provides only marginal gains and may even degrade the model’s fundamental perceptual capabilities.

We attribute this limitation to the insufficiency of next-token prediction as the fundamental policy objective in MLLM post-training. In typical MLLM architectures, visual inputs are encoded as tokens and projected into the textual embedding space to serve as context for generation. Accurately answering fine-grained questions requires the model to precisely identify and attend to task-relevant visual information. This process is governed by the Transformer’s attention mechanism, which must learn to assign high weights to salient multimodal tokens. Standard RLHF, however, optimizes for the result (the token) rather than the process (the internal information allocation).

Inspired by this observation, we reformulate the post-training policy for MLLMs to operate directly on the attention distribution during generation. This yields REINFORCED ATTENTION LEARNING (RAL), an algorithm designed to optimize the model toward high-utility attention trajectories. Unlike traditional methods, RAL treats the attention pattern itself as the policy: when a response receives a high reward, the algorithm encourages the underlying attention distribution by minimizing the divergence between the current and reference policies. Conversely, for low-reward responses, the model is penalized by increasing the divergence from those sub-optimal patterns. By shifting the optimization target from token likelihood to attention-based allocation, RAL fine-tunes MLLMs more directly for multimodal alignment. Our results indicate that RAL consistently outperforms Group Relative Policy Optimization (GRPO) across video and image benchmarks, particularly on perception-intensive tasks.

The efficacy of optimizing attention distributions naturally extends to **On-Policy Distillation**. While traditional distillation focuses on token-level probability alignment, we propose a dual-distillation approach that transfers knowledge via both token and attention distribution alignment.

In summary, the contributions of this chapter are:

- **Reinforced Attention Learning:** We propose RAL, a policy-gradient method that

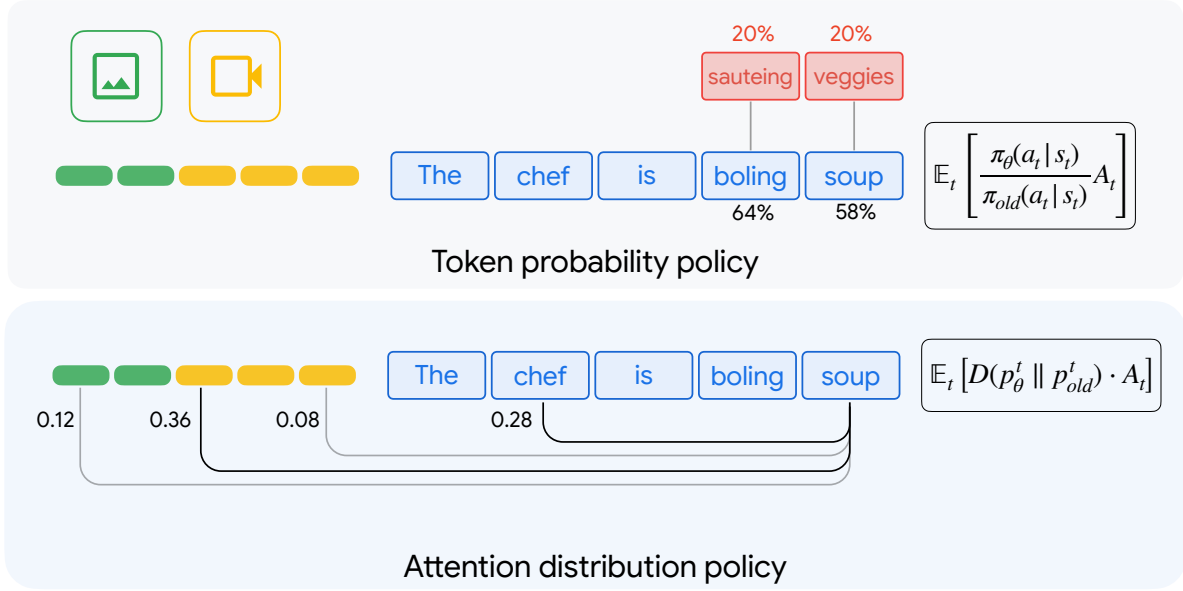


Figure 7.1: **Reinforced Attention Learning formulates internal attention distributions as a policy.** Unlike traditional methods that optimize next-token probabilities (“what to generate”), our approach prioritizes the selective allocation of information focus (“where to look”). Optimizing for the advantage, the model explores a high-rewarding attention policy that effectively isolates salient information from dense contexts.

shifts the optimization objective from next-token prediction to attention-distribution alignment, enabling direct reinforcement of visual grounding rather than indirect supervision through textual outputs.

- **On-Policy Attention Distillation:** We extend this framework to an on-policy distillation setting, which substantially improves a student model’s ability to inherit fine-grained perceptual and grounding behaviors from a teacher.
- **Empirical Validation:** Extensive experiments demonstrate that RAL consistently improves upon GRPO across diverse visual question answering benchmarks requiring fine-grained visual understanding and perception.

7.2 Reinforced Attention Learning

Traditional reinforcement learning for LLMs typically optimizes a policy by maximizing the expected return over the output token distribution. Modern off-policy algorithms like PPO and GRPO utilize a surrogate objective based on importance sampling:

$$\mathcal{L}_{\text{RL}} = \mathbb{E}_t \left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\text{old}}(a_t|s_t)} A_t \right] \quad (7.1)$$

where A_t denotes the advantage estimate. This formulation explicitly optimizes the **divergence between the current and the behavioral policy’s token distributions**. While effective for maximizing immediate rewards, this token-level optimization often precipitates **diversity collapse**. In such cases, the model overfits to specific high-reward surface forms, weakening its generalization across diverse reasoning patterns and leading to “reward hacking” of the linguistic structure rather than the underlying logic.

To mitigate this, we propose shifting the optimization target from the external output distribution to the **internal attention distributions**. By supervising how the model allocates its computational focus over contextual information, we provide a robust form of structural regularization. We frame the model’s aggregate attention over an input prompt as an **information-gathering policy**.

Aggregated Causal Attention Distribution Policy

We posit that the model’s internal attention mechanism constitutes an alternative, latent policy space. By reinforcing these internal distributions, we guide the model’s reasoning process—specifically, how it integrates both the original context and its own generated rationale—without strictly constraining the output tokens to a narrow, brittle distribution.

Let the total sequence be $S = (x_1, \dots, x_T)$, where $x_1, \dots, x_P$ represents the prompt and $x_{P+1}, \dots, x_T$ represents the generated response. Let $\alpha_{t,i}$ represent the attention weight from the generated token at position t to any preceding token at position i ($i < t$), extracted from the final layer and averaged across all heads. For each generated token $t \in [P+1, T]$, we define the causal Attention Distribution Policy p_θ^t as the distribution over all preceding positions:

$$p_\theta^t(i) = \frac{\alpha_{t,i}}{\sum_{j=1}^{t-1} \alpha_{t,j}}, \quad \forall i \in [1, t-1]. \quad (7.2)$$

This formulation captures how the model attends to its own emerging reasoning in addition to the initial instructions and visual input.

Advantage-Weighted Attention Divergence

To ensure the model retains an effective **information-gathering policy**, we derive a per-token loss function. Drawing inspiration from the importance sampling ratios in PPO/GRPO, this objective encourages attention patterns that correlate with high rewards. The objective is defined as the expected divergence of the current attention policy and the old attention policy:

$$L_{\text{AttnRL}} = \mathbb{E}_t [A_t \cdot D(p_\theta^t \parallel p_{\text{old}}^t)], \quad (7.3)$$

where A_t is the sequence-level advantage and $D(\cdot)$ is a symmetric, bounded divergence measure, such as Jensen-Shannon Divergence (JSD). Using JSD ensures training stability and behaves according to the sign of the advantage:

- If $A_t > 0$, minimizing L_{AttnRL} pulls the current policy p_θ toward the successful strategy p_{old} .
- If $A_t < 0$, the objective “pushes” p_θ away from the suboptimal strategy.

This per-token granularity prevents the “vanishing gradient” effect often encountered when averaging attention over long sequences, ensuring that even late-stage tokens in a long response are supervised by the internal attention objective.

Combined Optimization Objective

The final training objective integrates the standard token-level policy gradient with our internal attention regularizer:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{RL}} + \lambda_{\text{attn}} L_{\text{AttnRL}}, \quad (7.4)$$

where λ_{attn} is a hyperparameter balancing explicit output maximization with internal attention-level exploration. This dual-objective approach ensures the model remains linguistically flexible while maintaining a structured and reward-aligned reasoning process.

Gradient Derivation

The gradient of $\mathcal{L}_{\text{AttnRL}}$ with respect to the attention logits $e_{t,i}$ is derived via the chain rule. Let $J_t = \text{JSD}(p_\theta^t \| p_{\text{old}}^t)$.

Gradient w.r.t. Distribution. The partial derivative of JSD with respect to the current distribution p_θ^t is:

$$\nabla_{p_\theta^t} J_t = \frac{1}{2} \ln \left(\frac{2p_\theta^t}{p_\theta^t + p_{\text{old}}^t} \right).$$

Gradient w.r.t. Logits. Using the softmax Jacobian $\frac{\partial p_j}{\partial e_i} = p_i(\delta_{ij} - p_j)$, the gradient for a specific logit $e_{t,i}$ is:

$$\nabla_{e_{t,i}} J_t = p_\theta^t(i) \left(\nabla_{p_\theta^t(i)} J_t - \sum_j p_\theta^t(j) \nabla_{p_\theta^t(j)} J_t \right).$$

Total Parameter Update. The final gradient for the attention loss is the advantage-weighted accumulation across the sequence:

$$\nabla_\theta \mathcal{L}_{\text{AttnRL}} = \mathbb{E}_\tau \left[A_\tau \sum_{t=P+1}^T \sum_{i=1}^{t-1} (\nabla_{e_{t,i}} J_t) \nabla_\theta e_{t,i} \right].$$

When $A_\tau > 0$, the update minimizes divergence from successful patterns. When $A_\tau < 0$, it pushes the model to explore alternative attention fragments, penalizing the specific reasoning path that led to a low reward.

On-Policy Attention Distillation

Beyond reward-driven optimization, the attention policy framework extends to an on-policy distillation setting where the student model π_θ aims to inherit the attention distribution—the structural reasoning patterns—from a teacher model π_ϕ . In this regime, we provide dense supervision by minimizing the divergence between the student’s and teacher’s internal attention distributions over the student’s own generated trajectories.

Teacher-Student Alignment. For each token t in a trajectory τ sampled from the student policy π_θ , we define the distillation loss as the sum of divergences between the student attention policy p_θ^t and the teacher attention policy p_ϕ^t :

$$\mathcal{L}_{\text{AttnDistill}} = \mathbb{E}_{\tau \sim \pi_\theta} \left[\sum_{t=P+1}^T \text{JSD}(p_\theta^t \| p_\phi^t) \right]. \quad (7.5)$$

Crucially, this objective does not involve an advantage term A_t . The goal is pure structural imitation, ensuring that for any generated token, the student utilizes the same contextual evidence as the teacher. This provides a denser gradient signal than token-level KL divergence alone.

Unified Distillation Objective. The final objective for on-policy distillation combines the RL objective, the generalized knowledge distillation (GKD) on output logits, and our proposed attention alignment:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{RL}} + \mu \mathcal{L}_{\text{GKD}} + \gamma_{\text{attn}} \mathcal{L}_{\text{AttnDistill}}, \quad (7.6)$$

where $\mathcal{L}_{\text{GKD}}$ typically represents the reverse KL divergence between the output distributions $\pi_\theta(a_t|s_t)$ and $\pi_\phi(a_t|s_t)$. μ controls the strength of the distillation and γ_{attn} balances the strength of the attention distillation.

Gradient Flow. The gradient of $\mathcal{L}_{\text{AttnDistill}}$ with respect to student parameters θ is derived similarly to Section 7.2, but aligns purely toward the teacher’s distribution:

$$\nabla_\theta \mathcal{L}_{\text{AttnDistill}} = \mathbb{E}_{\tau \sim \pi_\theta} \left[\sum_{t=P+1}^T \sum_{i=1}^{t-1} (\nabla_{e_{t,i}} J_t) \nabla_\theta e_{t,i} \right], \quad (7.7)$$

where $J_t = \text{JSD}(p_\theta^t \| p_\phi^t)$. By sampling trajectories from the student ($\tau \sim \pi_\theta$), the model learns to maintain “teacher-like” attention patterns even when navigating states it would not encounter under the teacher’s original distribution, thereby mitigating exposure bias in internal representations.

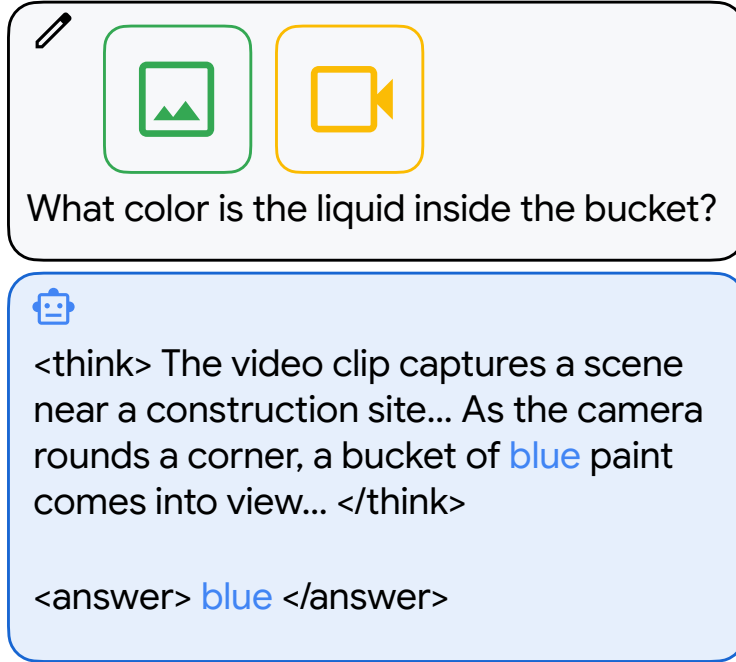


Figure 7.2: **Sample data of the SFT and RL training stages.** The SFT stage adapts the model to a “think-and-answer” paradigm, while the RL stage employs a reward function to verify the format and correctness of rollout responses.

7.3 Experiments

We empirically evaluate RAL for post-training MLLMs. Our experiments investigate both the standard RAL objective and the on-policy attention distillation variant, implemented on top of the GRPO framework.

Experimental Setup

Model Configurations. We adopt **Qwen-2.5-VL-7B** as our foundation MLLM. For the on-policy attention distillation experiments, we utilize **Qwen-2.5-VL-32B** as the teacher model. In all settings, the visual encoder and multimodal projector are kept frozen, with gradients updated only for the language model backbone.

Training Pipeline. Our training pipeline utilizes the VIDEO-R1 (Feng et al., 2025) dataset and consists of two stages, implemented using the **veRL** infrastructure (Sheng et al., 2025):

- **Supervised Fine-Tuning (SFT):** We perform SFT using the VIDEO-R1-COT-165K subset, which provides 165k instances of video-question pairs with structured Chain-of-Thought reasoning. Following the format in Figure 7.2, the “thinking” process and final answer are enclosed in `<think>` and `<answer>` tags, respectively. This stage requires roughly 10 hours on a cluster of $8 \times$ NVIDIA H100s.
- **Reinforcement Learning (RL):** We subsample the first 51.2k instances from the VIDEO-R1-260K dataset and train for a single epoch. For each input, the policy generates $G = 8$ rollouts for advantage estimation. This stage requires approximately 120 hours on $8 \times$ NVIDIA H100s.

Implementation Details

Visual Processing. For the visual encoder, we set the maximum resolution of images to $5120 \times 28 \times 28$ pixels and the minimum to $128 \times 28 \times 28$ pixels. Video frames are sampled at 1 fps, with the total number capped at 128. Each individual frame is constrained to a maximum of $128 \times 28 \times 28$ pixels.

Hyperparameters. We extract the attention weights of the last Transformer layer by patching its attention implementation with eager attention. The weights are averaged across all attention heads. For RL training, we maintain a fixed sampling temperature of $\tau = 0.9$ and a KL coefficient of $\beta = 0.04$. The learning rate is 5×10^{-6} for SFT and 1×10^{-5} for RL. Table 7.1 provides the hyperparameter ranges for RAL.

Baselines. We primarily compare against **GRPO**, which eliminates the need for a separate value model by computing the relative advantage of each response within a group of G sampled outputs. For a given response i , the advantage is:

$$\hat{A}_i = \frac{r_i - \frac{1}{G} \sum_{j=1}^G r_j}{\text{std}(r_1, \dots, r_G)}, \quad (7.8)$$

Parameter	Value
Rollout batch size	8
KL Coefficient (β)	0.04
Sampling Temperature (τ)	1
λ_{attn}	$\{0.5, 1, 5\}$
γ_{attn}	$\{0.05, 0.5, 1\}$

Table 7.1: Hyperparameters for RAL experiments.

where r_i represents the total reward. Our reward system is entirely rule-based, comprising two components: an **Accuracy Reward** (r_{acc}), which grants 1.0 for an exact match between the `<answer>` content and the ground truth; and a **Format Reward** (r_{fmt}), which grants 1.0 for perfect adherence to the `<think>...<answer>...` template. The final reward is $r_i = 0.9 \cdot r_{acc} + 0.1 \cdot r_{fmt}$. These settings remain consistent across all experiments.

We also include **Video-R1-7B** for comparison, which uses Temporal-GRPO—a variant incorporating a temporal coherence reward that perturbs frame ordering and provides auxiliary rewards for accuracy on correctly sequenced inputs. Since Video-R1-7B uses the same Qwen-2.5-VL-7B backbone, it facilitates controlled comparison.

Evaluation Benchmarks

We evaluate across a broad spectrum of vision-centric benchmarks.

Image QA Tasks. For static image understanding: (1) V^* Bench for fine-grained visual search and spatial reasoning; (2) MMMU Pro (Yue et al., 2024) for multi-step expert-level knowledge; (3) MME (Fu et al., 2025a) for basic perception and high-level cognition; (4) MuirBench (Wang et al., 2024a) for robustness across diverse image types; (5) ChartQA (Masry et al., 2022) for complex data extraction; (6) VizWiz (Gurari et al., 2018) for real-world visual grounding; (7) Blink (Fu et al., 2024) for foundational perception tasks; and (8) CVBench (Tong et al., 2024a) for core computer vision capabilities.

Video QA Tasks. For temporal reasoning and long-context integration: (1) LongVideoBench (Wu et al., 2024) for long-range referring reasoning; (2) NExT-QA (Xiao et al., 2021) for causal and temporal action explanation; (3) Video-MME (Fu et al., 2025b) for comprehensive multi-domain evaluation; (4) Video-MMMU (Hu et al., 2025) for expert-level knowledge acquisition; (5) LVBench (Wang et al., 2024b) for extreme long-form comprehension; (6) MVBench (Li et al., 2024d) for multi-task temporal perception; and (7) TempCompass (Liu et al., 2024d) for sensitivity to motion and temporal order.

Main Results

Tables 7.2 and 7.3 present a comprehensive comparison of RAL against GRPO and standard on-policy distillation. Overall, RAL consistently surpasses both the GRPO baseline and the base Qwen-2.5-VL-7B model, suggesting that optimizing internal attention distributions provides a more stable and effective learning signal than token-level policy gradients alone.

Image VQA. Under the reinforcement learning paradigm, RAL outperforms GRPO across all eight image benchmarks. We observe significant improvements on V^* (+5.8), MME (+94.1), ChartQA (+2.8), and VizWiz (+3.8). These gains indicate that attention-level supervision strengthens visual grounding and compositional reasoning, particularly for perception-intensive and document-based tasks. Crucially, RAL not only mitigates the performance degradation often introduced by GRPO relative to the base model (e.g., on V^* and VizWiz) but also exceeds the original Qwen-2.5-VL-7B performance across all metrics. This suggests that our approach fosters genuine generalization rather than narrow reward-model overfitting.

In the on-policy distillation regime, the integration of attention distillation yields superior results over standard distillation on 7 out of 8 benchmarks, with marked increases on V^* (+3.6) and MuirBench (+1.8). These findings demonstrate that supervising where a teacher model attends provides a crucial, complementary signal that simple output imitation lacks. Notably, the strong performance on V^* , which specifically probes fine-grained object attributes and spatial relationships, underscores RAL’s efficacy in resolving complex scene geometries.

Method	V^*	MMMUpro	MME	MuirBench	ChartQA	VizWiz	Blink	CVBench
Qwen-2.5-VL-7B	70.7	36.4	2309.3	44.9	84.0	71.2	56.3	77.6
<i>Reinforcement Learning</i>								
Video-R1	66.5	36.0	2266.2	40.9	84.2	63.5	47.4	72.3
GRPO	68.6	36.8	2258.7	43.9	81.7	67.9	54.9	78.1
RAL	73.3	37.8	2352.8	47.4	86.4	71.7	57.2	79.0
RAL-zero	72.3	38.4	2306.2	43.4	82.7	71.9	55.4	78.4
<i>On-Policy Distillation</i>								
On-policy Distillation	68.1	37.8	2344.7	39.9	81.7	71.2	56.3	78.7
+ Attention Distillation	72.3	38.4	2345.1	43.4	83.6	72.1	57.2	78.4

Table 7.2: **Performance comparison across image VQA benchmarks.**

Video VQA. On long-video benchmarks (Table 7.3), RAL outperforms GRPO on 6 out of 7 datasets. The most pronounced improvements occur on LongVideoBench (+2.2), NExTQA (+3.4), and MVBench (+1.5), all of which demand robust temporal understanding and multi-hop reasoning. While GRPO maintains a marginal lead on VideoMMMU, RAL remains highly competitive.

Under on-policy distillation, attention-enhanced alignment improves performance on every benchmark except for ties on LongVideoBench and MVBench. The substantial gains on NExTQA (+4.4) and VideoMME (+2.6) suggest that attention alignment is particularly potent for long-context temporal reasoning and precise event localization.

Implications. Across both modalities and training regimes, attention-centric learning provides consistent and well-distributed performance increments. In contrast to GRPO, which can exhibit benchmark-specific trade-offs or degrade base model capabilities, RAL delivers uniform gains. This supports our hypothesis that supervising internal information allocation offers a more stable and generalizable training signal than pure token-level gradients. Furthermore, the success of attention distillation confirms that attention distributions serve as a transferable and semantically rich representation of reasoning behavior.

In summary, these results validate attention-based policy optimization as a robust complement for, conventional RL and distillation in MLLM post-training, especially for perception-heavy and long-horizon video understanding.

Method	LongVideoBench	NExTQA	VideoMME	VideoMMMU	LVBench	MVBench	TempCompass
Qwen-2.5-VL-7B	57.0	73.7	61.6	47.6	40.5	63.4	69.5
<i>Reinforcement Learning</i>							
Video-R1	56.5	65.3	62.5	44.7	43.5	62.0	68.0
GRPO	57.9	70.7	62.0	49.7	43.9	64.0	68.3
RAL	60.1	74.1	63.4	48.6	44.2	65.5	70.0
RAL-zero	58.8	76.2	65.1	49.2	45.9	62.7	68.5
<i>On-Policy Distillation</i>							
Knowledge Distillation	59.7	70.9	61.3	47.3	43.7	65.5	69.2
+ Attention Distillation	59.7	75.3	63.9	48.5	44.7	65.5	70.3

Table 7.3: **Performance comparison across long-video benchmarks.** RAL and Attention Distillation consistently enhance multimodal understanding. RAL-zero denotes the variant where the explicit thinking process is removed, isolating the impact of the attention-based policy gradient.

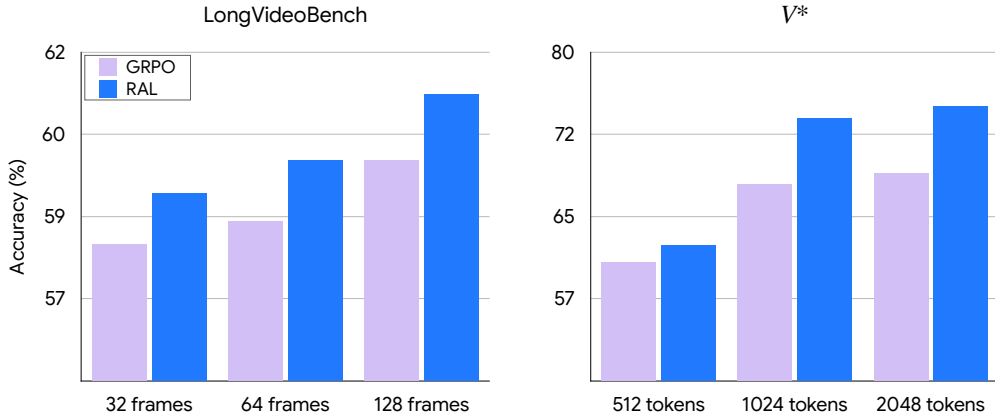


Figure 7.3: **RAL improves GRPO along increasing video frames or image resolution.**

Ablation Studies

RAL yields consistent improvements across varying visual resolutions and frame rates. To investigate the robustness of these gains relative to visual information density, we analyze the performance of RAL versus GRPO across different video sampling lengths and image resolutions. We utilize LongVideoBench and V^* as our evaluation suite: the former requires extracting sparse, salient information from extremely long video sequences, while the latter features high-resolution images of complex scenes that demand fine-grained reasoning. Together, these benchmarks facilitate a diagnostic analysis of cross-modal reasoning

capabilities across diverse temporal and spatial densities.

We evaluate on LongVideoBench using maximum frame counts of 32, 64, and 128. For V^* , we vary the maximum token budget per image (512, 1024, and 2048 tokens, where a unit image patch is 28×28 pixels, so 1024 tokens correspond approximately to an 896×896 image).

As illustrated in Figure 7.3, RAL consistently outperforms GRPO across all temporal scales, demonstrating the efficacy of attention-based policies in locating salient cues within dense temporal contexts. On V^* , the performance margin widens as image resolution increases, rising from a +1.6 gain at 512 tokens to a significant +6.3 gain at 2048 tokens. This trend indicates that the advantages of RAL become increasingly pronounced as visual information becomes more granular, suggesting superior scalability for high-fidelity multimodal understanding.

Is an Explicit Thinking Process Necessary for VQA? To isolate the impact of attention-level supervision, we investigate whether RAL improves cross-modal reasoning even in the absence of an explicit thinking process. We introduce RAL-zero, a variant where the thinking process is completely removed from both the SFT and RL stages. In this configuration, the model is trained to generate the final answer directly. By excluding high-reward text tokens (rationales), the training signal is dominated by the attention policy, allowing us to evaluate its intrinsic contribution to visual understanding.

We maintain all hyper-parameters and data volumes from the main experiments but modify the data format and reward functions. During SFT, the `<think>` blocks (see Figure 7.2) are stripped, forcing the model to produce only the final response. Correspondingly, the RL stage employs a format reward that penalizes any output different from the `<answer>...</answer>` structure.

As shown in Table 7.3, RAL-zero surpasses the base model on 5 out of 7 long-video benchmarks and outperforms the full GRPO baseline on 5 benchmarks. Notably, RAL-zero achieves state-of-the-art performance on NExTQA (temporal reasoning), VideoMME (comprehensive video understanding), and LVBench (long-video event understanding). These results demonstrate that the attention policy space is significantly underexplored; RAL effectively unlocks visual understanding capabilities by optimizing internal attention weight

distributions.

On image benchmarks (Table 7.2), RAL-zero improves upon the base model in 4 out of 8 cases, achieving the highest scores among RL-based methods on MMMU-Pro and VizWiz. Given that MMMU-Pro tests complex visual knowledge and VizWiz focuses on fine-grained object recognition, these gains suggest that optimizing attention distributions serves as a powerful bridge between modalities. This confirms that policy gradient methods applied directly to attention mechanisms can induce superior cross-modal reasoning, even without explicit verbalized logic.

7.4 Conclusion and Transition

This chapter introduced **Reinforced Attention Learning**, a MLLM post-training paradigm that shifts optimization from text token distributions to internal attention distributions. By treating attention as a policy, RAL directly reinforces visual grounding and perceptual focus, addressing a fundamental limitation of outcome-based RL methods that neglect the underlying cross-modal reasoning process.

Experiments across diverse image and long-video benchmarks demonstrate that RAL consistently outperforms the base Qwen-2.5-VL-7B and GRPO baselines. The approach provides more stable and uniform gains than token-level RL, which can occasionally degrade base model performance. These results validate our hypothesis that supervising internal information allocation yields a more reliable and generalizable training signal than next-token gradients alone. Furthermore, the attention-centric perspective extends naturally to on-policy distillation, where transferring “where to focus” provides a complementary and semantically rich signal that surpasses simple output imitation.

Ultimately, this work establishes attention distributions as a first-class optimization target for multimodal alignment. By reinforcing internal computation pathways, RAL offers a principled, process-aware alternative to standard RLHF.

Looking ahead. This chapter completes the five contributions of this dissertation. Taken together, they trace a coherent arc through the representation-centric study of machine

reasoning: from reformulating output spaces with rich semantic representations (LITE), through diagnosing the shortcut vulnerabilities that arise when internal representations are shallow (EUREQA), to improving the visual representations that enter the model’s context (SEMCLIP), enabling reasoning directly within the latent visual representation space (LVR), and finally optimizing the fundamental mechanism—attention—that governs how the model routes information across all of these representations (RAL). The next and final chapter synthesizes these contributions, discusses their limitations, and outlines directions for future research.

Chapter 8

Conclusion and Future Work

This dissertation has investigated the thesis that the quality of machine reasoning is fundamentally constrained by the quality of the underlying representations, and that systematic improvements to how models represent, select, and attend to information yield reasoning systems that are more efficient, more robust, and less prone to hallucination across both unimodal and multimodal settings. Across five interconnected contributions, we traced an arc from the limitations of categorical representations in natural language understanding, through the diagnosis of reasoning fragility in large language models, to the development of novel multimodal reasoning paradigms that operate in learned latent spaces and optimize the attention mechanisms that govern cross-modal information flow.

This chapter synthesizes the contributions (Section 8.1) and outlines directions for future research (Section 8.2), with particular emphasis on latent-space reasoning and self-evolving models that learn from unlabeled data and real-world interaction.

8.1 Summary of Contributions

LITE: Semantic Representations for Entity Typing (Chapter 3). We introduced LITE, the first framework to reformulate entity typing as a natural language inference problem. By transforming candidate types into textual hypotheses and scoring them via entailment, LITE preserves the semantic content of type labels that classification-based systems discard. This representational choice enables state-of-the-art performance on ultra-fine, fine-grained,

and zero-shot entity typing benchmarks, and demonstrates strong transferability across datasets with disjoint type inventories. The core insight—that reasoning over semantic descriptions rather than opaque indices unlocks generalization—recurs as a motif throughout the dissertation.

EUREQA: Diagnosing Reasoning Shortcuts (Chapter 4). We proposed EUREQA, a multi-hop question-answering benchmark designed to systematically strip away semantic shortcuts and expose the reasoning fragility of large language models. By replacing entities with recursive sub-queries and introducing deceptive semantic cues, EUREQA isolates genuine multi-step inference from distributional pattern matching. Our evaluation revealed a striking gap between human performance (95%) and the best LLM (62–64%), with model accuracy strongly correlated to entity similarity rather than reasoning depth. These findings established that the reasoning bottleneck in LLMs is representational—rooted in the model’s reliance on biased linguistic priors rather than grounded logical inference.

SEMCLIP: Semantic-Guided Visual Selection (Chapter 5). We developed SEMCLIP, a plug-and-play framework that uses textual query semantics to dynamically select the most task-relevant sub-image for a vision-language model. Without requiring any retraining of the VLM itself, SEMCLIP achieves performance gains that exceed both retraining-based baselines and models with substantially larger language backbones, while using 60% fewer visual tokens than the strongest baseline. The theoretical upper-bound analysis revealed a 15.4% average accuracy gap between current VLM performance and the potential achievable through optimal visual selection, demonstrating that the question of *which* visual representations enter the reasoning context is a high-leverage and under-explored dimension of multimodal reasoning.

LVR: Latent Visual Reasoning (Chapter 6). We introduced Latent Visual Reasoning, a paradigm in which the language model generates latent states that reconstruct question-relevant visual features directly in the shared embedding space, rather than translating all reasoning into text. Through a two-stage training pipeline combining supervised fine-tuning with a novel GRPO_{latent} algorithm that replays latent reasoning steps during policy gradient computation, LVR achieves state-of-the-art results on perception-intensive benchmarks,

outperforming both “thinking about images” (text-only CoT) and “thinking with images” (external tool) paradigms. LVR demonstrates that the shared latent space of multimodal LLMs can serve as a viable medium for reasoning, not merely a conduit for translating visual information into text.

RAL: Reinforced Attention Learning (Chapter 7). We proposed Reinforced Attention Learning, which shifts the post-training optimization objective from what the model generates (token distributions) to where it attends (attention distributions). By treating the Transformer’s internal attention pattern as a policy and applying advantage-weighted divergence optimization, RAL provides a more direct training signal for cross-modal alignment than token-level RL alone. RAL consistently outperforms GRPO across all image and video benchmarks, with gains that scale with visual information density. The framework extends naturally to on-policy attention distillation, where transferring “where to focus” from teacher to student provides a complementary signal that surpasses output-level imitation. Notably, RAL-zero—a variant without any explicit thinking process—achieves state-of-the-art results on several benchmarks, confirming that the attention policy space is a powerful and under-explored dimension of multimodal optimization.

8.2 Future Work

The contributions of this dissertation open several promising research directions. We organize the discussion around three themes: extending latent-space reasoning beyond its current boundaries, enabling models to self-evolve through interaction with unlabeled data and the real world, and unifying the representational insights of this dissertation into more general frameworks.

Machine Reasoning in Latent Space

The Latent Visual Reasoning paradigm introduced in Chapter 6 demonstrated that multimodal LLMs can reason productively in continuous embedding spaces, generating “visual thoughts” that reconstruct task-relevant features without decoding them into text. This result suggests

a far broader research program: developing models that reason natively in latent space across all modalities, rather than forcing all cognition through the bottleneck of discrete token generation.

Multi-step latent reasoning chains. LVR currently performs a single latent reasoning episode per query—the model enters LVR mode, reconstructs a set of visual features, and then exits to generate text. A natural extension is to enable multi-step latent reasoning, where the model alternates between latent and textual reasoning across multiple episodes within a single generation. This would allow the model to iteratively refine its understanding: an initial latent step might identify a region of interest, a textual step might formulate a sub-question about that region, and a subsequent latent step might zoom into the relevant details. Such interleaved reasoning would more closely approximate the iterative, multi-fixation nature of human visual problem-solving, where attention shifts dynamically between the global scene and local details in response to evolving task demands.

Cross-modal latent reasoning. The latent reasoning framework need not be restricted to the visual modality. In principle, the same mechanism could be applied to audio, video, 3D point clouds, or any modality whose continuous representations are projected into the LLM’s embedding space. A model capable of generating latent “auditory thoughts” or “spatial thoughts” would represent a significant step toward genuinely multimodal cognition. The key technical challenge is defining appropriate reconstruction targets for each modality—visual patches have a natural spatial correspondence that makes them amenable to MSE supervision, but analogous supervision signals for audio spectrograms or 3D scenes require further investigation.

Latent reasoning without explicit supervision. A fundamental limitation of the current LVR approach is its reliance on bounding box annotations to define which visual features the model should reconstruct during the supervised fine-tuning stage. This creates a dependency on labeled data that limits scalability. Future work should explore self-supervised objectives for latent reasoning, where the model learns which features to reconstruct based on the downstream utility of the reconstruction rather than explicit spatial annotations. One

promising direction is to combine the latent reasoning framework with the reinforcement learning approach of RAL: the model could be rewarded for generating latent states that improve answer accuracy, without any explicit specification of what those states should encode. The $\text{GRPO}_{\text{latent}}$ algorithm introduced in Chapter 6 takes a first step in this direction, but the RL stage still builds on a model that was initialized with supervised bounding box data. A fully unsupervised latent reasoning curriculum—perhaps progressing from simple reconstruction tasks to increasingly abstract reasoning—would significantly broaden the applicability of the paradigm.

Theoretical foundations of latent reasoning. The empirical success of LVR raises fundamental theoretical questions. Why does reasoning in continuous space outperform reasoning in discrete token space for perception-intensive tasks? One hypothesis is that continuous representations preserve fine-grained distinctions that are lost in the discretization process of tokenization—subtle differences in color, texture, or spatial arrangement that are critical for perception but that map to identical or similar token sequences. A related hypothesis is that the autoregressive generation of text tokens introduces a sequential bottleneck: the model must commit to a linear ordering of information that may not match the inherently parallel structure of visual scenes. Formalizing these intuitions—perhaps through information-theoretic analyses of the capacity of discrete versus continuous reasoning channels—would provide principled guidance for when latent reasoning should be preferred over textual reasoning, and how the two should be combined.

Self-Evolution: Learning from Unlabeled Data and Real-World Interaction

A recurring theme across the contributions of this dissertation is the tension between the richness of the models’ representational capabilities and the scarcity of the supervision signals available to develop them. LITE requires typed entity annotations, EUREQA relies on structured knowledge bases, SEMCLIP needs VQA labels, LVR depends on bounding box annotations, and RAL requires task-specific reward functions. In each case, the quality and diversity of the training signal constrains the model’s ability to develop robust reasoning

capabilities. The most important long-term direction for this research program is to move beyond these supervised and semi-supervised paradigms toward models that can *self-evolve*—improving their reasoning capabilities autonomously through interaction with unlabeled data and the real world.

Self-supervised reasoning from unlabeled text and images. The vast majority of text and images on the internet are unlabeled. Current pre-training paradigms exploit this data for representation learning (via next-token prediction or contrastive learning), but they do not explicitly train reasoning capabilities. A self-evolving model would go further: given an unlabeled document or image, it would formulate its own questions, generate candidate answers, verify them against internal consistency criteria, and update its reasoning mechanisms based on the outcomes. This vision connects to the growing literature on self-play and self-improvement in language models, but extends it to the multimodal setting where verification is both more challenging (visual facts are harder to check than textual ones) and more rewarding (the space of visual reasoning tasks is vastly larger than what any curated benchmark can cover).

Concretely, one could imagine a training loop in which the model: (i) encounters an unlabeled image, (ii) uses its current capabilities to generate a set of questions about the image (e.g., “What objects are present?” “What is the spatial relationship between X and Y?”), (iii) generates candidate answers using different reasoning strategies (e.g., text-only CoT versus LVR), (iv) compares the answers for consistency and assigns self-generated rewards, and (v) updates its parameters using the RAL framework to reinforce attention patterns that led to consistent, high-confidence answers. This self-supervised loop would require no human annotations and could scale to the entire distribution of images and text on the web.

Learning from interaction with the physical world. The ultimate test of a reasoning system is whether it can operate in the real world, where the consequences of reasoning failures are immediate and tangible. Embodied AI systems—robots that navigate, manipulate objects, and communicate with humans—provide a natural testbed for the representational principles developed in this dissertation. A robot that must locate a specific tool in a cluttered workshop faces precisely the challenge that SEMCLIP addresses: selecting the most task-relevant visual

region from a complex scene. A robot that must plan a multi-step assembly task faces the multi-hop reasoning challenge that EUREQA diagnoses. And a robot that must allocate its visual attention efficiently across a dynamic scene faces the attention optimization challenge that RAL addresses.

The key insight connecting this dissertation to embodied AI is that real-world interaction provides a natural reward signal for reasoning quality: a robot that reasons correctly about spatial relationships will successfully grasp the target object; one that takes semantic shortcuts will fail. This reward signal is richer, more diverse, and more aligned with genuine reasoning capability than the rule-based rewards used in our current experiments. Future work should investigate how the latent reasoning and attention optimization frameworks developed in this dissertation can be adapted to learn from embodied interaction, where the model receives continuous, grounded feedback from the physical world rather than discrete, annotation-based feedback from human evaluators.

Curriculum learning for reasoning self-evolution. A self-evolving model cannot improve its reasoning capabilities by simply processing more data at a fixed difficulty level—it must encounter progressively more challenging reasoning tasks that push the boundaries of its current abilities. This suggests the need for an automatic curriculum generation mechanism that identifies the frontier of the model’s reasoning capabilities and generates training instances just beyond that frontier. The EUREQA benchmark provides a template for such curriculum generation in the textual domain: by systematically increasing reasoning depth and removing semantic shortcuts, it creates a natural difficulty progression. Extending this idea, a self-evolving model could automatically generate its own EUREQA-style challenges—identifying the types of reasoning chains on which it currently fails and constructing targeted training instances to address those weaknesses.

In the multimodal domain, curriculum generation could take the form of progressively more challenging visual reasoning tasks: starting with simple object identification, progressing to spatial relationship reasoning, then to multi-object counting and attribute comparison, and finally to complex scene understanding that requires integrating information from multiple image regions. The latent visual reasoning framework provides a natural mechanism for this progression: as the model’s latent reasoning capabilities improve, it can reconstruct

increasingly detailed and abstract visual features, enabling it to tackle more sophisticated reasoning tasks.

Continual self-improvement without catastrophic forgetting. A persistent challenge for self-evolving models is catastrophic forgetting: as the model improves on new reasoning tasks, it may degrade on previously mastered capabilities. The attention-centric optimization framework of RAL offers a potential mitigation strategy. Because RAL optimizes attention patterns rather than output distributions, it may be possible to maintain a library of “attention templates”—successful attention patterns for different types of reasoning tasks—that can be selectively activated or composed during inference. This would allow the model to accumulate reasoning capabilities over time without overwriting previously learned attention strategies. More broadly, the decomposition of reasoning into representational components (what to represent), selection components (what to attend to), and generation components (what to output) may provide natural axes along which continual learning can be managed, with different components updated at different rates depending on the stability of the underlying reasoning patterns.

Toward Unified Reasoning Architectures

Integrating the five contributions. The five contributions of this dissertation were developed and evaluated independently, but they address complementary aspects of the reasoning pipeline. A natural direction for future work is to integrate them into a unified system. Such a system would use semantic representations for its output space (LITE), include mechanisms for detecting and avoiding semantic shortcuts (EUREQA-inspired verification), employ query-guided visual selection at the input level (SEMCLIP), perform hybrid latent-textual reasoning in the embedding space (LVR), and optimize its internal attention patterns through reinforcement learning (RAL). The technical challenge lies in managing the interactions between these components: for instance, the visual selection performed by SEMCLIP determines which visual tokens are available for LVR’s latent reasoning, which in turn affects the attention patterns that RAL optimizes. Jointly optimizing across these interdependencies is a non-trivial research problem, but one whose solution could yield

reasoning systems that are substantially more capable than any individual component.

Attention as a general optimization target. RAL demonstrated that attention distributions can be treated as a first-class optimization target for multimodal alignment. This principle extends naturally to other internal structures of modern architectures. Mixture-of-Experts (MoE) routing decisions, cross-modal fusion gates, and hierarchical memory access patterns all represent forms of internal information allocation that could benefit from direct optimization rather than being treated as implicit byproducts of output-level training. The mathematical framework of RAL—advantage-weighted divergence optimization over internal distributions—is sufficiently general to be applied to any differentiable distribution within the model, opening a broad design space for process-aware post-training methods.

From perception to world models. The multimodal reasoning capabilities developed in this dissertation operate on static snapshots: individual images or pre-recorded videos. The next frontier is reasoning about dynamic, interactive environments—that is, building world models that can predict how scenes will evolve in response to actions. Latent visual reasoning provides a natural foundation for world modeling: instead of reconstructing the current visual state, the model could be trained to predict future visual states in the embedding space, enabling planning and counterfactual reasoning without the computational expense of pixel-level video generation. The attention optimization framework of RAL could be adapted to this setting by rewarding attention patterns that lead to accurate predictions of future states, effectively teaching the model to focus on the aspects of the current scene that are most predictive of future dynamics.

8.3 Closing Remarks

This dissertation began with the observation that the bottleneck in machine reasoning is not merely computational but representational. Across five studies spanning entity typing, multi-hop question answering, vision-language modeling, latent-space reasoning, and attention optimization, we have provided evidence that this observation holds at multiple levels of the reasoning pipeline: in how output candidates are represented, in how reasoning chains

are structured, in how visual evidence is selected, in what space reasoning occurs, and in how internal computation is allocated. Each contribution demonstrated that targeted improvements to representations yield substantial gains in reasoning quality, efficiency, and robustness.

Looking forward, the most exciting prospect is the convergence of latent-space reasoning with self-evolving learning paradigms. If models can learn to reason in continuous representational spaces—freed from the sequential bottleneck of token generation—and if they can improve their reasoning capabilities autonomously through interaction with the vast, unlabeled, dynamic world around them, then the representational principles established in this dissertation may serve as foundations for reasoning systems of far greater scope and capability than those we have demonstrated here. The path from static benchmarks to dynamic, self-improving, world-interacting intelligence is long, but it begins with getting the representations right.

Bibliography

- Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The twelfth international conference on learning representations*, 2024.
- Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Vqa: Visual question answering. *International Journal of Computer Vision*, 123:4 – 31, 2015. URL <https://api.semanticscholar.org/CorpusID:3180429>.
- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.
- Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *Proceedings of the 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference*, ISWC’07/ASWC’07, page 722–735, Berlin, Heidelberg, 2007. Springer-Verlag. ISBN 3540762973.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Siboz Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaozhai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Siboz Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaozhai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL <https://arxiv.org/abs/2502.13923>.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Siboz Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025c.

Sule Bai, Mingxing Li, Yong Liu, Jing Tang, Haoji Zhang, Lei Sun, Xiangxiang Chu, and Yansong Tang. Univg-r1: Reasoning guided universal visual grounding with reinforcement learning. *arXiv preprint arXiv:2505.14231*, 2025d.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022. URL <https://arxiv.org/abs/2212.08073>.

Mahtab Bigverdi, Zelun Luo, Cheng-Yu Hsieh, Ethan Shen, Dongping Chen, Linda G Shapiro, and Ranjay Krishna. Perception tokens enhance visual reasoning in multimodal language

- models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3836–3845, 2025.
- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL <https://aclanthology.org/D15-1075>.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL <https://arxiv.org/abs/2005.14165>.
- Mu Cai, Jianwei Yang, Jianfeng Gao, and Yong Jae Lee. Matryoshka multimodal models. *arXiv preprint arXiv:2405.17430*, 2024.
- Rui Cai, Bangzheng Li, Xiaofei Wen, Muhao Chen, and Zhe Zhao. Diagnosing and mitigating modality interference in multimodal large language models. *arXiv preprint arXiv:2505.19616*, 2025.
- Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. Sft or rl? an early investigation into training rl-like reasoning large vision-language models. *arXiv preprint arXiv:2504.11468*, 2025a.
- Jifan Chen, Eunsol Choi, and Greg Durrett. Can NLI models verify QA systems’ predictions? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3841–3854, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.324. URL <https://aclanthology.org/2021.findings-emnlp.324>.

- Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm’s referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024a.
- Shuang Chen, Yue Guo, Zhaochen Su, Yafu Li, Yulun Wu, Jiacheng Chen, Jiayu Chen, Weijie Wang, Xiaoye Qu, and Yu Cheng. Advancing multimodal reasoning: From optimized cold start to staged reinforcement learning. *arXiv preprint arXiv:2506.04207*, 2025b.
- Tongfei Chen, Yunmo Chen, and Benjamin Van Durme. Hierarchical entity typing via multi-level learning to rank. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8465–8475, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.749. URL <https://aclanthology.org/2020.acl-main.749>.
- Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b.
- Jeffrey Cheng and Benjamin Van Durme. Compressed chain of thought: Efficient reasoning through dense representations. *arXiv preprint arXiv:2412.13171*, 2024.
- Eunsol Choi, Omer Levy, Yejin Choi, and Luke Zettlemoyer. Ultra-fine entity typing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 87–96, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1009. URL <https://aclanthology.org/P18-1009>.
- Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

- Jiwan Chung, Junhyeok Kim, Siyeol Kim, Jaeyoung Lee, Min Soo Kim, and Youngjae Yu. Don't look only once: Towards multimodal interactive reasoning with selective visual revisitation. *arXiv preprint arXiv:2505.18842*, 2025.
- Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In Joaquin Quiñonero-Candela, Ido Dagan, Bernardo Magnini, and Florence d'Alché Buc, editors, *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*, pages 177–190, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-33428-6.
- Hongliang Dai, Donghong Du, Xin Li, and Yangqiu Song. Improving fine-grained entity typing with entity linking. 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 6210–6215, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1643. URL <https://aclanthology.org/D19-1643>.
- Hongliang Dai, Yangqiu Song, and Haixun Wang. Ultra-fine entity typing with weak supervision from a masked language model. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1790–1799, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.141. URL <https://aclanthology.org/2021.acl-long.141>.
- Wenliang Dai, Junnan Li, D Li, AMH Tiong, J Zhao, W Wang, B Li, P Fung, and S Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. arxiv 2023. *arXiv preprint arXiv:2305.06500*, 2, 2023.
- Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. Words or vision: Do vision-language models have blind faith in text?, 2025a.
- Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025b.

- Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Open-vlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. *arXiv preprint arXiv:2503.17352*, 2025c.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL <https://aclanthology.org/N19-1423>.
- Ning Ding, Yulin Chen, Xu Han, Guangwei Xu, Pengjun Xie, Hai-Tao Zheng, Zhiyuan Liu, Juanzi Li, and Hong-Gee Kim. Prompt-learning for fine-grained entity typing. *arXiv preprint arXiv:2108.10604*, 2021.
- Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, et al. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. *arXiv preprint arXiv:2404.06512*, 2024.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025.
- Yu Feng, Ben Zhou, Haoyu Wang, Helen Jin, and Dan Roth. Generic temporal reasoning with differential analysis and explanation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12013–12029, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.671. URL <https://aclanthology.org/2023.acl-long.671>.

- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *NeurIPS Datasets and Benchmarks Track*, 2025a.
- Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *CVPR*, 2025b.
- Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*, 2024.
- Xingyu Fu, Minqian Liu, Zhengyuan Yang, John Corring, Yijuan Lu, Jianwei Yang, Dan Roth, Dinei Florencio, and Cha Zhang. Refocus: Visual editing as a chain of thought for structured image understanding. *arXiv preprint arXiv:2501.05452*, 2025c.
- Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.295. URL <https://aclanthology.org/2021.acl-long.295>.
- Vedant Gaur and Nikunj Saunshi. Reasoning in large language models through symbolic math word problems. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5889–5903, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.364. URL <https://aclanthology.org/2023.findings-acl.364>.
- Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

- Xinyu Geng, Peng Xia, Zhen Zhang, Xinyu Wang, Qiuchen Wang, Ruixue Ding, Chenxi Wang, Jialong Wu, Yida Zhao, Kuan Li, et al. Webwatcher: Breaking new frontiers of vision-language deep research agent. *arXiv preprint arXiv:2508.05748*, 2025.
- Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021.
- Dan Gillick, Nevena Lazic, Kuzman Ganchev, Jesse Kirchner, and David Huynh. Context-dependent fine-grained entity type tagging. *arXiv preprint arXiv:1412.1820*, 2014.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018.
- Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.
- Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yueze Wang, Tiejun Huang, and Bo Zhao. Efficient multimodal learning from data-centric perspective. *arXiv preprint arXiv:2402.11530*, 2024.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

- Minjie Hong, Zirun Guo, Yan Xia, Zehan Wang, Ziang Zhang, Tao Jin, and Zhou Zhao. Apo: Enhancing reasoning ability of mllms via asymmetric policy optimization. *arXiv preprint arXiv:2506.21655*, 2025.
- Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. 2025. URL <https://arxiv.org/abs/2501.13826>.
- Yanfeng Hu, Xue Qiao, Luo Xing, and Chen Peng. Diversified semantic attention model for fine-grained entity typing. *IEEE Access*, 9:2251–2265, 2020.
- Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *arXiv preprint arXiv:2406.09403*, 2024.
- Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427, 2024.
- Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models, 2025a.
- Zeyi Huang, Yuyang Ji, Anirudh Sundara Rajan, Zefan Cai, Wen Xiao, Haohan Wang, Junjie Hu, and Yong Jae Lee. Visualtoolagent (vista): A reinforcement learning framework for visual tool selection. *arXiv preprint arXiv:2505.20289*, 2025b.
- Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

- Chaoya Jiang, Yongrui Heng, Wei Ye, Han Yang, Haiyang Xu, Ming Yan, Ji Zhang, Fei Huang, and Shikun Zhang. Vlm-r 3: Region recognition, reasoning, and refinement for enhanced multimodal chain-of-thought. *arXiv preprint arXiv:2505.16192*, 2025.
- Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. TinyBERT: Distilling BERT for natural language understanding. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.372. URL <https://aclanthology.org/2020.findings-emnlp.372/>.
- Hailong Jin, Lei Hou, Juanzi Li, and Tiansi Dong. Fine-grained entity typing via hierarchical multi graph convolutional networks. 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 4969–4978, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1502. URL <https://aclanthology.org/D19-1502>.
- Mitchell Koch, John Gilmer, Stephen Soderland, and Daniel S. Weld. Type-aware distantly supervised relation extraction with linked arguments. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1891–1901, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1203. URL <https://aclanthology.org/D14-1203>.
- Kirby Kuznia, Swaroop Mishra, Mihir Parmar, and Chitta Baral. Less is more: Summary of long instructions is better for program synthesis. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4532–4552, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.301. URL <https://aclanthology.org/2022.emnlp-main.301>.
- Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36, 2024.

- Bangzheng Li, Wenpeng Yin, and Muhao Chen. Ultra-fine entity typing with indirect supervision from natural language inference. *Transactions of the Association for Computational Linguistics*, 10:607–622, 2022.
- Bangzheng Li, Ben Zhou, Fei Wang, Xingyu Fu, Dan Roth, and Muhao Chen. Deceptive semantic shortcuts on reasoning chains: How far can models go without hallucination? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7675–7688, 2024a.
- Bangzheng Li, Fei Wang, Wenxuan Zhou, Nan Xu, Ben Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. Semantic-clipping: Efficient vision-language modeling with semantic-guided visual selection. *arXiv preprint arXiv:2503.11794*, 2025a.
- Bangzheng Li, Jianmo Ni, Chen Qu, Ian Miao, Liu Yang, Xingyu Fu, Muhao Chen, and Derek Zhiyuan Cheng. Reinforced attention learning, 2026a. URL <https://arxiv.org/abs/2602.04884>.
- Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL <https://openreview.net/forum?id=j84WR50RsC>.
- Bo Li, Peiyuan Zhang, Jingkan Yang, Yuanhan Zhang, Fanyi Pu, and Ziwei Liu. Otterhd: A high-resolution multi-modality model. *arXiv preprint arXiv:2311.04219*, 2023a.
- Bo Li, Peiyuan Zhang, Kaichen Zhang, Fanyi Pu, Xinrun Du, Yuhao Dong, Haotian Liu, Yuanhan Zhang, Ge Zhang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Accelerating the development of large multimodal models, 2024b. URL <https://github.com/EvolvingLMMS-Lab/lmms-eval>.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024c.

- Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023b.
- Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought, 2025b. URL <https://arxiv.org/abs/2501.07542>.
- Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024d.
- Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024e.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore, December 2023c. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.20. URL <https://aclanthology.org/2023.emnlp-main.20/>.
- Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26763–26773, 2024f.
- Xiao Ling and Daniel S Weld. Fine-grained entity recognition. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.
- Xiao Ling, Sameer Singh, and Daniel S. Weld. Design challenges for entity linking. *Transactions of the Association for Computational Linguistics*, 3:315–328, 2015. doi: 10.1162/tacl_a_00141. URL <https://aclanthology.org/Q15-1023>.

- Dairu Liu, Ziyue Wang, Minyuan Ruan, Fuwen Luo, Chi Chen, Peng Li, and Yang Liu. Visual Abstract Thinking Empowers Multimodal Reasoning, May 2025a. URL <http://arxiv.org/abs/2505.20164>.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc., 2023.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024a.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024b. URL <https://llava-vl.github.io/blog/2024-01-30-llava-next/>.
- Qing Liu, Hongyu Lin, Xinyan Xiao, Xianpei Han, Le Sun, and Hua Wu. Fine-grained entity typing via label reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4611–4622, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.378. URL <https://aclanthology.org/2021.emnlp-main.378>.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024c.
- Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*, 2024d.

- Zeyu Liu, Yuhang Liu, Guanghao Zhu, Congkai Xie, Zhen Li, Jianbo Yuan, Xinyao Wang, Qing Li, Shing-Chi Cheung, Shengyu Zhang, et al. Infi-mmr: Curriculum-based unlocking multimodal reasoning via phased reinforcement learning in multimodal small language models. *arXiv preprint arXiv:2505.23091*, 2025b.
- Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025c.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. URL <https://openreview.net/pdf?id=Bkg6RiCqY7>.
- Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.
- Tengchao Lv, Yupan Huang, Jingye Chen, Yuzhong Zhao, Yilin Jia, Lei Cui, Shuming Ma, Yaoyao Chang, Shaohan Huang, Wenhui Wang, et al. Kosmos-2.5: A multimodal literate model. *arXiv preprint arXiv:2309.11419*, 2023.
- Yunze Man, De-An Huang, Guilin Liu, Shiwei Sheng, Shilong Liu, Liang-Yan Gui, Jan Kautz, Yu-Xiong Wang, and Zhiding Yu. Argus: Vision-centric reasoning with grounded chain-of-thought. In *CVPR*, 2025.
- Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.177. URL <https://aclanthology.org/2022.findings-acl.177>.
- Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhui Wang, Junjun He, Kaipeng Zhang, et al. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *CoRR*, 2025.

- Swaroop Mishra, Matthew Finlayson, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, Tanmay Rajpurohit, Oyvind Tafjord, Ashish Sabharwal, Peter Clark, and Ashwin Kalyan. LILA: A unified benchmark for mathematical reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5807–5832, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.392. URL <https://aclanthology.org/2022.emnlp-main.392>.
- Minheng Ni, Zhengyuan Yang, Linjie Li, Chung-Ching Lin, Kevin Lin, Wangmeng Zuo, and Lijuan Wang. Point-rft: Improving multimodal reasoning with visually grounded reinforcement finetuning. *arXiv preprint arXiv:2505.19702*, 2025.
- Rasha Obeidat, Xiaoli Fern, Hamed Shahbazi, and Prasad Tadepalli. Description-based zero-shot fine-grained entity typing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 807–814, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1087. URL <https://aclanthology.org/N19-1087>.
- Yasumasa Onoe and Greg Durrett. Learning to denoise distantly-labeled data for entity typing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2407–2417, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1250. URL <https://aclanthology.org/N19-1250>.
- Yasumasa Onoe and Greg Durrett. Interpretable entity representations through large-scale typing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 612–624, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.54. URL <https://aclanthology.org/2020.findings-emnlp.54>.
- Yasumasa Onoe, Michael Boratko, Andrew McCallum, and Greg Durrett. Modeling fine-grained entity types with box embeddings. In *Proceedings of the 59th Annual Meeting of the*

Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 2051–2064, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.160. URL <https://aclanthology.org/2021.acl-long.160>.

OpenAI. Gpt-4 technical report, 2023.

Timothy Ossowski and Junjie Hu. Retrieving multimodal prompts for generative visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2518–2535, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.158. URL <https://aclanthology.org/2023.findings-acl.158>.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1244. URL <https://aclanthology.org/D16-1244>.

Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*, 2025.

Pouya Pezeshkpour, Moin Aminnaseri, and Estevam Hruschka. Mixed signals: Decoding vlms’ reasoning and underlying bias in vision-language conflict, 2025.

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.

- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL <http://jmlr.org/papers/v21/20-074.html>.
- Xiang Ren, Wenqi He, Meng Qu, Clare R Voss, Heng Ji, and Jiawei Han. Label noise reduction in entity typing by heterogeneous partial-label embedding. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1825–1834, 2016.
- Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets, 2015. URL <https://arxiv.org/abs/1412.6550>.
- Oscar Sainz, Oier Lopez de Lacalle, Gorka Labaka, Ander Barrena, and Eneko Agirre. Label verbalization and entailment for effective zero and few-shot relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1199–1212, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.92. URL <https://aclanthology.org/2021.emnlp-main.92>.
- Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models, 2024a.

- Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models, 2024b.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024c.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024d.
- Chuming Shen, Wei Wei, Xiaoye Qu, and Yu Cheng. Satori-r1: Incentivizing multimodal reasoning with spatial grounding and verifiable rewards. *arXiv preprint arXiv:2505.19094*, 2025a.
- Jiaming Shen, Wenda Qiu, Yu Meng, Jingbo Shang, Xiang Ren, and Jiawei Han. TaxoClass: Hierarchical multi-label text classification using only class names. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4239–4249, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.335. URL <https://aclanthology.org/2021.naacl-main.335>.
- Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. Codi: Compressing chain-of-thought into continuous space via self-distillation. *arXiv preprint arXiv:2502.21074*, 2025b.
- Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297, 2025.
- Baifeng Shi, Ziyang Wu, Maolin Mao, Xin Wang, and Trevor Darrell. When do we not need larger vision models? In *European Conference on Computer Vision*, pages 444–462. Springer, 2025.

- Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022.
- Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhui Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 2025a.
- Zhaochen Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, et al. Openthinking: Learning to think with images via visual tool reinforcement learning. *arXiv preprint arXiv:2505.08617*, 2025b.
- Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, et al. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*, 2025c.
- Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient knowledge distillation for bert model compression, 2019. URL <https://arxiv.org/abs/1908.09355>.
- Didac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128*, 2023.
- Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning. *arXiv preprint arXiv:2503.20752*, 2025.
- Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. doi: 10.48550/arXiv.2405.09818. URL <https://github.com/facebookresearch/chameleon>.
- Gemini Team. Gemini: A family of highly capable multimodal models, 2023.
- Gemini Team. Gemini: A family of highly capable multimodal models, 2025. URL <https://arxiv.org/abs/2312.11805>.

- Rahul Thapa, Kezhen Chen, Ian Covert, Rahul Chalamala, Ben Athiwaratkun, Shuaiwen Leon Song, and James Zou. Dragonfly: Multi-resolution zoom-in encoding enhances vision-language models, 2024.
- Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms, 2024a.
- Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024b.
- Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms, 2024c.
- Harsh Trivedi, Niranjana Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL <https://aclanthology.org/2022.tacl-1.31>.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.
- Luke Vilnis, Xiang Li, Shikhar Murty, and Andrew McCallum. Probabilistic embedding of knowledge graphs with box lattice measures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 263–272, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1025. URL <https://aclanthology.org/P18-1025>.
- Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*, 2024a.

- Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhui Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*, 2025a.
- Jiacong Wang, Zijian Kang, Haochen Wang, Haiyong Jiang, Jiawen Li, Bohong Wu, Ya Wang, Jiao Ran, Xiao Liang, Chao Feng, et al. Vgr: Visual grounded reasoning. *arXiv preprint arXiv:2506.11991*, 2025b.
- Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1405–1418, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.121. URL <https://aclanthology.org/2021.findings-acl.121>.
- Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. Lvbench: An extreme long video understanding benchmark, 2024b.
- Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, et al. Visualprm: An effective process reward model for multimodal reasoning. *arXiv preprint arXiv:2503.10291*, 2025c.
- Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025d.
- Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788, 2020.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

- Zhaochen Wang, Bryan Hooi, Yiwei Wang, Ming-Hsuan Yang, Zi Huang, and Yujun Cai. Text speaks louder than vision: ASCII art reveals textual biases in vision-language models. In *Second Conference on Language Modeling*, 2025e. URL <https://openreview.net/forum?id=naEyNVTLsh>.
- Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, et al. Perception-aware policy optimization for multimodal reasoning. *arXiv preprint arXiv:2507.06448*, 2025f.
- Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, et al. Perception-aware policy optimization for multimodal reasoning. *arXiv preprint arXiv:2507.06448*, 2025g.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.
- Lai Wei, Yuting Li, Kaipeng Zheng, Chen Wang, Yue Wang, Linghe Kong, Lichao Sun, and Weiran Huang. Advancing multimodal reasoning via reinforcement learning with cold start. *arXiv preprint arXiv:2505.22334*, 2025a.
- Yana Wei, Liang Zhao, Kangheng Lin, En Yu, Yuang Peng, Runpei Dong, Jianjian Sun, Haoran Wei, Zheng Ge, Xiangyu Zhang, et al. Perception in reflection. *arXiv preprint arXiv:2504.07165*, 2025b.
- Aaron Steven White, Pushpendre Rastogi, Kevin Duh, and Benjamin Van Durme. Inference is everything: Recasting semantic resources into a unified evaluation framework. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 996–1005, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing. URL <https://aclanthology.org/I17-1100>.
- Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

- Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL <https://aclanthology.org/N18-1101>.
- Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding, 2024. URL <https://arxiv.org/abs/2407.15754>.
- Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. *arXiv preprint arXiv:2506.09965*, 2025a.
- Mingyuan Wu, Jingcheng Yang, Jize Jiang, Meitang Li, Kaizhuo Yan, Hanchao Yu, Minjia Zhang, Chengxiang Zhai, and Klara Nahrstedt. Vtool-r1: Vlms learn to think with images via reinforcement learning on multimodal tool use. *arXiv preprint arXiv:2505.19255*, 2025b.
- Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. *arXiv preprint arXiv:2312.14135*, 2023.
- Jiaer Xia, Yuhang Zang, Peng Gao, Yixuan Li, and Kaiyang Zhou. Visionary-r1: Mitigating shortcuts in visual reasoning with reinforcement learning. *arXiv preprint arXiv:2505.14677*, 2025.
- Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9777–9786, June 2021.
- Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025.
- Wenhan Xiong, Jiawei Wu, Deren Lei, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. Imposing label-relational inductive bias for extremely fine-grained entity typing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short*

- Papers*), pages 773–784, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1084. URL <https://aclanthology.org/N19-1084>.
- Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024.
- Peng Xu and Denilson Barbosa. Neural fine-grained entity type classification with hierarchy-aware loss. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 16–25, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1002. URL <https://aclanthology.org/N18-1002>.
- Yi Xu, Chengzu Li, Han Zhou, Xingchen Wan, Caiqi Zhang, Anna Korhonen, and Ivan Vulić. Visual planning: Let’s think only with images, 2025. URL <https://arxiv.org/abs/2505.11409>.
- Zenan Xu, Daya Guo, Duyu Tang, Qinliang Su, Linjun Shou, Ming Gong, Wanjuan Zhong, Xiaojun Quan, Daxin Jiang, and Nan Duan. Syntax-enhanced pre-trained model. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5412–5422, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.420. URL <https://aclanthology.org/2021.acl-long.420>.
- Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025a.
- Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. 2025b. URL <https://arxiv.org/abs/2506.17218>.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop

question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium, October–November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL <https://aclanthology.org/D18-1259>.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of Thoughts: Deliberate problem solving with large language models, 2023.

Semih Yavuz, Izzeddin Gur, Yu Su, Mudhakar Srivatsa, and Xifeng Yan. Improving semantic parsing via answer type inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 149–159, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1015. URL <https://aclanthology.org/D16-1015>.

Wenpeng Yin, Jamaal Hay, and Dan Roth. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. 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 3914–3923, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1404. URL <https://aclanthology.org/D19-1404>.

Wenpeng Yin, Nazneen Fatema Rajani, Dragomir Radev, Richard Socher, and Caiming Xiong. Universal natural language processing with limited annotations: Try few-shot textual entailment as a start. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8229–8239, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.660. URL <https://aclanthology.org/2020.emnlp-main.660>.

Dani Yogatama, Daniel Gillick, and Nevena Lazic. Embedding methods for fine grained entity type classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 291–296, Beijing, China, July

2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-2048. URL <https://aclanthology.org/P15-2048>.
- En Yu, Kangheng Lin, Liang Zhao, Jisheng Yin, Yana Wei, Yuang Peng, Haoran Wei, Jianjian Sun, Chunrui Han, Zheng Ge, et al. Perception-rl: Pioneering perception policy with reinforcement learning. *arXiv preprint arXiv:2504.07954*, 2025.
- Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhui Chen, and Graham Neubig. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024.
- Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023.
- Guanghao Zhang, Tao Zhong, Yan Xia, Zhelun Yu, Haoyuan Li, Wangui He, Fangxun Shu, Mushui Liu, Dong She, Yi Wang, et al. Cmmcot: Enhancing complex multi-image comprehension via multi-modal chain-of-thought and memory augmentation. *arXiv preprint arXiv:2503.05255*, 2025a.
- Jingyi Zhang, Jiaying Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*, 2025b.
- Qizhe Zhang, Aosong Cheng, Ming Lu, Renrui Zhang, Zhiyong Zhuo, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. Beyond text-visual attention: Exploiting visual cues for effective token pruning in vlms. *arXiv preprint arXiv:2412.01818*, 2025c.
- Xintong Zhang, Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaowen Zhang, Yang Liu, Tao Yuan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, et al. Chain-of-focus: Adaptive visual search and zooming for multimodal reasoning via rl. *arXiv preprint arXiv:2505.15436*, 2025d.

- Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Lllavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023a.
- Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruva Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are visually-grounded language models bad at image classification? *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.
- Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1139. URL <https://aclanthology.org/P19-1139>.
- Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multi-modal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023b.
- Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.
- Ben Zhou, Kyle Richardson, Xiaodong Yu, and Dan Roth. Learning to decompose: Hypothetical question decomposition based on comparable texts. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2223–2235, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.142. URL <https://aclanthology.org/2022.emnlp-main.142>.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.
- Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao,

Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL <https://arxiv.org/abs/2504.10479>.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *Advances in Neural Information Processing Systems*, 36, 2024.