
Data Efficient Deep Multimodal Learning



Shen Meng

College of Computing and Data Science

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy (Ph.D.)

2025

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

01/24/2025

.....

Date

NTU NTU NTU NTU NTU NTU NTU NTU
NTU NTU NTU NTU NTU NTU NTU NTU
NTU NTU *Shen Meng* NTU NTU NTU NTU
NTU NTU NTU NTU NTU NTU NTU NTU
.....

Shen Meng

Authorship Attribution Statement

This thesis contains material from three papers published/submitted in the following peer-reviewed conferences in which I am listed as an author.

Chapter 3 is published as **Meng Shen**, Yizheng Huang, Jianxiong Yin, Heqing Zou, Deepu Rajan, and Simon See, “Towards Balanced Active Learning for Multimodal Classification”. In Proceedings of the 31st ACM International Conference on Multimedia (MM '23).

The contributions of the co-authors are as follows:

- I proposed the idea of this paper, designed the algorithm and wrote the manuscript.
- Yizheng Huang helped design more solid experiments to validate the algorithms.
- Prof. Deepu Rajan and Heqing Zou provided insightful comments and helped revise the manuscript.
- Jianxiong Yin and Prof. Simon See provided computing resources to help finish experiments.

Chapter 4 is published as **Meng Shen**, Yake Wei, Jianxiong Yin, Deepu Rajan, Di Hu, and Simon See, “Enhancing Modality Representation and Alignment for Multimodal Cold-start Active Learning”. In Proceedings of the 6th ACM International Conference on Multimedia in Asia (MM Asia '24).

The contributions of the co-authors are as follows:

- I found the research problem, developed the algorithm, conducted the experiments and wrote the manuscript.
- Yake Wei provided critical comments on the algorithm design and helped revise the manuscript.
- Prof. Deepu Rajan and Prof. Di Hu provided useful comments on the paper and revised the introduction.
- Yin Jianxiong and Prof. Simon See provided computing resources to help finish experiments.

01/24/2025

.....

Date

NTU NTU NTU NTU NTU NTU NTU NTU
NTU NTU NTU NTU NTU NTU NTU NTU
Shen Meng
NTU NTU NTU NTU NTU NTU NTU NTU
NTU NTU NTU NTU NTU NTU NTU NTU
.....

Shen Meng

Acknowledgements

I would like to express my deepest gratitude to my advisor, Prof. Deepu Rajan, for your guidance, advice, and support throughout my PhD journey at NTU. Your advice was instrumental in shaping my research work and expanding my thinking. I am always thankful for your patience which allowed me to overcome challenges and finish my thesis.

I am also sincerely thankful to Prof. Chng Eng Siong for your encouragement and support so that I can restart and finish my study. It was your firm support that made this journey significantly smoother.

I am grateful to my friends and coauthors, whose assistance and collaboration made this process more enriching. Your support during this period has meant a great deal to me.

To my parents, thank you for your encouragement and belief in me. Your support has been a source of strength throughout this journey.

Finally, to my wife, I owe the deepest appreciation. It is your love, patience, and companionship that carried me through the toughest times. Your presence has been my greatest source of strength and motivation, and I could not have completed this PhD study without you.

Abstract

Multimodal learning, which enables neural networks to process and integrate information from various sensory modalities such as vision, language, and sound, has become increasingly important in applications ranging from affective computing and healthcare to advanced multimodal chatbots. Despite its potential, multimodal learning faces significant challenges, particularly in the area of data efficiency. The requirement for large, high-quality datasets from multiple modalities presents a substantial barrier, limiting the scalability and accessibility of large multimodal models. This dissertation addresses several key issues in data-efficient deep multimodal learning, focusing on the imbalanced multimodal data selection, the cold-start problem in multimodal active learning, and the mitigation of hallucinations in large vision-language models.

Firstly, we analyze the limitations of conventional active learning strategies, which tend to favor dominant modalities, leading to unbalanced multimodal models that neglect weaker modalities. To overcome this, we propose a gradient embedding modulation method that ensures a more equitable data selection process across modalities, resulting in models that fairly utilize both weak and strong modalities.

Building on our work in warm-start active learning, we tackle the cold-start problem in multimodal active learning, where no initial labels are available for warm-start data selection. We develop a two-stage approach that first reduces the modality representation gap through multimodal self-supervised learning, utilizing unimodal prototypes to harmonize representations across modalities. In the subsequent data selection stage, we introduce a regularization term to maximize modality alignment, leading to improved model performance using the same amount of data compared to existing methods.

Extending our focus from data selection to the usage of training data, we address the challenge of hallucinations in large vision-language models, where the models generate content that is incorrect in the context of input images. We investigate

the relationship between hallucinations and visual dependence of tokens, revealing that certain tokens contribute disproportionately to these hallucinations. Based on this insight, we propose an approach that adjusts training weights according to the visual dependence of tokens, thereby reducing the hallucination rate without requiring additional training data or inference costs.

The contributions of this thesis offer significant advancements in the field of data-efficient multimodal learning. By developing novel methods for balancing multimodal data selection, addressing cold-start problem in multimodal active learning, and mitigating hallucinations in large vision-language models, this work paves the way for more practical and scalable multimodal learning systems that require less data and computational effort while achieving superior performance.

Contents

| | |
|---|--------------|
| Acknowledgements | ix |
| Abstract | xi |
| List of Figures | xvii |
| List of Tables | xxi |
| Symbols and Acronyms | xxiii |
| 1 Introduction | 1 |
| 1.1 Background and Motivation | 1 |
| 1.2 Objectives | 3 |
| 1.3 Contributions | 3 |
| 1.4 Thesis Outline | 4 |
| 2 Literature Review | 5 |
| 2.1 Deep Multimodal Learning | 5 |
| 2.2 Data Efficiency | 7 |
| 2.2.1 Labeling Data Efficiency | 7 |
| 2.2.2 Data Usage Efficiency | 8 |
| 2.3 Data-efficient Multimodal Learning | 9 |
| 2.3.1 Active Learning | 9 |
| 2.3.2 Balanced Multimodal Learning | 12 |
| 2.3.3 Multimodal Self-supervised Learning | 13 |
| 2.3.4 Hallucination Mitigation in LVLMS | 14 |
| 3 Towards Balanced Active Learning for Multimodal Classification | 17 |
| 3.1 Introduction | 17 |
| 3.2 Methodology | 19 |
| 3.2.1 Multimodal Learning Framework | 19 |
| 3.2.2 Analysis of Imbalance in AL | 19 |
| 3.2.3 Guidelines to Design Balanced MMAL | 21 |
| 3.2.4 Estimate Modality Contribution | 23 |

| | | |
|----------|--|-----------|
| 3.2.5 | Proposed Method | 24 |
| 3.3 | Experiment | 26 |
| 3.3.1 | Dataset | 26 |
| 3.3.2 | Baseline | 26 |
| 3.3.3 | Experiment Setting | 27 |
| 3.3.4 | AL Performance | 28 |
| 3.3.5 | Large-scale Active Learning | 33 |
| 3.3.6 | Ablation Studies | 34 |
| 3.4 | Chapter Summary | 37 |
| 4 | Enhancing Modality Representation and Alignment for Multi-modal Cold-start Active Learning | 39 |
| 4.1 | Introduction | 39 |
| 4.2 | Proposed Method | 42 |
| 4.2.1 | Multimodal Cold-start Active Learning Framework | 42 |
| 4.2.2 | Multimodal Contrastive Learning with Uni-modal Prototypes | 43 |
| 4.2.3 | Select Data with Good Modality Alignment for Cold-start AL | 46 |
| 4.3 | Experiments | 50 |
| 4.3.1 | Dataset | 50 |
| 4.3.2 | Baseline | 51 |
| 4.3.3 | Experiment Setting | 54 |
| 4.3.4 | Supervised Cold-start AL | 54 |
| 4.3.5 | Semi-supervised Cold-start AL | 55 |
| 4.3.6 | Ablation Study of λ_{align} | 56 |
| 4.4 | Data Selection Preference of AL | 57 |
| 4.4.1 | Rank the Samples by Confidence | 57 |
| 4.4.2 | Analyze the Data Preference of AL | 59 |
| 4.5 | Implementation Details | 60 |
| 4.5.1 | Multimodal Self-supervised Learning | 60 |
| 4.5.2 | Supervised Cold-start AL | 61 |
| 4.5.3 | Semi-supervised Cold-start AL | 61 |
| 4.5.4 | Optimize the Parametric Selection Model | 62 |
| 4.6 | Chapter Summary | 62 |
| 5 | Hallucination Reduction in Large Vision-language Models via Emphasizing Image-negative Tokens | 63 |
| 5.1 | Introduction | 63 |
| 5.2 | Analysis of Hallucination in LVLMS | 65 |
| 5.2.1 | Visual Dependence | 66 |
| 5.2.2 | Hallucinations in Generated Text Tokens | 67 |
| 5.2.3 | Information Flow across Tokens | 68 |
| 5.3 | Adjust Loss Weights During Training | 71 |
| 5.4 | Experiments | 73 |

| | | |
|----------|--|------------|
| 5.4.1 | Experiment Setup | 73 |
| 5.4.2 | Baselines | 74 |
| 5.4.3 | Main Results | 74 |
| 5.4.4 | Further Training Fine-tuned LVLMS | 78 |
| 5.4.5 | Filtering Data by Visual Dependence | 78 |
| 5.4.6 | The Distribution of Visual Dependence | 79 |
| 5.4.7 | Ablation Studies of Hyperparameters | 79 |
| 5.4.8 | Performance on MME and POPE | 82 |
| 5.4.9 | Experiment Settings and Computational Cost | 82 |
| 5.5 | Chapter Summary | 84 |
| 6 | Conclusions and Future Work | 85 |
| 6.1 | Conclusions | 85 |
| 6.2 | Future Work | 87 |
| A | Appendix of Chapter 3 | 89 |
| B | Appendix of Chapter 5 | 93 |
| | List of Publications | 101 |
| | Bibliography | 103 |

List of Figures

| | | |
|-----|--|----|
| 2.1 | Multimodal learning latent space representation. The image modality and text modality are processed and mapped into a latent space where modality-specific information is preserved. Additionally, modality-shared information from the two different modalities is extracted to produce richer latent representations compared to uni-modal representations. | 6 |
| 2.2 | General active learning process. The dashed lines represent model training. The solid lines represent data selection. | 10 |
| 3.1 | Modality contribution Φ across different AL iterations on the Food101 test set. | 28 |
| 3.2 | Modality contribution Φ across different AL iterations on the KineticsSound test set. | 28 |
| 3.3 | Performance comparison between proposed method and other conventional AL strategies with concatenation fusion method. The metric selected is top-1 accuracy (Top-1) on multimodal and unimodal classification. | 29 |
| 3.4 | Pairwise comparison of all active learning strategies. Each element in the matrix $P_{i,j}$ represents the number of times strategy i outperforms strategy j . A strategy is considered better if its row-wise value is larger, indicating that it beats other strategies more often. On the other hand, a strategy is better if its column-wise value is smaller, meaning it is rarely beaten by other strategies. The maximum value of each cell is 5, which is the total number of experimental settings. The bottom row displays the column-wise average values (lower is better). | 31 |
| 3.5 | Average weight for the weaker modality in a sub-dataset dominated by the other stronger modality. | 32 |
| 3.6 | Multimodal and unimodal classification performance comparison with NL-gate fusion method on the VGGSound dataset. | 34 |
| 3.7 | Top 10 improved and dropped classes based on the improvement of BMMAL to BADGE on multimodal classification accuracy on KineticsSound with 5K labeled samples. Bars represent multimodal classification accuracy. Stems represent unimodal classification accuracy. | 35 |

| | | |
|-----|--|----|
| 3.8 | Performance comparison between proposed method and other conventional AL strategies with Summation fusion method. The metric selected is top-1 accuracy (Top-1) on mulitmodal and unimodal classification. | 36 |
| 4.1 | Multimodal Cold-start Active Learning Framework (using video-audio data as an example): In Stage 1, multimodal self-supervised learning is applied to a large unlabeled dataset to derive concatenated multimodal representations from unimodal features. In Stage 2, a data selection strategy samples informative multimodal data pairs, which are then annotated by human oracles. The labeled samples are subsequently used to train a downstream task model. | 40 |
| 4.2 | Our method (use audio/video as an example): In Stage 1, we employ uni-modal prototypical loss and cross-modal contrastive loss; In Stage 2, our selection reduces distribution gap while maintaining diversity and modality alignment. | 43 |
| 4.3 | Preference for data selection of different AL strategies with 5% labeling budget on Food101. | 58 |
| 4.4 | Preference for data selection of different AL strategies with 5% labeling budget on KineticsSound. | 58 |
| 4.5 | Preference for data selection of different AL strategies with 5% labeling budget on VGGSound. | 59 |
| 5.1 | The left figure displays the input image and the middle figure is the noisy image with diffusion noise at time step 900. The model (fine-tuned LLaVA-v1.5-7b) is asked to generate a descriptive response given the image. The right figure shows the probability of each token y_t when provided with a clean image v and a noisy image v^* . The image-positive , image-invariant and image-negative tokens are marked as red , black and green | 66 |
| 5.2 | The number of grounded objects and hallucinated objects in different groups of tokens categorized by visual dependence. Tokens with lower visual dependence tend to be more image-negative, while tokens with higher visual dependence are more likely to be image-positive. | 68 |
| 5.3 | The significance of information flow per attention layer from image tokens and text tokens to different target tokens. | 69 |
| 5.4 | The saliency scores of each token in the attention matrix of LLaVA-v1.5-7b of example from Figure 1. | 70 |
| 5.5 | The hallucinated content are often found to be near the image-invariant tokens. | 71 |
| 5.6 | The distribution of the sum of visual dependence of each sample from LLaVA-Instruct 150k. | 80 |

| | | |
|-----|--|-----|
| A.1 | The implementation of NL-gate. We use the 3D video feature as query and the 2D concatenated audio and video feature as key and value. | 91 |
| B.1 | Qualitative results of LLaVA-v1.5-7b trained with vanilla loss baseline and ours approach with w_{neg} on LLaVA-Instruct 150k dataset. Hallucinations are marked as red. | 94 |
| B.2 | Qualitative results of LLaVA-v1.5-7b trained with vanilla loss baseline and ours approach with w_{neg} on LLaVA-Instruct 150k dataset. Hallucinations are marked as red. | 95 |
| B.3 | Qualitative results of PaliGemma-3b trained with vanilla loss baseline and ours approach with w_{neg} on LLaVA-Details 23k dataset. Hallucinations are marked as red. | 96 |
| B.4 | Qualitative results of Bunny-v1.1-8b trained with vanilla loss baseline and ours approach with w_{pos} and w_{neg} on SVIT-Details 71k dataset. Hallucinations are marked as red, the correct details that are not discovered in baseline are marked green. | 97 |
| B.5 | Qualitative results of Bunny-v1.1-8b trained with vanilla loss and ours approach with w_{pos} and w_{neg} on SVIT-Details 71k dataset. Hallucinations are marked as red, the correct details that are not discovered in baseline are marked green. | 98 |
| B.6 | The top left sub-figure is the input image, and the top right sub-figure is the generated image description from LLaVA-v1.5-7b fine-tuned from scratch with LLaVA-Details 23k dataset. The bottom bar charts display the probability of each token with clean image input v and with noisy image input v^* . The hallucinated objects are marked red, the grounded objects are marked green. | 99 |
| B.7 | The top left sub-figure is the input image, and the top right sub-figure is the generated image description from LLaVA-v1.5-7b fine-tuned from scratch with LLaVA-Details 23k dataset. The bottom bar charts display the probability of each token with clean image input v and with noisy image input v^* . The hallucinated objects are marked red, the grounded objects are marked green. | 100 |

List of Tables

| | | |
|-----|--|----|
| 3.1 | AL performance on VGGSound dataset with budget size of 5,000. The best results are highlight in bold. | 34 |
| 4.1 | The modality gap and the expectation of average cosine uni-modal and cross-modal similarity of all samples in three different multi-modal datasets. <i>w/o</i> represents without prototypes, and <i>w/</i> represents with prototypes. | 47 |
| 4.2 | The supervised cold-start AL experiment results. We report the mean and std of Top-1 accuracy with multiple runs. The best is marked bold , the second best is underlined. <i>w/o</i> and <i>w/</i> represent without and with uni-modal prototypes. Red ↑ and blue ↓ represent the performance is improved or degraded after integrating uni-modal prototypes. | 52 |
| 4.3 | The semi-supervised cold-start AL results. We report the mean and std of Top-1 accuracy with multiple runs. | 53 |
| 4.4 | The ablation study of λ_{align} | 56 |
| 5.1 | Main results of three variants of LVLMS: LLaVA-v1.5 (7b and 13b), PaliGemma-3b and Bunny-v1.1-8b on CHAIR and FaithScore (Faith). <i>The rows in italics</i> are excluded from the comparison because their object recognition recall is significantly lower than the others, making the comparison unfair. The best results are marked as bold and the second best are <u>underlined</u> | 75 |
| 5.2 | The experimental results of continue fine-tuning of instructional tuned models of three variants of LVLMS: LLaVA-v1.5 (7b and 13b), PaliGemma-3b and Bunny-v1.1-8b on CHAIR benchmark. | 77 |
| 5.3 | The experimental results of data filtering. Data Len. represents the average data sentence length of filtered data, and Response Len. represents the average response length. | 79 |
| 5.4 | The performance on CHAIR with different temperature τ in fine-tuning pre-trained LLaVA-v1.5-7b and PaliGemma-3b with LLaVA-Details 23k dataset. | 81 |
| 5.5 | The performance on CHAIR with different start times of loss re-weighting in fine-tuning pre-trained LLaVA-v1.5-7b and PaliGemma-3b with LLaVA-Details 23k dataset. For example, 0% means that starting loss re-weighting from the beginning. | 81 |

| | | |
|-----|---|----|
| 5.6 | The performance on CHAIR with different noise step T in fine-tuning pre-trained LLaVA-v1.5-7b and PaliGemma-3b with LLaVA-Details 23k dataset. | 82 |
| 5.7 | The performance of LLaVA-v1.5-7b on MME and POPE of vanilla MLE loss, selective EOS supervision and our approach of emphasizing image-negative tokens. MME_P and MME_C represent the perception and cognition scores. $POPE_{Acc}$ and $POPE_{F1}$ are the average accuracy and f1score across three POPE subsets: random, popular and adversarial. | 83 |
| 5.8 | The hyper-parameters for experiment of fine-tuning LVLMs from scratch. | 83 |
| 5.9 | Computational cost for each experiment. | 84 |

Symbols and Acronyms

Symbols

| | |
|----------------|--|
| X^L | input data in the labeled dataset |
| X^S | input data in the selected dataset |
| X^U | input data in the unlabelled dataset |
| Y^L | labels in the labeled dataset |
| W | the linear layer weight matrix |
| A | the attention weight matrix |
| $\ \cdot\ $ | the l2-norm of a vector or matrix in Euclidean space |
| \odot | the Hadamard (component-wise) product |
| \oplus | the concatenation |
| \mathcal{L} | the objective function |
| \mathbb{R}^D | the D -dimensional real vector space |
| σ | the softmax function |
| τ | the temperature |
| θ | the weights of a neural network model |
| Φ | the contribution function |

Acronyms

| | |
|--------|---------------------------------------|
| AL | Active Learning |
| KGC | K-greedy Center |
| LVLN | Large Vision-language Model |
| MMCSAL | Multimodal Cold-start Active Learning |

| | |
|-------|--------------------------------------|
| MMSSL | Multi-modal Self-supervised Learning |
| SSL | Self-supervised Learning |
| VCD | Visual Contrastive Decoding |

Chapter 1

Introduction

1.1 Background and Motivation

Humans perceive the world through various sensory modalities, such as sound, visual signals, and sense of touch. This multi-sensory integration allows for a rich and nuanced understanding of the environment. Deep multimodal learning aims to equip neural networks with such a capability to process and integrate information from multiple modalities. This capability has significant implications for various applications, leading to advancements in fields such as affective computing [1, 2], healthcare [3–5], and multimodal chat-bots [6, 7]. Deep multimodal learning offers better performance by exploiting unique and complementary information from different modalities.

However, one of the primary challenges in multimodal learning is the significant data requirement to effectively train models [8–10]. Unlike unimodal learning, which relies on data from a single source, multimodal learning requires large, well-annotated datasets from multiple sources, such as images, text, and audio. This necessity for extensive and diverse data can be particularly burdensome, as collecting and curating high-quality multimodal datasets is both time-consuming and tedious. Moreover, the complexity increases with the need to ensure synchronization and alignment between different modalities, further exacerbating the data demands. This data-hungry nature of multimodal learning poses a significant barrier, limiting its accessibility and scalability, particularly in domains where high-quality data is scarce or expensive to obtain. Addressing these data efficiency issues is crucial

for advancing multimodal learning technologies and making them more practical and widely applicable.

In this dissertation, we cover two mainstream data-efficient learning algorithms within the context of multimodal learning: active learning (AL) [11, 12] and self-supervised learning (SSL) [13]. Additionally, we explore an application of multimodal learning that optimizes data usage to reduce multimodal hallucinations in large vision-language models (LVLMs) [14, 15]. Active learning is an iterative data labeling strategy designed to save labeling budgets for training a model. Self-supervised learning leverages large amounts of unlabeled data by generating supervision signals from the data itself. Despite their advantages in reducing the requirement for labeled data, these approaches are not fully optimized for multimodal learning. Conventional active learning strategies often exhibit data selection bias, tending to select data pairs based primarily on dominant modalities. This bias leads to unbalanced multimodal learning, where the multimodal model is reduced to a unimodal model, ignoring the information from weaker modalities. Furthermore, when integrating active learning with self-supervised learning to address the cold-start problem, where no initial annotations are available, a natural representation gap between different modalities arises. This gap affects the distance calculation of multimodal data pairs due to disparities between unimodal and cross-modal distances. In the context of large vision-language models, the primary issue is multimodal hallucinations, where models generate ungrounded content that does not exist in the input images. Approaches to mitigate hallucinations involve collecting more high-quality data with fewer hallucinations and filtering out training data with false content through human expert review. However, these methods are expensive and require substantial effort of data engineering.

We are motivated by the goal of reducing data cost associated with multimodal learning. We aim to address the imbalance in data selection within conventional warm-start active learning by punishing the gradient embeddings of strong modalities. Additionally, to reduce the modality gap in cold-start multimodal active learning, we introduce a novel multimodal self-supervised learning method and enhance modality alignment during data selection. We also explore strategies to mitigate hallucinations without incurring additional training data costs. Specifically, we propose adjusting the training weights for different tokens according to their visual dependence to prevent the model from learning hallucinated content.

By developing these data-efficient algorithms for multimodal learning, we can build deep multimodal networks with efficient usage of data and shorter training times.

1.2 Objectives

Our objectives are as follows:

- Address the issue of imbalanced data selection in conventional warm-start multimodal active learning. Our goal is to prevent algorithms from favoring dominant modalities by ensuring fairer data selection across all modalities, thereby enabling the training of a more balanced multimodal model that effectively utilizes information from both strong and weak modalities.
- Tackle the cold-start problem in multimodal active learning by minimizing the modality gap that arises from current multimodal self-supervised learning methods. This gap negatively impacts the calculation of distances among multimodal data pairs. Additionally, we aim to enhance modality alignment during data selection to improve data quality for downstream tasks.
- Mitigate hallucinations in large vision-language models without incurring additional training data costs. We aim to understand how different training samples influence hallucinations at the token level and adjust training weights accordingly to prevent the model from learning hallucinated content.

1.3 Contributions

Our main contributions can be stated as follows:

- Analyze the shortcomings of conventional active learning algorithms in selecting balanced multimodal data pairs. To address this, we introduce a gradient embedding modulation method that effectively reduces the bias toward dominant modalities in multimodal active learning. This approach results in more balanced multimodal models that better utilize information from both weak and strong modalities.

- Develop a two-stage method for cold-start multimodal active learning, enabling effective data selection in the absence of initial labels. In the first stage, we reduce the representation gap across modalities using uni-modal prototypes during multimodal self-supervised learning. In the second stage, we enhance modality alignment by introducing a regularization term during data selection. Our method achieves superior model performance with the same amount of data compared to baseline approaches.
- Explore the relationship between hallucinations and visual dependence, discovering that different tokens contribute variably to hallucinations. Based on this insight, we propose a novel approach to mitigate hallucinations in large vision-language models by adjusting the training weights of tokens according to their visual dependence. Our method reduces the hallucination rate by emphasizing tokens that are negatively correlated with the input images, all without requiring additional training data.

1.4 Thesis Outline

The outline of this thesis is as followings. Chapter 2 introduces the background of deep multimodal learning and data efficiency, along with a review of related works. Chapter 3 focuses on the issue of unbalanced multimodal data selection in conventional active learning strategies and proposes a gradient embedding modulation method to mitigate the bias toward selecting the dominant modality over the weaker modality. Chapter 4 addresses the cold-start problem in multimodal active learning, where no initial labels are available for warm-start data selection. We develop a two-stage method that closes the modality representation gap during multimodal self-supervised learning and maximizing the alignment during data selection. Chapter 5 studies the hallucination problem in large vision-language models and introduces a loss re-weighting method and a data filtering strategy to mitigate hallucinations without requiring additional training data or increasing inference computation costs. Finally, Chapter 6 concludes the thesis and discusses about the future work for data-efficient multimodal learning.

Chapter 2

Literature Review

This chapter presents a literature review for this thesis. We begin by introducing deep multimodal learning and discussing the associated data efficiency: labeling data efficiency and efficiency of training data usage. Next, we discuss about the related works and challenges in developing data efficient algorithms for different multimodal learning applications, including balanced warm-start multimodal active learning, cold-start multimodal active learning which integrates multimodal self-supervised learning, and hallucination reduction for large vision-language models.

2.1 Deep Multimodal Learning

Deep multimodal learning mimics the human cognitive process by integrating several modalities to produce desired outcomes through the fusion of different modalities. One of the ultimate goals of multimodal learning is to generate well-aligned multimodal representations that are spatially and temporally coherent, consistent, and naturally clustered [16, 17]. Using image-text as an example, we illustrate the multimodal latent space in Figure 2.1. The image and text pairs are processed by a multimodal model and mapped into a joint latent space, where modality-specific information is preserved and modality-shared information is extracted. Compared to single modality latent spaces, the multimodal latent space is semantically richer, and well-aligned multimodal representations from this latent space are beneficial in various applications such as cross-modal retrieval [18], multi-modal reasoning [19], and image-to-text generation [20]. In this thesis, we study three learning schemes

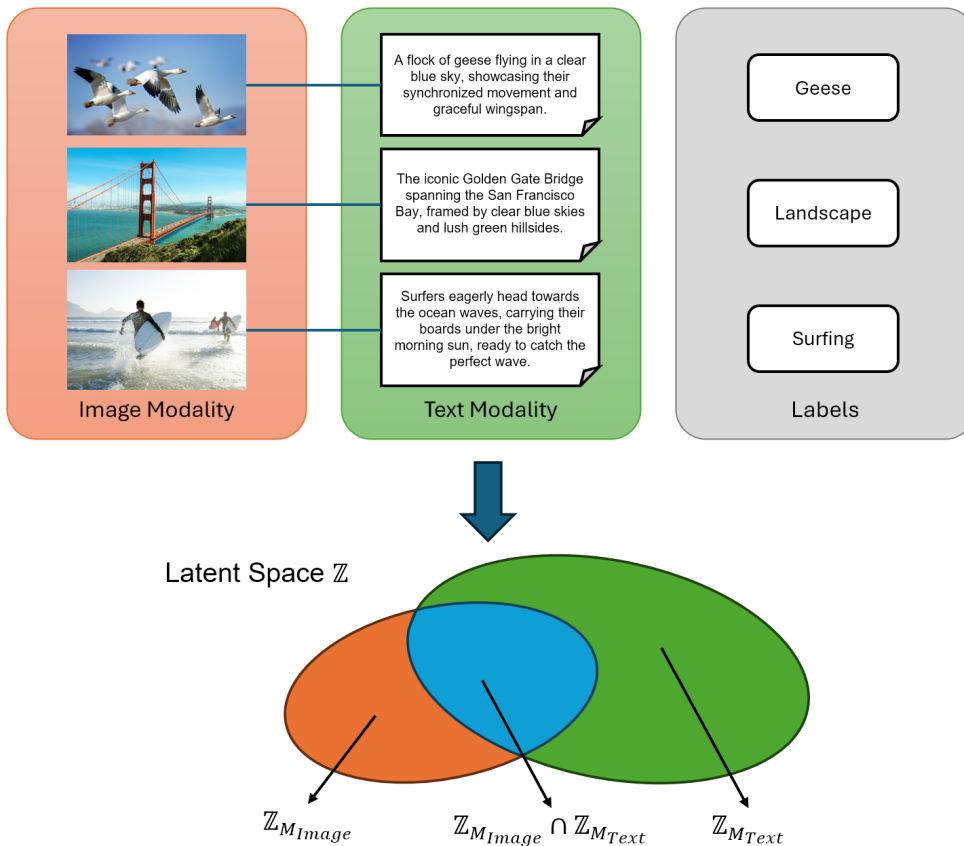


FIGURE 2.1: Multimodal learning latent space representation. The image modality and text modality are processed and mapped into a latent space where modality-specific information is preserved. Additionally, modality-shared information from the two different modalities is extracted to produce richer latent representations compared to uni-modal representations.

of deep multimodal learning based on the presence or absence of labels. We refer to the learning scheme with labels as supervised multimodal learning, the one without labels as self-supervised multimodal learning, and the one with partially labeled data as semi-supervised multimodal learning. In supervised learning, the labels provide supervision signals to guide the model during training. In self-supervised multimodal learning, the multimodal model is primarily supervised by the pairing information across modalities. Semi-supervised multimodal learning is a borderline case where only some of the data are labeled while the rest remain unlabeled. Typically, the model is pre-trained using the more abundant unlabeled data and then fine-tuned with the labeled data to enhance performance.

2.2 Data Efficiency

In this section, we introduce labeling data efficiency which aims to reduce annotation cost when training a network with annotated data; and data usage efficiency whose objective is to optimize the training data usage with re-weighted sampling to train a better model using the same training dataset.

2.2.1 Labeling Data Efficiency

In this thesis, we investigate labeling data efficiency for supervised and semi-supervised multimodal learning where labels are important for the model to gain knowledge from. In supervised multimodal learning, we train the model using data samples from multiple modalities along with their corresponding labels [16, 17]. However, assigning labels is an expensive process that requires hiring human annotators to spend significant time examining each sample individually and marking them with labels. For multiple modalities, this process can be even more time-consuming due to increased data complexity compared to single modality labeling. Therefore, in supervised multimodal learning, we define labeled data efficiency as the number of labeling budgets required to achieve a desired model performance. Achieving the target performance with fewer labeling budgets indicates better data efficiency. Initially, we are given a M modalities dataset and N data samples, $X^U = \{(x_{m_1}, \dots, x_{m_M})_{1..n}\}$, without any labels. To maximize data efficiency, we need to identify the most valuable unlabeled data samples and assign labels to them. This strategy aims to maximize the model’s performance, using accuracy (Acc) as an example:

$$\begin{aligned}
 X^L = \arg \max_{X^L \subseteq X^U} & \quad Acc(f(X^L, Y^L; \theta)), \\
 \text{where } X^L = & \quad \{(x_{m_1}, \dots, x_{m_M})_{1..B}\} \\
 \text{and } Y^L = & \quad \{y_{1..B}\}.
 \end{aligned} \tag{2.1}$$

B is the number of selected data samples, f is the model, θ is the parameters of the model, X^L are the selected data samples from the unlabeled dataset and Y^L are the assigned labels. In semi-supervised multimodal learning [21], we firstly train the model with solely unlabeled data samples to obtain a pre-trained model with

weights θ^* and then fine-tune the model with labeled data samples. Therefore, we define the data efficiency in semi-supervised learning as the same as in supervised learning, which is to locate a subset of unlabeled data samples but with pre-trained model weights θ^* instead of randomly initialized model weights θ .

2.2.2 Data Usage Efficiency

In this thesis, we also focus on improving efficiency of data usage in multimodal learning. Traditional learning methods typically treat all training samples equally when computing the loss and updating model parameters. In these methods, a mini-batch of training data is sampled from the entire dataset, and the average loss is calculated and used to update the network:

$$\mathcal{L}(x_1, \dots, x_B) = \frac{1}{B} \sum_{i=1}^B \mathcal{L}(x_i). \quad (2.2)$$

However, this uniform approach may not be optimal, as individual data samples can vary significantly in terms of difficulty and noise [22]. To address this, we define efficiency of data usage as the quality of the trained model achieved with a given amount of training data. In other words, a method is more data usage efficient if it produces a better model with the same amount of training data. One effective way to enhance efficiency of data usage is by re-weighting the sampled training data based on their difficulty or quality [23]. We denote the re-weighting function as $w(x_i; \theta)$, which measures the importance of each sample x_i and adjusts the final loss accordingly:

$$\mathcal{L}(x_1, \dots, x_B) = \frac{1}{B} \sum_{i=1}^B w(x_i; \theta) \mathcal{L}(x_i). \quad (2.3)$$

As training progresses, the importance of each sample may change depending on the model's parameters. Therefore, we dynamically estimate the weight for each sample using the current model θ to ensure more accurate weight adjustments. It is important to distinguish between labeled data efficiency and data usage efficiency.

While labeled data efficiency focuses on reducing costs by selecting the most relevant samples, data usage efficiency aims to optimize how the data is used, ensuring that the model’s performance is improved without the need for additional data.

2.3 Data-efficient Multimodal Learning

In this section, we begin by discussing the general active learning framework, focusing on related works within two key settings: warm-start and cold-start. Following this, we explore the concept of balanced multimodal learning and examine the challenges associated with warm-start active learning in this context. We then delve into multimodal self-supervised learning, highlighting its applications and the issues it presents in cold-start multimodal active learning. Finally, we focus on the problem of multimodal hallucinations in large vision-language models and review existing methods for hallucination mitigation and the challenges related to their training data usage.

2.3.1 Active Learning

The general active learning process is shown in **Figure 2.2**. Initially, we are given a large unlabeled data pool $X_0^U = \{(x_{m_1}, \dots, x_{m_M})_{1..n}\}$ of n input data with M modalities and an empty labeled data pool $X_0^L = \emptyset$. The labeling budget of each round is set to B . In the first round of active learning, since there is no trained model to evaluate with, a subset X_1^L containing B multimodal data is randomly selected from X_0^U , and they will be assigned with true labels Y_1^L . After data selection, the unlabeled dataset becomes $X_1^U = X_0^U \setminus X_1^L$. The training dataset for the first round of model training consists of X_1^L and Y_1^L . Starting from the second round, an active learning strategy S_{AL} evaluates the trained model and unlabeled data in the last round using an acquisition function, and selects a batch of candidates for label assignment to construct a new training dataset for the current round of model training. The processes of data selection and model training continue until the total labeling budget is run out or the target performance of the trained model is reached.

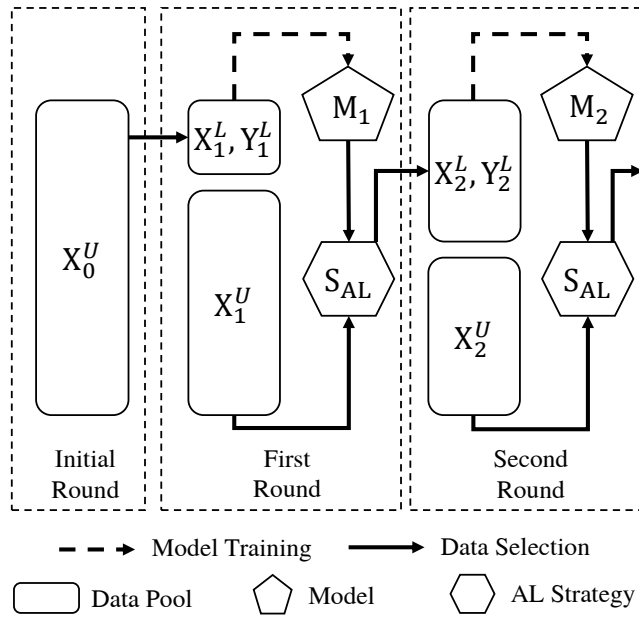


FIGURE 2.2: General active learning process. The dashed lines represent model training. The solid lines represent data selection.

2.3.1.1 Warm-start Active Learning

We study warm-start active learning, where labeled data is available in the first round to train a model. This model can then be used to assess the importance of unlabeled data. Existing warm-start active learning strategies can be grouped into three categories: uncertainty-aware, diversity-aware, and hybrid strategies.

Uncertainty-aware strategies attempt to utilize the data uncertainty or the model uncertainty as a criterion to locate unlabeled data points that the current model has less confidence about. One strategy is to utilize the posterior classification probability distribution by measuring its entropy [24, 25], or the margin between the most confident class and the second most confident class [26]. In addition, uncertainty can be evaluated as the variance of predictions generated by an ensemble of models [27] or by multiple inferences with Monte-Carlo dropout as an alternative Bayesian approximation for static networks [28]. Moreover, ALFA-Mix [29] evaluates unlabeled samples by mixing their features with labeled samples and observing whether there is inconsistency among predictions from mixed features. DFAL [30] incorporates adversarial attack techniques [31] to select unlabeled data samples located close to the classification boundaries.

Diversity-aware strategies tend to select unlabeled data points whose features are as diverse as possible to minimize data redundancy. [32] utilizes K-medoid algorithm [33] to select representative data centroids that minimize the total distance from other data samples to the nearest centroids. CoreSet [34] greedily selects unlabeled data samples that have maximum distances from their nearest neighbors. [35] adopts the determinantal point process (DPP) to evaluate the diversity by calculating the determinant of the similarity matrix. Diversity-aware strategies can also be considered in the context of distribution matching, which aims to reduce the gap between the distributions of labeled and unlabeled samples in latent space or feature space. VAAL [36] trains a variational auto-encoder to construct the latent distribution of labeled samples and an adversarial network to distinguish labeled samples and unlabeled samples in the latent space. Moreover, the maximum mean discrepancy (MMD) [37], the \mathcal{H} -divergence [38] and the Wasserstein distance [39] are used to measure the distribution gap.

To achieve a better trade-off between informativeness and diversity, hybrid methods are developed with an awareness of both. Since diversity-aware strategies are orthogonal to most of uncertainty-aware strategies [40], they could be easily combined together. ALFA-Mix [29] adopts K-means clustering to further filter out samples to enhance diversity. BADGE [41] represents unlabeled data samples via gradient embedding of parameters of the last classifier layer and applies K-means++ [42] to form a diverse data selection which still carries high uncertainty.

2.3.1.2 Cold-start Active Learning

Unlike warm-start active learning, cold-start active learning aims to select the initial batch of data samples in circumstances where no labels are available. An effective cold-start AL strategy should identify a subset of samples that are more informative and diverse compared to a randomly chosen set. Due to the absence of labels, self-supervised learning has been widely adopted to mitigate the cold-start problem in active learning. ALPS [43] utilizes self-supervised masked language modeling based on BERT [44] to create surprisal embeddings for each sentence sample, identifying more significant samples. The study in [45] examines both self-supervised and clustering methods to form the initial data selection for active learning on image classification task. They find that complicated initial data selection strategies based on SimCLR [46], SCAN [47] and VAE [48] are not strongly

competitive with random initialization, showing that the cold-start AL problem is challenging. PT4AL [49] performs a pretext task before data selection and identifies significant positive correlations between rotation prediction and image classification losses, and between colorization and segmentation losses. It ranks samples based on their pretext task loss, selecting the most challenging ones for the initial cold-start batch. Moreover, they find that the contrastive loss in SimSiam [50] shows weak correlation with the downstream task loss, making contrastive learning as an unsuitable pretext task in their framework. They suspect that contrastive loss is class-biased due to its behavior of pulling positive samples together and pushing away negative samples. TypiClust [51] has proven that under a very low label budget, the representativeness of samples is more critical than their uncertainty. Therefore they employ a self-supervised deep clustering algorithm, SCAN [47] to pick typical samples. Additionally, they perform representation learning methods DINO [52] and SimCLR [46] followed by K-means to group data samples into clusters, choosing the most dense samples in each cluster. CSVAL [53] uses MoCo-v2 [54] as a self-supervised learning strategy to obtain image representations, then generates pseudo labels through K-means to train a classification model. It adopts Dataset Map [55] to select out easy-to-learn samples to form the first batch. These samples are also found to be hard-to-contrast samples as they are hard to be distinguished from others, indicating their typicality. ActiveFT [56] proposes to utilize data representations provided by pretrained models such as DINO [52] to select a subset of data for label annotations and subsequent model fine-tuning. It minimizes the data distribution gap between the selected subset and the entire unlabeled dataset and regularizes the diversity to keep data selection from collapsing to centroid points.

2.3.2 Balanced Multimodal Learning

While existing warm-start active learning methods have improved the efficiency of training deep networks, they are primarily applied in unimodal settings. When these methods are used to select multimodal data, the resulting multimodal models often underperform on weaker modalities compared to models trained with randomly selected data. This poor performance may be due to unbalanced multimodal learning, where the selected data is biased toward dominant modalities.

It has been observed that the best unimodal networks could potentially outperform multimodal networks regardless of fusion mechanisms or regularization methods [57]. Recent works show that the degradation of multimodal learning could be due to unbalanced optimization among different modalities. In [58], the failure of multimodal learning is attributed to modality competition where only dominant modalities are fully explored by joint training. Similarly, [59] demonstrates that multimodal learning greedily optimizes the dominant modalities and chooses to balance their training speeds. [57] proposes to blend gradients with weights that are disproportional to the overfitting and generalization ratio of each modality so that each modality could be optimized in a balanced manner. [60] finds that fusion mechanisms such as concatenation and summation encourage the dominant modality to learn faster and thus develops gradient modulation to adaptively balance the training speed of each modality.

In our thesis, we analyze the selection process of existing active learning methods for multimodal data samples, identifying the reasons behind their bias toward certain modalities that hinder balanced multimodal learning. To address this issue, we introduce a strategy that penalizes the dominant modalities during data selection, leading to the development of more balanced and effective multimodal models.

2.3.3 Multimodal Self-supervised Learning

As previously discussed, in cold-start active learning settings, where no labels are available at the beginning, self-supervised learning is employed to generate useful data features, enabling the selection of important data samples without requiring labels. In this context, we examine existing multimodal self-supervised learning methods.

Multimodal self-supervised learning (MMSSL) constructs semantically rich representations across multiple modalities without relying on human annotations. These methods can be broadly categorized based on their learning objectives into three main approaches: contrastive discrimination, clustering, and masked token prediction [61]. **Contrastive** methods treat paired multimodal samples as positive pairs and unpaired ones as negative pairs, aiming to minimize the representation distance between positive pairs while maximizing it for negative pairs during training. Models trained with contrastive learning, such as CLIP [9], an image-text

model, have produced image and text representations with aligned cross-modal semantic information, demonstrating robust zero-shot capabilities. Furthermore, contrastive learning has been successfully applied to various modality combinations, including video-text [62, 63], audio-text [10], and video-audio [64, 65]. **Clustering** aims to group data with similar semantic information. In the realm of visual self-supervised learning, SwAV [66] performs online clustering, assigning multiple instances to prototypes and learning representations by predicting the cluster assignment of one view from another view. PCL [67] combines contrastive learning for extracting useful visual features with cluster assignment prediction to enhance the semantic structure of the representations. In the multimodal domain, XDC [68] uses the clustering assignments of one modality as prediction targets for another, thereby deriving multimodal representations. AV-HuBert [69] performs clustering on unmasked visual and auditory sequences, forcing the masked visual and auditory tokens to predict the clustering results. **Masked token prediction** methods are often used in the image-text domain. These methods such as VL-BEiT [70] and ALBEF [71] involve recovering masked image or text tokens given the context of unmasked image and text, facilitating interaction between image and text to yield meaningful representations.

In this thesis, we focus on the contrastive learning-based method CLIP [9], which is notable for its scalability to large datasets and its ability to produce semantically rich data features. However, researchers have identified a modality gap [72] in its feature space between image and text modalities. This gap implies that features within the same modality are closer together, while features from different modalities are further apart. We observed a similar modality gap when applying the same training method to video-audio datasets. This gap complicates the process of selecting data pairs, as it affects the calculation of both unimodal and multimodal distances. To address this issue, we propose a new multimodal self-supervised learning method aimed at reducing the modality gap, thereby improving data selection results.

2.3.4 Hallucination Mitigation in LVLMs

In the context of large vision-language models (LVLMs), we focus on image hallucination, which occurs when the model generates responses that are inconsistent

with the visual input. The causes of hallucination can be understood from several perspectives. **Training Data:** Misaligned visual instructional training data [73] and data containing hallucinations [74] can introduce hallucination into the model during training. **Decreased Visual Information:** As the length of the generated response increases, the model is more likely to produce hallucinated content due to the gradual loss of visual information [75–77]. **Language Prior:** Hallucination can occur when the model’s language processing ability exceeds its vision processing ability, leading the model to overlook the image content [77–79]. **Visual Encoder:** The current design of LVLMs often involves extracting hidden states from a pre-trained vision model and feeding them into a language model. Therefore, when the visual representations are not good enough, the model is not able to generate consistent responses [80, 81]. The current hallucination reduction methods can be broadly categorized into three classes based on the stage they target: data preparation, training and inference. Data curation methods [74] focus on collecting high-quality data with less hallucination, ensuring that the trained model produces more robust responses. Training methods [76, 82] concentrate on manipulating the training process by either adjusting the generation length or training a plug-in module to control the level of hallucination. Inference-based methods [78, 83] aim to prevent the generation of hallucinated content through techniques such as contrastive decoding or rolling back when summary tokens are detected.

In this thesis, we aim to mitigate hallucination from data perspective. We argue that each data token contributes differently to image hallucinations, therefore, uniformly treating every token from training dataset might not be the best choice. We adopt the idea of re-weighting training data to prevent the model learning from potentially harmful or incorrect information in training dataset. In this way, we can achieve hallucination reduction during training using the same amount of training data and do not need for collecting additional data.

Chapter 3

Towards Balanced Active Learning for Multimodal Classification

3.1 Introduction

As discussed in Chapter 2, deep learning strategies have been implemented to train deep multimodal neural networks [16, 84]. In this chapter, we focus on multimodal classification which is one of the classical multimodal learning tasks that aims to exploit complementary information inherent in multimodal data to achieve better classification performance. However, such deep multimodal networks require an enormous amount of data to learn from, given their huge number of parameters. To reduce data cost, active learning (AL) is used to select a subset of more informative and distinctive unlabeled data samples for label assignment by oracles. Consequently, large networks can maintain performance while utilizing a smaller labeling budget. Most existing active learning algorithms are designed for unimodal tasks such as image classification [27, 34], object detection [85, 86] and language modeling [87, 88]. The objective is to select samples that have high uncertainty in them, carry novel knowledge for model training and those with distinctive features. However, there has been significantly less research reported on the design of effective active learning strategies for multimodal learning [89].

In this chapter¹, we initially examine the performance of existing active learning strategies in selecting multimodal data. Our experiments reveal that these strategies tend to focus more on the dominant modality rather than fairly considering all modalities. For instance, in an image-text classification task, if the text contributes more to model optimization, active learning strategies may exhibit a bias towards the more distinguishable text modality by selecting valuable text samples and disregarding the informativeness of image samples. As a result, the selected multimodal dataset could become unbalanced, with insufficient information from the image modality, potentially leading to a degraded image model backbone. Recent works [57–60] point out that balancing the training and optimization of all modalities is a key factor for successful multimodal learning. Similarly, it is crucial to design active learning strategies that can select multimodal data with fairness among all modalities to assist balanced multimodal learning.

Based on our findings, we develop an algorithm for **Balanced Multimodal Active Learning (BMMAL)** that selects multimodal data by fairly considering each modality present in the data. In our approach, we choose the gradient embedding of model parameters, as it reflects the impact on model training and captures the diversity of data samples. However, we examine how the previous gradient embedding method [41] fails to select balanced multimodal data. To ensure fairness, we individually assess the contribution of each modality feature by examining the Shapley value, which attributes its contribution to the final multimodal prediction. We then apply modulation on the gradient embedding to penalize samples with dominant modalities. Lastly, a clustering seed initialization algorithm is employed to select diverse multimodal data with a significant influence on model training.

In summary, our empirical studies show that most existing active learning strategies fail to select a balanced multimodal dataset. We analyze how to improve the current gradient embedding based active learning strategy to rectify this. To mitigate unfair data selection, we propose a method to modulate the gradient embedding on sample-level to select more balanced multimodal candidates. We conduct experiments on three multimodal datasets to show that our proposed method treats multimodal data more equally and achieves better performance.

¹The work in this chapter has been published as **Meng Shen**, Yizheng Huang, Jianxiong Yin, Heqing Zou, Deepu Rajan, and Simon See, “Towards Balanced Active Learning for Multimodal Classification”. *In Proceedings of the 31st ACM International Conference on Multimedia (MM ’23)*.

3.2 Methodology

In this section, we firstly introduce the framework of multimodal learning, and then provide analysis of imbalanced data selection in current state-of-the-art active learning strategy, BADGE [41]. Finally, we propose our method to mitigate the unfair data selection by punishing gradient embeddings of dominant modality.

3.2.1 Multimodal Learning Framework

We introduce our multimodal learning framework for classification task. x_{m_1} and x_{m_2} represent the input data from two different modalities. They are processed through encoders φ_{m_1} and φ_{m_2} respectively to extract unimodal features $z_{m_1} \in \mathbb{R}^{D_{m_1}}$ and $z_{m_2} \in \mathbb{R}^{D_{m_2}}$. We adopt concatenation, a widely used early-fusion mechanism [16], to construct multimodal features $z_{mm} = z_{m_1} \oplus z_{m_2}$.² The unimodal and multimodal features are fed to unimodal classifiers C_{m_1} , C_{m_2} and multimodal classifier C_{mm} respectively to produce logits f_{m_1} , f_{m_2} and f_{mm} for classification. The final loss is the average cross-entropy loss \mathcal{L}_{CE} of unimodal and multimodal logits with true labels y :

$$\mathcal{L}_{final} = \frac{1}{3}[\mathcal{L}_{CE}(f_{m_1}, y) + \mathcal{L}_{CE}(f_{m_2}, y) + \mathcal{L}_{CE}(f_{mm}, y)]. \quad (3.1)$$

Once the model is trained, the unlabeled data samples are evaluated using an acquisition function and filtered for labeling.

3.2.2 Analysis of Imbalance in AL

We introduce one of the popular active learning algorithms BADGE [41] and provide analysis of its imbalanced data selection over multimodal data samples. BADGE was the first to propose the replacement of features for embedding with the gradient of the weight of the last FC layer, which acts as the classifier. In our case, the last FC layer for multimodal classification is the multimodal classifier C_{mm} . The weight W of the classifier is a 2-dimensional matrix of size $K \times D_{mm}$,

²Other fusion mechanisms such as summation and NL-gate are implemented in our further experiments.

where K is the number of classes and D_{mm} is the dimension of concatenated multimodal feature $D_{m_1} + D_{m_2}$. The corresponding multimodal cross-entropy loss can be expanded as

$$\begin{aligned}\mathcal{L}_{mm} &= - \sum_{i=1}^K y_i \cdot \log \sigma(f_{mm})_i \\ &= - \sum_{i=1}^K y_i \cdot \log \frac{e^{z_{mm} \cdot W_i^T}}{\sum_{i=1}^K e^{z_{mm} \cdot W_i^T}},\end{aligned}\tag{3.2}$$

where σ is softmax function and $z_{mm} \cdot W_i^T$ is the i^{th} element of logits f_{mm} . The gradient embedding is defined as $g = \frac{\partial \mathcal{L}_{mm}}{\partial W}$, and it is a 2-D matrix of size $K \times D_{mm}$ where the i^{th} row is

$$g_i = (f_i - 1_{\hat{y}_{mm}=i})z_{mm},\tag{3.3}$$

where $\hat{y}_{mm} = \underset{i \in [K]}{\operatorname{argmax}}[(f_{mm})_i]$ is the pseudo label for unlabeled data samples. The gradient embedding is flattened into a vector for sampling. It not only carries the uncertainty of classification from the margin between logits f_i and pseudo labels \hat{y}_{mm} , but also is representative enough due to the information present in z_{mm} .

However, in multimodal learning settings, identifying the source of uncertainty can be challenging. Upon examining the calculation of multimodal logits $f_i = z_{mm} \cdot W_i^T = z_{m_1} \cdot (W_i)_{m_1}^T + z_{m_2} \cdot (W_i)_{m_2}^T$, where W_i is divided into two matrices $(W_i)_{m_1}$ and $(W_i)_{m_2}$, it is difficult to determine which modality carries more uncertainty and which carries less. To illustrate, for a visual event such as drawing, the visual modality contains more information and contributes more to multimodal logits by generating a larger output. The multimodal uncertainty calculation is thus skewing the visual uncertainty instead of considering both visual and auditory uncertainties fairly. From **Section 3.3.4**, we find that BADGE does pay more attention to the dominant modality, which might potentially damage the performance of joint multimodal learning. Another limitation of BADGE is its inability to distinguish modality contributions. For instance, given two data samples with identical logits, we should prioritize the one with a more balanced contribution during data selection to facilitate balanced multimodal learning. However, the current BADGE algorithm cannot achieve this. Similarly, as discussed in Section 2.3.1.1, most conventional active learning algorithms are designed for unimodal tasks and do not account for the need to balance data selection across multiple modalities.

Hence, we develop a balanced multimodal active learning method that could avoid

biased data selection towards the dominant modality to mitigate modality competition and assure that the trained multimodal network would not easily degenerate to the dominant modality. While our designed method is encouraged to pay more attention to the weaker modality, it is essential to ensure that it does not overly lean towards the weaker modality, as this may also harm the multimodal classification performance.

3.2.3 Guidelines to Design Balanced MMAL

To make existing AL strategies more suitable for balanced multimodal learning, it is necessary to inspect the individual modality contribution and reduce the contribution gap among different modalities. We empirically propose three guidelines for designing active learning strategies that treat each modality more equally. Let $\Phi_{m_i}(x)$ represent the contribution of the i^{th} modality of data sample x to the final model outcome, which should satisfy:

$$\sum_{i=1}^M \Phi_{m_i}(x) = 1. \quad (3.4)$$

We introduce the dominance degree $\rho(x)$ to quantify how severely a data sample x is dominated by the strongest modality:

$$\rho(x) = \sum_{i=1}^M [\max(\Phi_{m_1}(x), \dots, \Phi_{m_M}(x)) - \Phi_{m_i}(x)]. \quad (3.5)$$

We further partition the entire unlabeled dataset into multiple subsets for the ease of discussion: $X = \{X_1, \dots, X_M\}$. In each subset X_i , modality m_i contributes the most:

$$\Phi_{m_i}(x) \geq \Phi_{m_j}(x), i \neq j, \forall x \in X_i. \quad (3.6)$$

Guideline 1: Acquisition function a in an active learning strategy is used to score a data point to represent its importance [90]. For two multimodal data samples x_i and x_j , if their acquisition scores of conventional active learning (CAL) strategies

are equal, the one with more balanced unimodal contributions should have higher acquisition scores of balanced multimodal active learning strategies,

$$a_{BMMAL}(x_i, \rho_i) > a_{BMMAL}(x_j, \rho_j), \rho_i < \rho_j, \quad (3.7)$$

where $a_{CAL}(x_i) \approx a_{CAL}(x_j), i \neq j$.

Here, $a_{CAL}(x_i) \approx a_{CAL}(x_j)$ reflects the scenario where two samples have nearly identical acquisition scores under conventional active learning. While exact equality is unlikely due to the continuous nature of a_{CAL} , this illustrative condition highlights the emphasis of BMMAL on selecting samples with more balanced unimodal contributions.

By following Guideline 1, data samples with more equal unimodal contributions are more likely to be selected. However, this does not guarantee that the stronger modality will be suppressed, nor does it ensure that the weaker modality will not be overly encouraged. Therefore, we introduce two additional guidelines.

Guideline 2: To avoid biased data selection favoring the stronger modality, the gap in the relative proportions of acquisition scores between data samples dominated by the weaker modality and the stronger modality should be reduced. In the two-modality case where m_1 is the weaker modality and m_2 is the stronger modality (i.e. the average contribution of m_1 over the entire dataset is less than that of m_2 , $\frac{1}{|X|} \sum_{x \in X} \Phi_{m_1}(x) < \frac{1}{|X|} \sum_{x \in X} \Phi_{m_2}(x)$), this can be expressed as:

$$\frac{\frac{1}{|X_1|} \sum_{x \subseteq X_1} a_{CAL}(x)}{\frac{1}{|X_2|} \sum_{x \subseteq X_2} a_{CAL}(x)} < \frac{\frac{1}{|X_1|} \sum_{x \subseteq X_1} a_{BMMAL}(x)}{\frac{1}{|X_2|} \sum_{x \subseteq X_2} a_{BMMAL}(x)}. \quad (3.8)$$

This inequality indicates that balanced multimodal active learning (BMMAL) adjusts acquisition scores such that the relative contribution of the weaker modality is enhanced, promoting a more balanced selection of data points across modalities.

Guideline 3: Lastly, to prevent biased data selection towards the weaker modality, it is necessary to ensure that the contribution of each modality to the acquisition score function a_{BMMAL} is still proportional to its modality contribution to the model outcome on the sample-level. It ensures that the data samples are selected in a way that fairly represents the contributions of each modality to the actual model outcome.

In summary, Guideline 1 prioritizes the samples with more equal unimodal contributions. Guideline 2 and 3 work together to punish the stronger modality on the dataset-level but maintain the relationship between strong and weak modality on the sample-level, avoiding biases towards either the stronger or weaker modalities.

3.2.4 Estimate Modality Contribution

We show how we compute modality contribution Φ . In the context of multimodal classification, balanced active learning should select data samples that fairly contribute to the performance of all modalities. To achieve this, it is essential to estimate the degree to which each modality of a given data sample contributes to the final multimodal prediction. One approach involves assessing modality importance by computing the disparity in model performance before and after the incorporation of a particular modality. Researchers have proposed various techniques to remove the information of one modality, such as masking [91], permutation [92], and empirical multimodally-additive projection (EMAP) [93]. Nonetheless, these attribution methods are ill-suited for active learning as they require ground truth labels to calculate model performance metrics, such as accuracy. As a result, these methods cannot be employed for estimating modality contribution for unlabeled data due to the absence of ground truth labels.

Therefore, we choose to use the Shapley value to estimate modality contribution without the need for true labels. The Shapley value [94] was proposed to fairly attribute payouts among group of cooperative players based on their contributions to the total payout in game theory. In deep learning, SHapley Additive exPlanations (SHAP) value [95] considers each feature as a player and the model prediction as the total payout to estimate feature contributions. Let $\mathcal{M} = \{z_{m_1}, \dots, z_{m_M}\}$ represent the set of all modality features, \mathcal{S} denote the subset, and V symbolize the model outcome. Here, we use features instead of raw data inputs since features are utilized in active learning. To estimate the Shapley value of i^{th} modality feature z_{m_i} , we compute the marginal contribution to the subset \mathcal{S} and average over all possible subset selections:

$$\phi(z_{m_i}) = \sum_{\mathcal{S} \subseteq \mathcal{M} \setminus \{z_{m_i}\}} \frac{|\mathcal{S}|!(|\mathcal{M}| - |\mathcal{S}| - 1)!}{|\mathcal{M}|!} [V(\mathcal{S} \cup \{z_{m_i}\}) - V(\mathcal{S})]. \quad (3.9)$$

We use the largest predicted class probability $p_{\hat{y}}$ provided by f_{mm} as the model outcome V , where \hat{y} is the pseudo class. For the most common two-modality case, the Shapley values of modality features can be computed as follows (\emptyset represents a zero vector):

$$\begin{aligned}\phi(z_{m_1}) &= \frac{1}{2}[V(z_{m_1}, z_{m_2}) - V(\emptyset, z_{m_2}) + V(z_{m_1}, \emptyset) - V(\emptyset, \emptyset)], \\ \phi(z_{m_2}) &= \frac{1}{2}[V(z_{m_1}, z_{m_2}) - V(z_{m_1}, \emptyset) + V(\emptyset, z_{m_2}) - V(\emptyset, \emptyset)].\end{aligned}\tag{3.10}$$

The complexity of **Equation 3.9** is $O(2^n)$, where n is the number of modalities. The computation could be expensive if the number of modalities is large. In our work, we focus on the common two-modality case, where the computational cost is manageable. However, for scenarios involving more than two modalities, the complexity can indeed become a concern. To address this, a potential solution is to use Monte Carlo sampling, which can approximate the computation efficiently and reduce the complexity from $O(2^n)$ to $O(n)$.

The Shapley value could be positive, negative or zero. While the sign indicates in which direction each modality contributes, our primary interest lies in the extent of its contribution. Hence, we define modality contribution as follows:

$$\Phi_{m_i} = \frac{|\phi(z_{m_i})|}{\sum_{i=1}^M |\phi(z_{m_i})|}.\tag{3.11}$$

3.2.5 Proposed Method

Following the proposed guidelines, we redesign the BADGE for multimodal classification scenarios with k modalities $\{m_1, \dots, m_k\}$, to achieve more balanced data selection. The i^{th} row of gradient embedding in **Equation 3.3** could be derived as concatenation of k unimodal gradient embeddings:

$$g_i = (f_i - 1_{\hat{y}_{mm}=i})z_{m_1} \oplus \dots \oplus (f_i - 1_{\hat{y}_{mm}=i})z_{m_k}.\tag{3.12}$$

We then design weights w_{m_k} for k^{th} modality, and scale each unimodal gradient embedding by them respectively:

$$w_{m_k} = 1 - \rho_{m_k}(x), \quad (3.13)$$

$$g'_i = w_{m_1}(f_i - 1_{\hat{y}_{mm=i}})z_{m_1} \oplus \cdots \oplus w_{m_k}(f_i - 1_{\hat{y}_{mm=i}})z_{m_k}. \quad (3.14)$$

Here, $\rho_{m_k}(x) = [\max(\Phi_{m_1}, \dots, \Phi_{m_k}) - \Phi_{m_k}]$ is the difference between contributions of the dominant modality and k^{th} modality. Note that the gradient embedding of larger l2 norm will be selected more easily by K-Means++ algorithm [41]. In K-Means++, the probability of selecting a data point is directly proportional to its distance from the nearest, previously chosen points. This mechanism promotes diversity among the selected samples and, in our case, helps prioritize samples with larger gradient magnitudes, which are likely to be more informative for the task. Therefore, by multiplying with these weights, the magnitude of gradient embedding will be suppressed more if their unimodal contributions are more unbalanced. It aligns with our Guideline 1 where we want to punish the samples with unbalanced contributions.

Moreover, we observe that the average ρ of the subset in which the weaker modality dominates is smaller than that of the subset where the stronger modality dominates. See **Figure 3.5** and our discussion in **Section 3.3.6**. If m_1 is the weaker modality regarding the entire dataset, then we will have $\frac{1}{|X_1|} \sum_{x \in X_1} \rho(x) < \frac{1}{|X_2|} \sum_{x \in X_2} \rho(x)$ for two subsets X_1 and X_2 dominated by m_1 and m_2 respectively. It means that the subset where the stronger modality dominates will be suppressed more, and it follows our Guideline 2 to punish the stronger modality on the dataset-level.

Finally, the Guideline 3 is also adhered to. For each sample, the modality with a higher contribution to the model outcome is always assigned a greater weight, resulting in a higher magnitude of unimodal gradient embedding. This ensures that the contribution to data selection is proportional to the contribution to the model outcome and model optimization if selected.

In the end, we perform K-Means++ over the scaled gradient embedding to select candidates for labeling. As a result, our BMMAL strategy could achieve more balanced active learning on multimodal classification than BADGE. It could prevent biased selection towards either the stronger or weaker modalities, thus benefiting multimodal learning.

3.3 Experiment

In this section, we introduce the datasets, baselines and experiment settings of evaluation. We compare AL performance of our method against baselines and provide ablation studies.

3.3.1 Dataset

Food101 [25] is a multi-class food recipe dataset with 101 kinds of food. Each recipe consists of a food image and textual recipe description. The dataset consists of 45,719 samples for training and 15,294 samples for testing.

KineticsSound [96] is a sub-dataset containing 31 action classes selected from Kinetics-400 [97]. These action classes are considered to be correlated to both visual and auditory content. This dataset contains 14,739 clips for training and 2,594 clips for testing.

VGGSound [98] is a large-scale video dataset with 309 classes. Each video clip is 10-second and captures the object making the sound. We are only able to download 180,911 clips for training and 14,843 clips for testing due to the unavailability of YouTube videos.

3.3.2 Baseline

We consider seven existing active learning strategies as baselines. **Random** selects the data samples randomly from the unlabeled data pool. **Entropy** [24] selects data samples with the highest entropy of multimodal classification probabilities. **CoreSet** [34] filters out a subset of unlabeled data with representative multimodal features via K-center greedy algorithm. **BADGE** [41] is a hybrid method that selects diverse data samples by K-means++ sampler over their gradient embedding of multimodal classifier. **BALD** [99] is a Bayesian method to evaluate the mutual information between model predictions and model parameters. Since our model is static, we run five rounds of model forwarding with dropout enabled to obtain the entropy of model parameters. **DeepFool** [30] adopts an adversarial-like approach that adds small perturbations over multimodal features and selects data whose

predictions are flipped. **GCNAL** [100] learns an extra graph convolution network to distinguish labelled and unlabelled samples and selects unlabelled samples that are sufficiently different from labelled ones.

3.3.3 Experiment Setting

Image-text Classification. For the Food101 dataset, we adopt ResNet-101 pre-trained on ImageNet as the image backbone and pre-trained Bert-base model [44] as the text backbone. All unimodal and multimodal classifiers are single FC layers. We use AdamW [101] as the optimizer and train the model for 15 epochs in each AL round and adopt random crop, random horizontal flip and random grey scale for image augmentation.

Video Classification. For VGGSound and KineticsSound, we utilize ResNet-2P1D-18 [102] as visual backbone. The difference in the way it is used for the two datasets is that the model is pre-trained on Kinetics-400 for VGGSound and is randomly initialized for KineticsSound since it is a subset of Kinetics-400. We use the randomly initialized ResNet-18 as an auditory backbone whose input channel is modified from 3 to 1. The video is uniformly sampled into 10 frames at the rate of one frame per second. The audio clip is transformed into a spectrogram with a window length of 512 and an overlap length of 353. For video augmentation, we randomly sample 5 frames out of 10 frames and apply image augmentation techniques on each frame. For audio augmentation, we randomly extract a 5-second audio fragment from the whole audio clip. We use Adam as optimizer and train the model for 45 epochs in each round.

The experiment is repeated 5 times for image-text classification and 3 times for video classification to remove the randomness of the initial querying. For multimodal fusion, we apply concatenation which is a widely used multimodal fusion mechanism on all tasks. In addition, we implement summation and NL-gate [103] that is similar to multi-head attention [104] in further experiments.

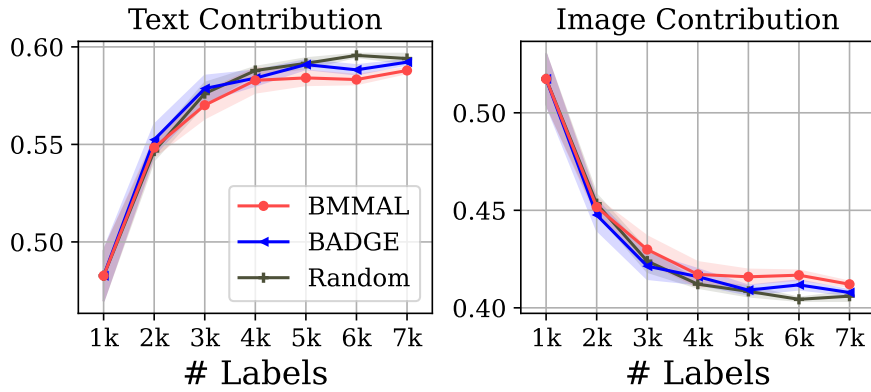


FIGURE 3.1: Modality contribution Φ across different AL iterations on the Food101 test set.

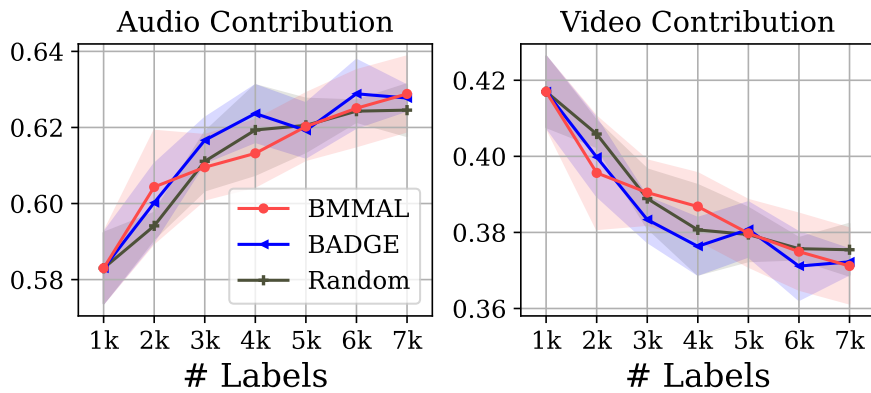
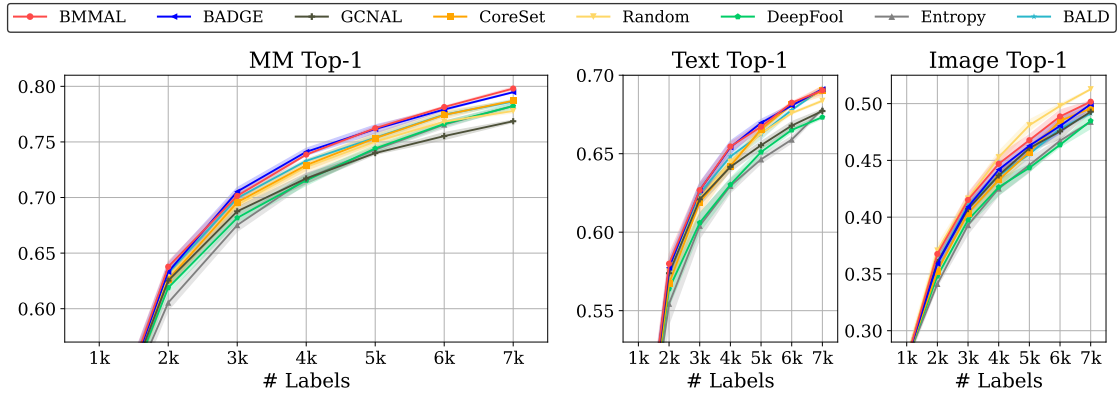


FIGURE 3.2: Modality contribution Φ across different AL iterations on the KineticsSound test set.

3.3.4 AL Performance

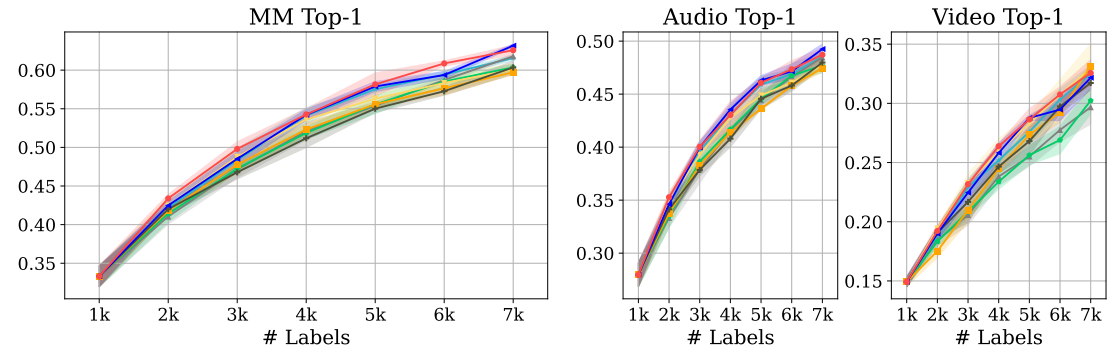
A fair and good AL strategy ought to select important multimodal data that could contribute to multimodal tasks and, simultaneously, pay fair attention to weaker modalities and strong modalities to prevent the trained multimodal network from degenerating into only a good unimodal network. We run conventional active learning strategies along with our proposed method BMMAL on several multimodal datasets, and compare their multimodal and unimodal classification accuracy.

We firstly draw the trend of modality contributions to the predicted probability over the ground truth class on test dataset across different active learning iterations in **Figure 3.1** and **Figure 3.2**. As shown in the figures, the textual modal contributes more than the imagery modal on the Food101 after second iteration, and the auditory modal contributes more than the visual modal on the KineticsSound.



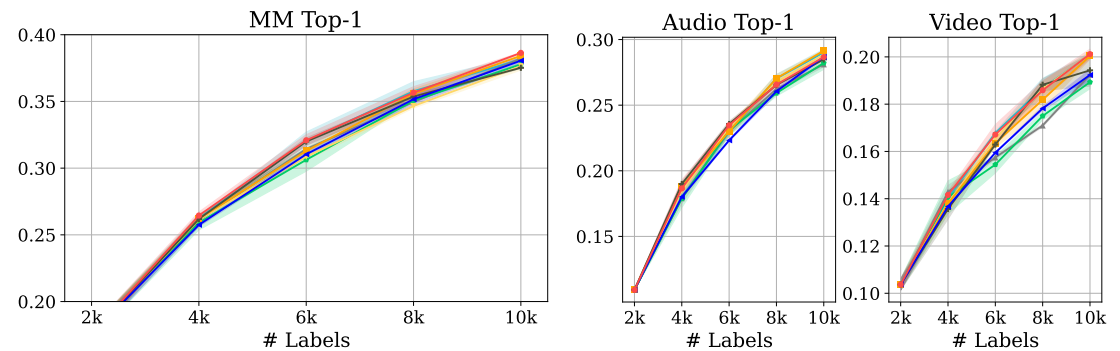
(A) Multimodal performance comparison across AL iterations on Food101.

(B) Unimodal performance comparison across AL iterations on Food101.



(C) Multimodal performance comparison across AL iterations on KineticsSound.

(D) Unimodal performance comparison across AL iterations on KineticsSound.



(E) Multimodal performance comparison across AL iterations on VGGSound.

(F) Unimodal performance comparison across AL iterations on VGGSound.

FIGURE 3.3: Performance comparison between proposed method and other conventional AL strategies with concatenation fusion method. The metric selected is top-1 accuracy (Top-1) on multimodal and unimodal classification.

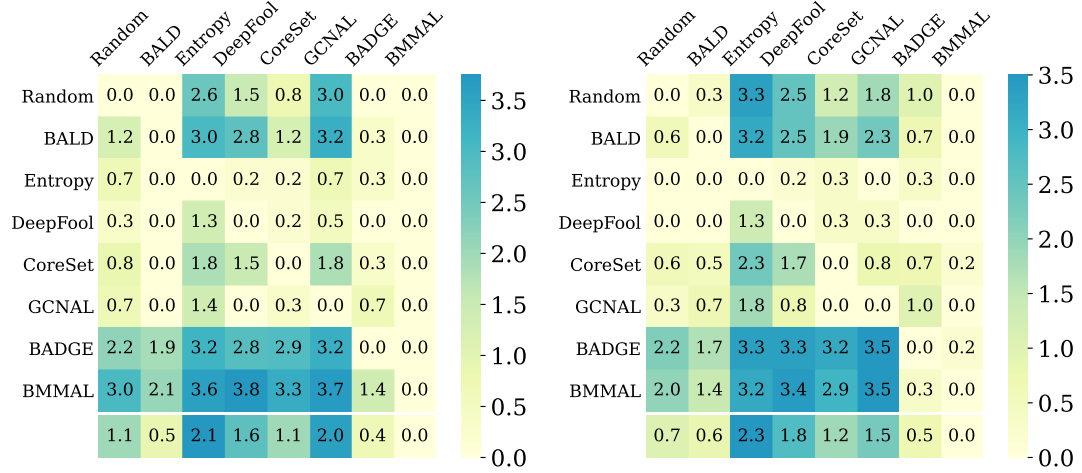
More importantly, the difference between two unimodal contributions of BMMAL is overall smaller than both BADGE and Random. It means that two modalities contribute more equally in the models trained by the data selected by BMMAL.

The performance comparison of each AL iteration on the Food101 dataset is shown in **Figure 3.3a** and **Figure 3.3b**. Note that textual modality is the stronger modality since iteration 2. Our method outperforms all baselines except BADGE in multimodal classification. In text classification, BMMAL, BADGE and CoreSet achieve good performance. In image classification, our method is superior to most of the baselines except Random. From the above comparison, we can tell that BADGE and CoreSet mainly focus on selecting valuable samples over the stronger text modality and ignore the weaker image modality. Although Random uniformly selects multimodal data without any weighting in image classification, it is considered unfair concerning the text modality.

The performance comparison of each AL iteration on the KineticsSound dataset is shown in **Figure 3.3c** and **Figure 3.3d**. Note that auditory modality is the stronger modality. Our method outperforms all baselines in multimodal classification. BADGE performs the best on audio classification on many iterations, However, its performance declines on video classification indicating that biased data selection might negatively affect multimodal classification. It shows that BADGE tends to assign more importance to audio modality during data selection and such behavior might negatively affect multimodal joint training.

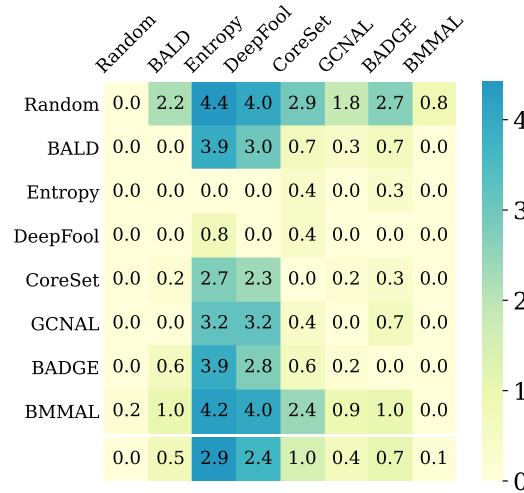
The performance comparison of each AL iteration on the VGGSound dataset is shown in **Figure 3.3e** and **Figure 3.3f**. Note that auditory modality is the stronger modality. Our method outperforms BADGE in not only multimodal classification but also in two unimodal classification by an obvious margin.

Findings. Our first finding is that AL methods such as BADGE and BALD which win at classification of the stronger modality could stand a good chance of failing at classification of the weak modality. This may be due to biased data selection towards the stronger modality, and it is undesirable for balanced multimodal learning. Our second finding is that Random and CoreSet could perform better in the weaker modality, whereas they are inferior in multimodal classification because random selection treats every sample with absolute fairness and CoreSet focuses too much on the weak modality which are both unfair concerning the stronger



(A) Pairwise comparison on multimodal classification.

(B) Pairwise comparison on unimodal classification of stronger modalities.



(C) Pairwise comparison on unimodal classification of weaker modalities.

FIGURE 3.4: Pairwise comparison of all active learning strategies. Each element in the matrix $P_{i,j}$ represents the number of times strategy i outperforms strategy j . A strategy is considered better if its row-wise value is larger, indicating that it beats other strategies more often. On the other hand, a strategy is better if its column-wise value is smaller, meaning it is rarely beaten by other strategies. The maximum value of each cell is 5, which is the total number of experimental settings. The bottom row displays the column-wise average values (lower is better).

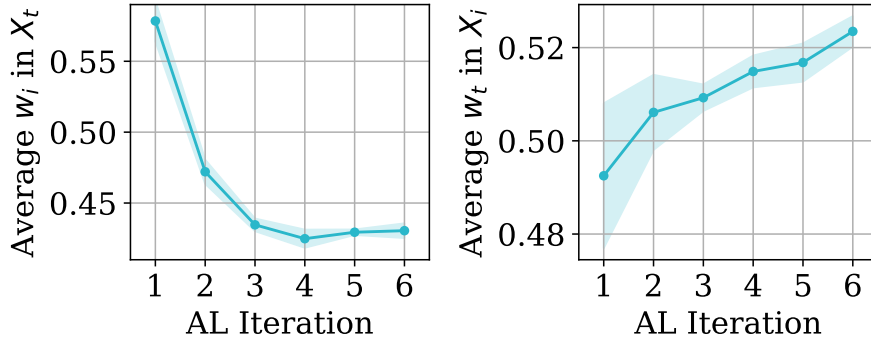
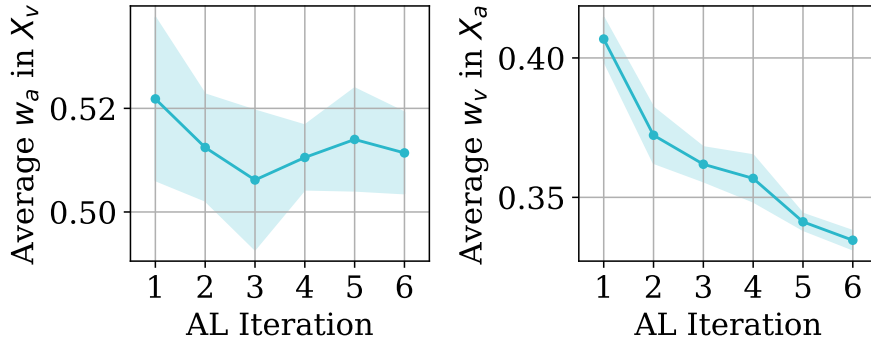
(A) Average weight w_i in X_t and average weight w_t in X_i on the Food101 dataset.(B) Average weight w_a in X_v and average weight w_v in X_a on the KineticsSound dataset.

FIGURE 3.5: Average weight for the weaker modality in a sub-dataset dominated by the other stronger modality.

modality. Finally, our method achieves a fairer multimodal data selection with a better trade-off between weak and strong modalities.

Pairwise Comparison. We illustrate the results across various experimental settings in matrix P in **Figure 3.4** [41]. We compute the t-score for each repeated experiment and use the two-sided t-test to compare the performance of paired strategies on the test set with a 0.9 confidence interval. If strategy i significantly outperforms strategy j , we add $1/L$ to $P_{i,j}$, where L is the total number of iterations for a single experiment setting. The maximum cell value equals the total number of experiment settings. $P_{i,j}$ indicates the number of times strategy i significantly outperforms strategy j . We compute the matrix for both multimodal and unimodal classification for stronger (text for Food101, audio for KineticsSound and VGGSound) and weaker modalities (image for Food101, video for KineticsSound and VGGSound). The three matrices demonstrate that our proposed method outperforms most baselines across settings. Specifically, BMMAL surpasses BADGE in multimodal classification and unimodal classification on weaker modalities, while

performing comparably with BADGE in unimodal classification on stronger modalities. This suggests that the performance improvement of BMMAL in multimodal classification mainly stems from enhancing weaker modalities while maintaining stable performance in stronger modalities.

Dominance Degree. As described in **Equation 3.6**, we divide the entire unlabeled dataset into multiple sub-datasets in which modality m_i contributes the most. The Food101 dataset is divided into X_t and X_i dominated by text and image modality, respectively. In **Figure 3.5a**, the average weight values of the weaker modality are showed. As shown before in **Figure 3.1**, text modality is the stronger one starting from the second iteration. The average value of w_i in X_t accordingly becomes less than that of w_t in X_i from the second iteration, meaning that the average difference value ρ between two unimodal contributions in X_t is larger than in X_i . The KineticsSound dataset is divided into X_v and X_a dominated by video and audio modality, respectively. In **Figure 3.5b**, the average weight values of the weaker modality are showed. Similarly, the average difference value ρ between two unimodal contributions in X_a is larger than in X_v . Consequently, on the dataset-level, the sub-dataset dominated by the weaker modality receives less punishment compared to the sub-dataset dominated by the stronger modality.

3.3.5 Large-scale Active Learning.

We conduct experiment on VGGSound with larger budget size of 5,000 to validate our method on large-scale active learning for multimodal video classification³. The results are averaged and shown in **Table 3.1**. On video classification, the performance of BADGE degrades and becomes worse than random selection, while our method achieves improvement over BADGE and random selection. On audio classification, BADGE and our method are comparable and are both better than random selection. As a result, our method performs better than BADGE and can save around 5k labels compared with random selection if target multimodal classification top-1 accuracy is set to 0.435.

³Due to its large dataset size, we have to split the data pool to store the gradient embeddings. The details can be found in Appendix A.

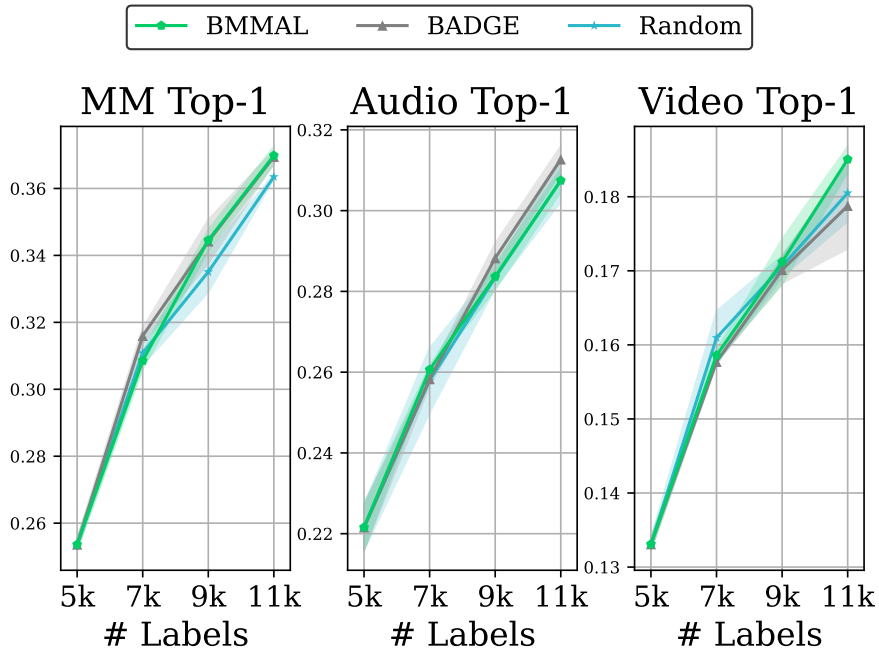


FIGURE 3.6: Multimodal and unimodal classification performance comparison with NL-gate fusion method on the VGGSound dataset.

| | #Labels | 5k | 10k | 15k | 20k | 25k |
|-------------|---------|-------|--------------|--------------|--------------|--------------|
| MM-Top-1 | Random | 0.261 | 0.340 | 0.387 | 0.418 | 0.435 |
| | BADGE | 0.261 | 0.355 | 0.406 | 0.433 | 0.451 |
| | BMMAL | 0.261 | 0.352 | 0.407 | 0.437 | 0.458 |
| Audio-Top-1 | Random | 0.189 | 0.251 | 0.295 | 0.318 | 0.334 |
| | BADGE | 0.189 | 0.262 | 0.307 | 0.332 | 0.345 |
| | BMMAL | 0.189 | 0.261 | 0.308 | 0.333 | 0.350 |
| Video-Top-1 | Random | 0.145 | 0.178 | 0.203 | 0.220 | 0.229 |
| | BADGE | 0.145 | 0.184 | 0.206 | 0.218 | 0.225 |
| | BMMAL | 0.145 | 0.180 | 0.208 | 0.222 | 0.231 |

TABLE 3.1: AL performance on VGGSound dataset with budget size of 5,000. The best results are highlight in bold.

3.3.6 Ablation Studies

Different Fusion Mechanisms. We perform experiment by changing the fusion method from concatenation into summation on Food101 and KineticsSound, while keeping other settings unchanged. We include the performance comparison in the

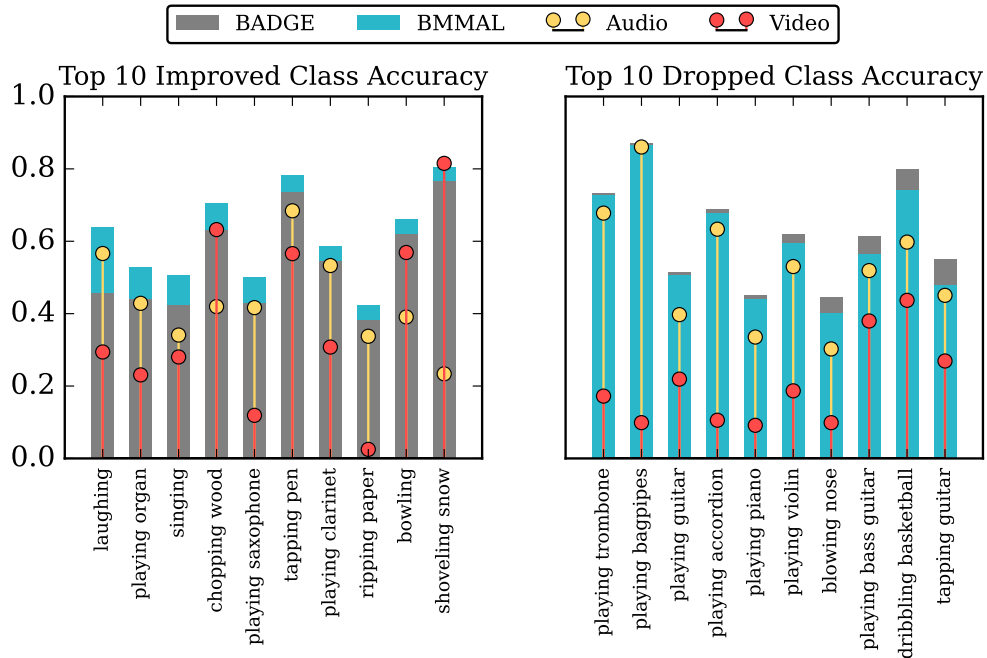


FIGURE 3.7: Top 10 improved and dropped classes based on the improvement of BMMAL to BADGE on multimodal classification accuracy on KineticsSound with 5K labeled samples. Bars represent multimodal classification accuracy. Stems represent unimodal classification accuracy.

pairwise comparison and present the iterative comparison. Furthermore, we change concatenation to NL-gate⁴ for mixing video and audio features on the VGGSound dataset, setting the initial budget to 5,000 and the AL budget for each round to 2,000, as NL-gate requires more data to demonstrate its efficiency in fusion. As shown in **Figure 3.6**, our method achieves comparable multimodal classification performance to BADGE and becomes worse on auditory classification. However, for the weaker visual classification, our method outperforms the others, demonstrating its effectiveness in balancing weak and strong modalities.

Classwise Performance Comparison. We show the classwise performance comparison on the KineticsSound dataset. As shown in **Figure 3.7**, the gain is more significant than the drop. Moreover, improved classes such as 'chopping wood', 'bowling' and 'shoveling snow' carry more visual information, and dropped classes are mostly dominated by the auditory modality. Note that KineticsSound is a dataset where audio contributes more than vision, which means that BMMAL

⁴The implementation of NL-gate can be found in Appendix A.

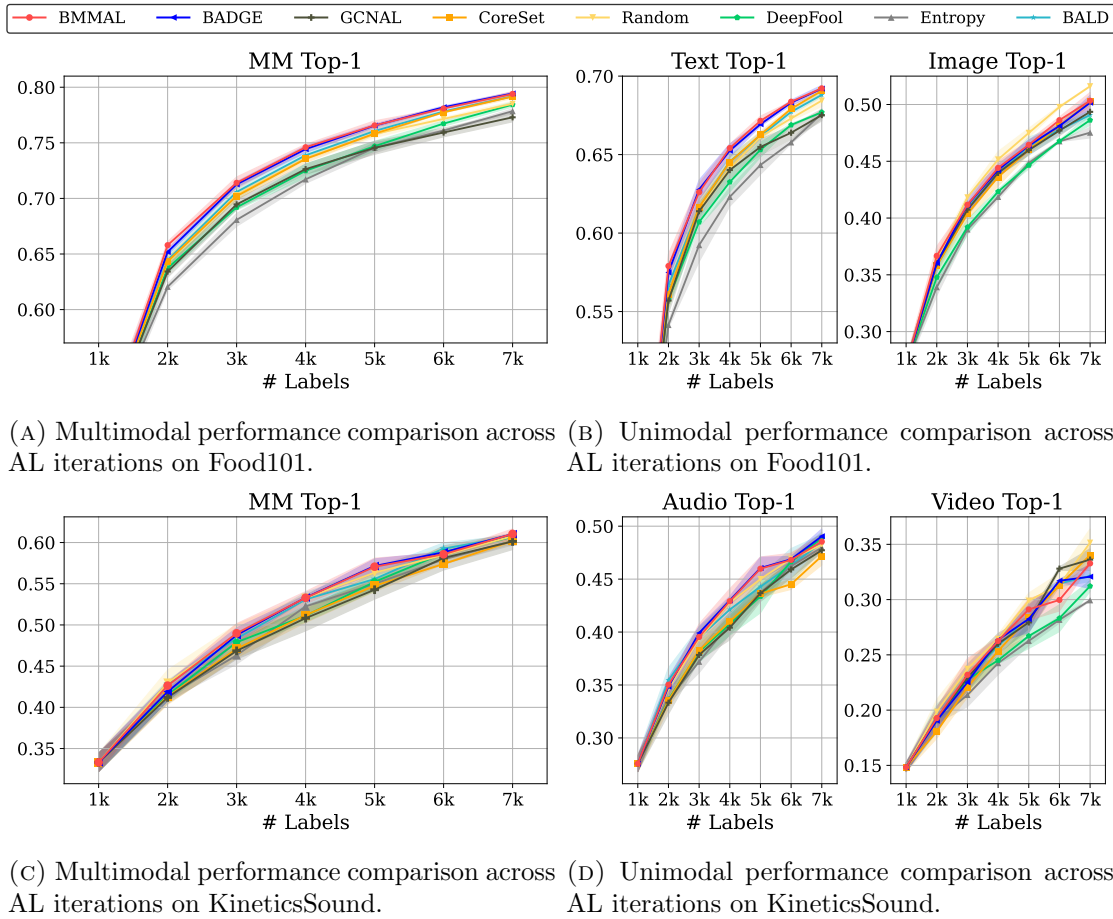


FIGURE 3.8: Performance comparison between proposed method and other conventional AL strategies with Summation fusion method. The metric selected is top-1 accuracy (Top-1) on multimodal and unimodal classification.

avoids biased selection over auditory modality and focuses more on the weaker visual modality.

AL Performance with Summation. We visualize the performance comparison of all baselines with our proposed method in all AL rounds on Food101 and KineticsSound with fusion mechanism of summation in **Figure 3.8**. As shown in the figures, our proposed method outperforms BADGE on Food101 and achieves more balanced unimodal contribution than BADGE. While on KineticsSound, our proposed method is comparable with BADGE, and it may be due to the weak fusion ability of summation.

3.4 Chapter Summary

In this chapter, we evaluate how existing active learning strategies perform on multimodal classification. Our empirical studies show that they might treat different modalities unfairly, and it could lead to performance degradation for multimodal learning. We propose BMMAL to mitigate this unfairness by separately scaling unimodal gradient embeddings, which avoids mixing all unimodal information and well retain characteristics of each modality. The method performs well on multiple datasets and can be potentially applied on large-scale multimodal active learning.

Chapter 4

Enhancing Modality

Representation and Alignment for Multimodal Cold-start Active Learning

4.1 Introduction

In Chapter 3, we introduce an effective active learning (AL) algorithm that can reduce data labeling cost by selectively annotate the most significant data samples with balanced modality contributions. It mitigates the unbalanced multimodal learning from the perspective of data and achieves better performance across three various multimodal benchmarks. However, most AL strategies [105], including the one we proposed in previous chapter, initiate with a randomly selected subset of data, obtaining label annotations for these samples to start the data evaluation and selection, a process known as **Warm-start**. Nevertheless, in practice, assembling a multimodal dataset often starts with limited or even no labels. This poses a challenge for uncertainty-based, diversity-based, and hybrid AL strategies. These strategies rely on initial labels to construct a model capable of assessing the uncertainty and diversity of the data. A lack of sufficient initial labels leads to the **Cold-start** problem in AL, where there is inadequate information to initiate the

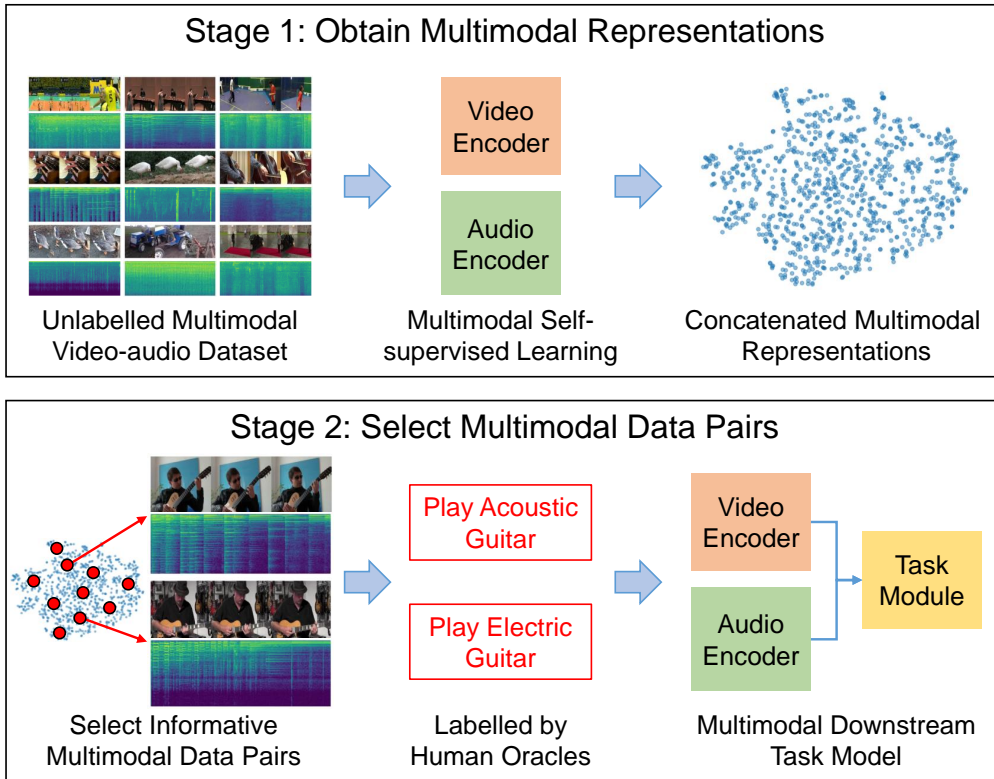


FIGURE 4.1: Multimodal Cold-start Active Learning Framework (using video-audio data as an example): In Stage 1, multimodal self-supervised learning is applied to a large unlabeled dataset to derive concatenated multimodal representations from unimodal features. In Stage 2, a data selection strategy samples informative multimodal data pairs, which are then annotated by human oracles. The labeled samples are subsequently used to train a downstream task model.

active learning loop effectively [43, 53]. In this work¹, we present a novel algorithm to produce high-quality multimodal representations and select well-aligned multimodal data pairs to further reduce label annotation cost under a cold-start AL setting.

Researchers have explored to address the cold-start problem in uni-modal active learning, introducing a two-stage framework incorporating self-supervised learning (SSL) with data selection [43, 49, 51, 53]. However, there exists specific problems in multimodal scenarios expected to be resolved. In **Stage 1**, SSL processes the entire unlabeled dataset to produce structured data representations, aiding in the identification of informative samples and evaluation of sample diversity. In the

¹The work in this chapter has been published as **Meng Shen**, Yake Wei, Jianxiang Yin, Deepu Rajan, Di Hu, and Simon See, “Enhancing Modality Representation and Alignment for Multimodal Cold-start Active Learning”. In *Proceedings of the 6th ACM International Conference on Multimedia in Asia (MM Asia ’24)*.

multimodal domain, SSL typically leverages inherent cross-modal pairing information to guide model training without labels. However, we observe that sole reliance on cross-modal information in multimodal SSL creates a modality gap, where the centroid of each uni-modal representations is far away from another. This modality gap can skew the distance measurements among multimodal data pairs, affecting the data selection process.

In **Stage 2**, after completing SSL to generate semantically rich data representations, the initial batch of data samples is selected for annotation by human oracles. These newly labeled samples are then used to train downstream task models. Existing methods, however, predominantly focus on uni-modal data, such as text sentences [43] or images [49, 51, 53, 56], and may not effectively handle multimodal data pairs. To the best of our knowledge, current cold-start AL approaches do not account for modality alignment in the selection of multimodal data pairs. This oversight neglects the potential for harnessing more complementary information, which could enhance the performance of multimodal downstream tasks.

To overcome the limitation in **Stage 1**, we propose enhancing multimodal self-supervised learning (MMSSL) for AL, aimed at bridging the modality gap and improving distance estimation for multimodal data pairs. Our approach integrates uni-modal prototypes into cross-modal contrastive learning, yielding more balanced representations for both uni-modal and cross-modal feature densities and narrowing the modality gap. This enhancement allows for more precise distance measurements between multimodal data pairs by considering both uni-modal and cross-modal distances, thus improving the data selection process. In **Stage 2**, addressing cross-modal alignment, which has not been previously explored in AL, is crucial for selecting multimodal data pairs. We introduce a regularization term dedicated to optimizing the modality alignment of the selected data subset. **Figure 4.1** illustrates the two stages for addressing the cold-start problem in multimodal AL. Empirical findings suggest that enhancing cross-modal similarity during data selection benefits the training of downstream multimodal tasks by leveraging more complementary information. Consequently, we develop a novel method, **MultiModal Cold-Start Active Learning (MMCSAL)**, which strategically addresses these challenges.

In this chapter, we introduce a multimodal self-supervised learning method that incorporates cross-modal contrastive learning and uni-modal prototypical learning.

This approach aims to reduce the modality gap and provide more accurate distance estimation for multimodal data selection. Moreover, we propose a novel multimodal data selection strategy that maximizes the modality alignment of the selected data subset, leading to improved performance for downstream tasks. Through empirical evaluation, our method shows enhanced ability in selecting multimodal data pairs for cold-start multimodal AL, leading to improved results in both supervised and semi-supervised learning across three multimodal classification datasets: Food101, KineticsSound and VGGSound, covering textual, auditory and visual modalities.

4.2 Proposed Method

In this section, we first introduce the multimodal cold-start active learning framework. Next, we propose a novel multimodal self-supervised learning method that integrates uni-modal prototypes to close the multimodal representation gap. In addition, we propose to enhance modality alignment during data selection to further select more informative multimodal data pairs for cold-start active learning.

4.2.1 Multimodal Cold-start Active Learning Framework

We illustrate the multimodal cold-start active learning framework in **Figure 4.1** and provide details about our method in **Figure 4.2**. In the first stage, we are provided with an unlabeled multimodal dataset $X^U = \{(x_i^{m_1}, x_i^{m_2})\}_{i \in [N]}$, consisting of N data pairs across modalities m_1 and m_2 . Multimodal self-supervised learning is used to train a two-stream multimodal feature encoder, $f^{m_1}(x_i^{m_1})$ and $f^{m_2}(x_i^{m_2})$, for each modality, mapping data samples to representations r^{m_1} and r^{m_2} in the normalized high-dimensional feature space \mathbb{R}^C . The concatenated multimodal representations, $\langle r_i^{m_1} \oplus r_i^{m_2} \rangle$, encapsulate the combined characteristics of both modalities.

In the second stage, we allocate a label budget B , typically ranging from 1% to 10% of the total dataset size N , for acquiring labels from human oracles. Using the multimodal representations $R^U = \{\langle r_i^{m_1} \oplus r_i^{m_2} \rangle\}_{i \in [N]}$ from the first stage, we apply an AL strategy S_{AL} to select a subset $X^S = \{(x_j^{m_1}, x_j^{m_2})\}_{j \in [B]}$ from the entire unlabeled dataset X^U . This budget B is used to assign labels $\{y_j\}_{j \in [B]}$ to the selected

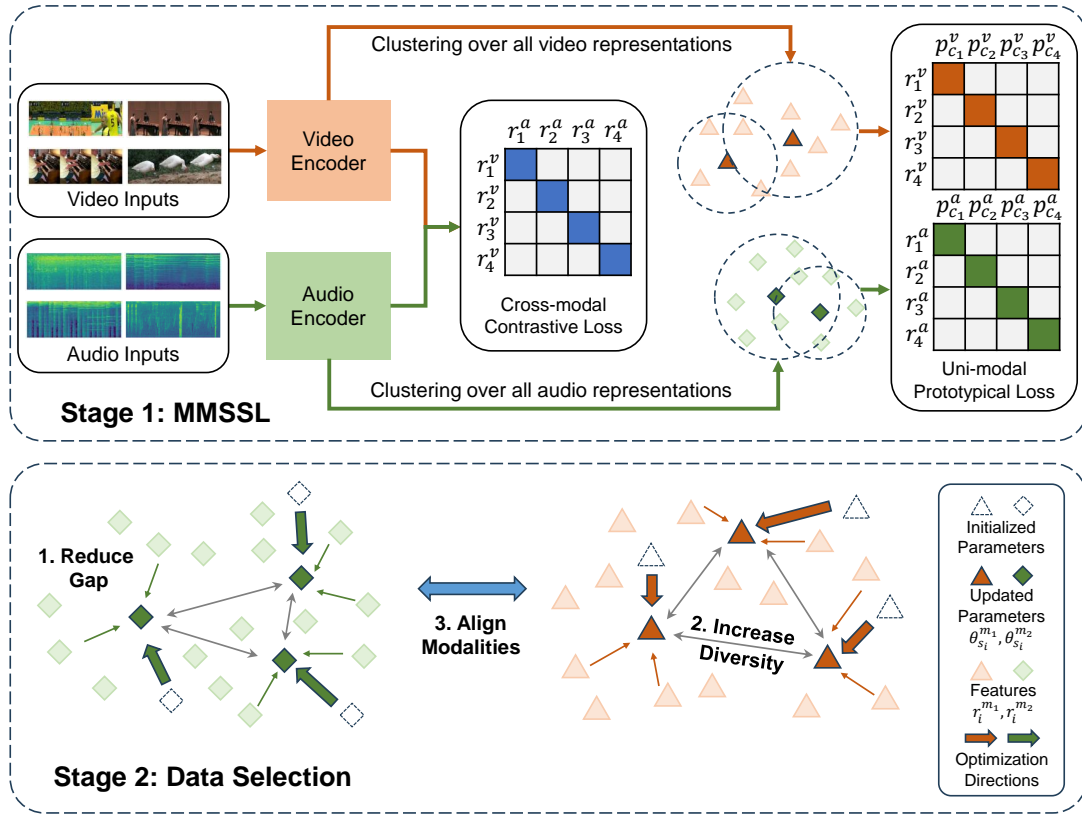


FIGURE 4.2: Our method (use audio/video as an example): In Stage 1, we employ uni-modal prototypical loss and cross-modal contrastive loss; In Stage 2, our selection reduces distribution gap while maintaining diversity and modality alignment.

subset, forming a labeled subset $X^L = \{(x_j^{m_1}, x_j^{m_2}, y_j)\}_{j \in [B]}$. The labeled subset is then used to train a downstream task model, such as an audio-video classification model. Our primary objective is to devise an effective two-stage method. This method will explore a suitable MMSSL for generating multimodal representations in **Stage 1** to provide precise distance estimation among multimodal data pairs, and develop a data selection strategy in **Stage 2** that judiciously select multimodal data pairs, optimizing downstream task performance.

4.2.2 Multimodal Contrastive Learning with Uni-modal Prototypes

The objective of MMSSL is to produce high-quality representations for multimodal data that can accurately capture both the uni- and cross-modal relationships. On the one hand, it is easier to measure the similarity between two multimodal data

pairs by examining their uni-modal components. For example, one would check if two video clips are visually similar and sound similar, rather than comparing the visual aspect of one clip with the audio of another. On the other hand, cross-modal pairing information is crucial as it connects different modalities. Moreover, when conducting contrastive learning with only cross-modal information, we observe a modality gap, i.e., a non-negligible distance between the centroids of two unimodal representations, as shown in Table 4.1. This gap biases our distance estimation of multimodal data pairs. Therefore, we propose using both uni- and cross-modal information to improve representation quality.

To first learn the cross-modal connections, we can seamlessly adopt cross-modal contrastive learning methods such as CLIP [9]. The idea is to bring the representations of two modalities from the same pair closer and push away those from the unpaired data. Specifically, for a batch of K data pairs, the process involves maximizing the cosine similarity between L2-normalized modality representations for the K paired samples and minimizing it for the $K \times (K - 1)$ unpaired modality combinations. We optimize the InfoNCE loss [106], with the temperature parameter τ set to 0.07, as recommended by studies such as [54, 107]:

$$\mathcal{L}_{\text{Cross}} = - \mathbb{E}_{i \in [K]} \left[\frac{1}{2} \log \frac{\exp(\cos(r_i^{m_1}, r_i^{m_2})/\tau)}{\sum_{j \in [K]} \exp(\cos(r_i^{m_1}, r_j^{m_2})/\tau)} + \frac{1}{2} \log \frac{\exp(\cos(r_i^{m_2}, r_i^{m_1})/\tau)}{\sum_{j \in [K]} \exp(\cos(r_i^{m_2}, r_j^{m_1})/\tau)} \right]. \quad (4.1)$$

However, measuring the distance between multimodal sample pairs based solely on cross-modal distance may not be sufficiently accurate due to the modality domain gap. As shown in [72], there is a modality gap in representations trained by contrastive loss, affecting the accuracy of distance measurement between samples which is crucial in our data selection. This modality gap phenomenon is also observed in our experiments across three datasets. As detailed in **Table 4.1**, the Euclidean distance between the centroids of two modality representations Δ_{gap} is larger without uni-modal prototypes. This modality gap is reduced by incorporating uni-modal prototypes that reshape the multimodal representations learnt by cross-modal contrastive learning. Uni-modal prototypes are proposed in PCL [67] and also utilized in PMR [108] to balance multimodal learning. Here, we follow

the implementation from PCL [67] and incorporate it into our method to facilitate reducing the modality gap and restructuring multimodal representations. We visualize our final MMSSL method in the left side of **Figure 4.2**.

We construct uni-modal prototypes individually for each modality. For example, for the first modality m_1 , we perform K-means clustering on its uni-modal representations $\{r_i^{m_1}\}_{i \in [N]}$ to produce C clusters, and the centroid of each cluster is referred to as a prototype p^{m_1} . The same process is performed for the other modality individually. We assign the representation $r_i^{m_1}$ to the prototype $p_{c_i}^{m_1}$ of its cluster c_i as a positive pair. Conversely, we assign it to the prototypes $\{p_{c_j}^{m_1}\}_{j \neq i}$ from other clusters $\{c_j\}_{j \neq i}$ as negative pairs. Based on this, we define the uni-modal prototypical loss as:

$$\mathcal{L}_{\text{Uni}} = - \mathbb{E}_{i \in [K]} \left[\frac{1}{2} \log \frac{\exp(\cos(r_i^{m_1}, p_{c_i}^{m_1}) / \phi_i^{m_1})}{\sum_{j \in [K]} \exp(\cos(r_i^{m_1}, p_{c_j}^{m_1}) / \phi_i^{m_1})} + \frac{1}{2} \log \frac{\exp(\cos(r_i^{m_2}, p_{c_i}^{m_2}) / \phi_i^{m_2})}{\sum_{j \in [K]} \exp(\cos(r_i^{m_2}, p_{c_j}^{m_2}) / \phi_i^{m_2})} \right], \quad (4.2)$$

where ϕ estimates and balances the concentrations ϕ of formed uni-modal clusters. The value of ϕ for each cluster is calculated based on the number of samples in this cluster and the average distance between the representations of this cluster and its prototype. The ϕ for each modality is then defined as:

$$\phi_i^m = \frac{\mathbb{E}_{r_j^m \in C_i^m} \|r_j^m - p_i^m\|_2}{\log(|C_i^m| + \alpha)}, m \in \{m_1, m_2\}, \quad (4.3)$$

where $|C_i^m|$ is the number of samples in i^{th} cluster and $\alpha = 10$ is a smoothing parameter to prevent producing an overly-large ϕ . All the ϕ values are normalized to have a mean of $\tau = 0.07$, ensuring that the uni-modal loss is of similar magnitude to the cross-modal loss. Note that we do not use the technique of hierarchy clustering or over clustering from PCL [67] since our multimodal dataset scale has not reached the level of millions. Our final MMSSL objective is:

$$\mathcal{L} = \frac{1}{2} \mathcal{L}_{\text{Cross}} + \frac{1}{2} \mathcal{L}_{\text{Uni}}. \quad (4.4)$$

To validate that the uni-modal prototypes effectively bridge the modality gap and balance the distance calculations between uni-modal and cross-modal representations, we assess the modality gap [72] as the Euclidean distance between the centroid representations of each modality. Additionally, we measure the representation densities by calculating the expectation of the average cosine similarity within and between modalities across the entire dataset:

$$\begin{aligned}
\mathcal{S}_{\text{Uni}}^{m_1} &= \mathbb{E}_{i \in [N]} \left[\frac{1}{N} \sum_{j \in [N]} \cos(r_i^{m_1}, r_j^{m_1}) \right] \\
\mathcal{S}_{\text{Uni}}^{m_2} &= \mathbb{E}_{i \in [N]} \left[\frac{1}{N} \sum_{j \in [N]} \cos(r_i^{m_2}, r_j^{m_2}) \right] \\
\mathcal{S}_{\text{Cross}} &= \mathbb{E}_{i \in [N]} \left[\frac{1}{N} \sum_{j \in [N]} \cos(r_i^{m_1}, r_j^{m_2}) \right] \\
\Delta_{\text{gap}} &= \left\| \frac{1}{N} \sum_{i \in [N]} r_i^{m_1} - \frac{1}{N} \sum_{i \in [N]} r_i^{m_2} \right\|_2.
\end{aligned} \tag{4.5}$$

These metrics allow us to quantify the effectiveness of the uni-modal prototypes in minimizing the modality gap and enhancing representation similarity assessments. The results are shown in **Table 4.1**. As reflected in the table, multimodal contrastive learning without uni-modal prototypes shows a modality gap between the centroids of the representations of each modality, along with a discrepancy between uni- and cross-modal similarities. With the integration of uni-modal prototypes, we restructure the modality representations, reducing both the modality gap and the disparities between average uni-modal and cross-modal similarities. Consequently, we can more accurately calculate the distances between two multimodal sample pairs, considering both uni-modal and cross-modal distances. Furthermore, in **Section 4.3.4**, we conduct an experiment to compare the cold-start AL performance with and without prototypes.

4.2.3 Select Data with Good Modality Alignment for Cold-start AL

After obtaining the learned multimodal representations through MMSSL, we then conduct multimodal data selection to form the initial batch for cold-start AL. As proposed in ActiveFT [56], the selected subset X^S should have two characteristic: (1) the distribution gap between the representations of the selected subset R^S and

| Datasets | Proto | Δ_{gap} | $\mathcal{S}_{\text{Uni}}^{m_1}$ | $\mathcal{S}_{\text{Uni}}^{m_2}$ | $\mathcal{S}_{\text{Cross}}$ |
|---------------|-------|-----------------------|----------------------------------|----------------------------------|------------------------------|
| Food101 | w/o | 0.2701 | 0.0204 | 0.0256 | -0.0134 |
| | w/ | 0.1203 ↓ | 0.0144 | 0.0159 | 0.0079 |
| KineticsSound | w/o | 0.1643 | 0.0188 | 0.0107 | 0.0012 |
| | w/ | 0.0979 ↓ | 0.0283 | 0.0201 | 0.0199 |
| VGGSound | w/o | 0.1327 | 0.0058 | 0.0044 | -0.0037 |
| | w/ | 0.0483 ↓ | 0.0073 | 0.0052 | 0.0051 |

TABLE 4.1: The modality gap and the expectation of average cosine uni-modal and cross-modal similarity of all samples in three different multimodal datasets. *w/o* represents without prototypes, and *w/* represents with prototypes.

those of the entire unlabeled dataset R^U should be minimized and **(2)** the diversity of the selected subset R^S should be maintained to prevent collapsed selection. We argue that, when selecting data pairs with multiple modalities, the selected subset X^S should have an additional characteristic: **(3)** the cross-modal representation alignment should be maximized within the selected subset for the downstream task model to exploit more shared information across modalities. We first introduce how ActiveFT select data samples by optimizing the parametric model. Then, following the two guidelines from ActiveFT and the third guideline for multimodal data, we propose our method to improve the modal alignment of the initial batch as shown in the right side of **Figure 4.2**.

Since it is hard to directly select data samples in a discrete way, ActiveFT alternatively models the data samples to be selected with the labelling budget B as a parametric model $\Theta^S = \{\theta_j\}_{j \in [B]}$. In our case with two modalities, we model them as $\Theta^S = \{\theta_j^{m_1} \oplus \theta_j^{m_2}\}_{j \in [B]}$, where each parameter is concatenated by two uni-modal parameters, and it will be a multimodal representation corresponding to one selected multimodal data sample pair. Assume the model is optimized to satisfy all the characteristics, then the sample x_j will be selected if its representation $\langle r_i^{m_1} \oplus r_i^{m_2} \rangle$ is closest to the parameter $\langle \theta_i^{m_1} \oplus \theta_i^{m_2} \rangle$, so that we will have a selected subset accordingly. The original optimization goal in ActiveFT [56] is:

$$\Theta_{\text{opt}}^S = \arg \min_{\Theta^S} \mathcal{D}(R^U, \Theta^S) - \lambda_{\text{div}} \mathcal{R}(\Theta^S), \quad (4.6)$$

where $\mathcal{D}(R^U, \Theta^S)$ is the measurement of distribution gap between the representations of the entire unlabeled dataset and our data selection parametric model,

$\mathcal{R}(\Theta^S)$ is the diversity of the parameters and λ_{div} is a hyper-parameter set as 1 in ActiveFT that controls the contribution of the diversity. The L2-norms of the uni-modal parameters $\|\theta_j^{m_1}\|_2$ and $\|\theta_j^{m_2}\|_2$ are always constrained to 1 to align with the normalized data representations.

Given the multimodal representations $\{\langle r_i^{m_1} \oplus r_i^{m_2} \rangle\}_{i \in [N]}$ of N unlabeled samples and the parametric model with B , the budget size, multimodal parameters $\{\langle \theta_j^{m_1} \oplus r_j^{m_1} \rangle\}_{j \in [B]}$, the first term in **Equation 4.6**, $\mathcal{D}(R^U, \Theta^S)$, is calculated as the expectation of the similarity between the representation r_i from the unlabeled dataset and its closest parameter θ_{s_i} with temperature $\tau = 0.07$:

$$\begin{aligned} \mathcal{D}(R^U, \Theta^S) &= - \mathbb{E}_{r_i \in R^U} \left[\frac{\cos(r_i^{m_1}, \theta_{s_i}^{m_1}) + \cos(r_i^{m_2}, \theta_{s_i}^{m_2})}{2\tau} \right], \\ s_i &= \arg \max_{j \in [B]} [\cos(r_i^{m_1}, \theta_j^{m_1}) + \cos(r_i^{m_2}, \theta_j^{m_2})]. \end{aligned} \quad (4.7)$$

When minimizing $\mathcal{D}(R^U, \Theta^S)$, the parameters will be updated to have higher similarity towards the entire dataset, thus reducing the distribution gap between selected subset with the unlabeled dataset. To prevent selected data sample collapse into the centroids of unlabeled dataset, we need to increase the diversity of the selected subset. The second term of diversity is then defined as the similarity between the parameters in the parametric model. We add the exponential operation, same as ActiveFT, to stabilize the optimization process.

$$\mathcal{R}(\Theta^S) = - \mathbb{E}_{j \in [B]} \left[\log \sum_{k \neq j} \exp \left(\frac{\cos(\theta_j^{m_1}, \theta_k^{m_1}) + \cos(\theta_j^{m_2}, \theta_k^{m_2})}{2\tau} \right) \right]. \quad (4.8)$$

However, when optimizing with **Equation 4.6**, the cross-modal alignment is not guaranteed to be maximized. Although the selected subset can perfectly satisfy the objective of ActiveFT, the absence of cross-modal interaction during the optimization process means some of the selected data pairs may not have good multimodal alignment, meaning that their cross-modal similarity is low. In this case, when feeding not well-aligned multimodal data to train a downstream task model, the advantage of multimodal network is not fully exploited because of less complementary information the selected data provided with.

To overcome this issue, the most straightforward way is to add the cross-modal distribution gap into the uni-modal distribution gap term in **Equation 4.7**. The term then becomes:

$$\mathcal{D}_{\text{MM}}(R^U, \Theta^S) = - \mathbb{E}_{r_i \in R^U} \left[\frac{\cos(r_i^{m_1}, \theta_{s_i}^{m_1}) + \cos(r_i^{m_2}, \theta_{s_i}^{m_2})}{2\tau} + \frac{\cos(r_i^{m_1}, \theta_{s_i}^{m_2}) + \cos(r_i^{m_2}, \theta_{s_i}^{m_1})}{2\tau} \right], \quad (4.9)$$

where the s_i is updated to the index of the parameter that has the least uni- and cross-modal distances to the representation r_i . By adding cross-modal similarity into consideration, when minimizing the distribution gap term, both cross-modal and uni-modal similarities are expected to increase across the unlabeled dataset. However, this term will be highly similar to our self-supervised training objective which has already been optimized during the representation learning process. Therefore, the solution of optimizing the distribution gap may be trivial. Another important factor is that we care more about the how well the modalities are aligned in the selected subset than with the entire unlabeled dataset. We show its cold-start AL performance in **Table 4.2**. It outperforms ActiveFT when the labeling budget is low, showing that choosing data with better modality alignment will help for downstream multimodal task. However, it becomes less advantageous when more budget becomes available due to the issues discussed above.

To address the issues of the naive multimodal version of ActiveFT, which is limited by its focus on optimizing already-learned objectives, aligning modalities only within selected subsets, and diminishing effectiveness as labeling budgets increase, we propose a novel regularization term that specifically maximizes the modality alignment only for the selected subset, independently calculated from the original uni-modal distribution gap term and the diversity term. We design this cross-modal alignment regularization term as the average cross-modal similarity between each parameter with all the other parameters in the data selection parametric model:

$$\mathcal{A}(\Theta^S) = \mathbb{E}_{j \in [B]} \left[\log \sum_{k \in [B]} \exp(\cos(\theta_j^{m_1}, \theta_k^{m_2})/\tau) \right]. \quad (4.10)$$

We use the same exponential operation to ensure numerical stabilization. By including this alignment regularization term with a hyper-parameter λ_{align} , controlling the contribution of cross-modal alignment, our final optimization objective then becomes:

$$\begin{aligned}
\mathcal{L} &= \mathcal{D}(R^U, \Theta^S) - \lambda_{\text{div}} \mathcal{R}(\Theta^S) + \lambda_{\text{align}} \mathcal{A}(\Theta^S) \\
&= - \mathbb{E}_{r_i \in R^U} \left[\frac{\cos(r_i^{m_1}, \theta_{s_i}^{m_1}) + \cos(r_i^{m_2}, \theta_{s_i}^{m_2})}{2\tau} \right] \\
&\quad - \lambda_{\text{div}} \mathbb{E}_{j \in [B]} \left[\log \sum_{k \neq j} \exp \left(\frac{\cos(\theta_j^{m_1}, \theta_k^{m_1}) + \cos(\theta_j^{m_2}, \theta_k^{m_2})}{2\tau} \right) \right] \\
&\quad + \lambda_{\text{align}} \mathbb{E}_{j \in [B]} \left[\log \sum_{k \in [B]} \exp (\cos(\theta_j^{m_1}, \theta_k^{m_2}) / \tau) \right].
\end{aligned} \tag{4.11}$$

After optimizing the parametric model with this objective, we select the samples exhibiting the smallest uni-modal distances with the parameters $\Theta^S = \{\theta_j^{m_1} \oplus \theta_j^{m_2}\}_{j \in [B]}$. These samples form our labeled subset $X^S = \{(x_j^{m_1}, x_j^{m_2})\}_{j \in [B]}$, which is then used to train a multimodal downstream task model.

4.3 Experiments

In this section, we design experiments to validate the effectiveness of our method. We evaluate our method across three different multimodal benchmark datasets. We compare the performance of our method with both warm-start and cold-start AL baselines under fully-supervised setting and semi-supervised setting.

4.3.1 Dataset

We choose three different multimodal classification datasets:

KineticsSound [96] is a subset of Kinetics-400 [97]. It consists of 31 action classes which are all correlated to both video and audio signals. To obtain more data pairs for us to perform self-supervised learning, we include all available video clips of these 31 action classes from Kinetics-400, resulting in 22,588 video clips for training and 3,012 video clips for testing.

Food101 [25] is an image-text classification dataset for food recipe recognition with 101 kinds of food, where each recipe is composed of a food image and a text recipe description. There are 45,719 sample pairs for training and 15,294 sample pairs for testing.

VGGSound [98] is a large video-audio dataset with 309 classes where each video clip captures the object that makes the sound. We are only able to download 180,911 clips for training and 14,843 clips for testing. Some of the video clips could not be downloaded due to their unavailability on the YouTube website.

4.3.2 Baseline

We compare our method with random selection, four cold-start AL strategies and four warm-start AL methods. **Random** selects the data samples randomly from the unlabeled data pool.

Cold-start AL. (1) **K-means** forms B clusters over multimodal representations and picks the center sample of each cluster as queried samples. (2) **KGC** (K-greedy-center) [112] greedily selects the sample that has maximum distance from previously selected samples. The Euclidean distance over concatenated multimodal representations is utilized as the distance metric. (3) **ActiveFT** [56] models the data selection process as an optimization process that can minimize the distribution gap between the selected subset and the entire unlabeled dataset while still maintaining the diversity. (4) **MMActiveFT** adds the cross-modal distance into the distribution gap calculation as in **Equation 4.9** to minimize both uni-modal and cross-modal distribution gap between selected subset and the unlabeled dataset.

Warm-start AL. (1) **BADGE** [109] extracts the gradient embedding of multimodal classifier and employs K-means++ initialization method to select diverse and informative data samples for label annotations. (2) **BMMAL** [110] achieves balanced multimodal data selection by modulating the gradient embedding of each uni-modal classifier with modality contribution scores. (3) **GCNAL** [100] builds a graph convolution network that learns how to classify labeled and unlabeled samples. It chooses those unlabeled samples that are difficult to distinguish. (4)

| AL Method | Proto | Food101 | | | |
|----------------|-------|------------------|------------------|------------------|------------------|
| | | 1% | 2% | 5% | 10% |
| Random | - | 27.9±1.5 | 48.1±1.4 | 67.7±0.2 | 75.5±0.1 |
| BADGE [109] | - | - | 48.4±1.1 | 68.0±0.3 | 76.0±0.3 |
| BMMAL [110] | - | - | 48.0±1.8 | 68.1±0.4 | 76.0±0.1 |
| GCNAL [100] | - | - | 43.0±1.1 | 59.5±2.2 | 69.5±2.3 |
| ALFA-Mix [111] | - | - | 46.2±1.0 | 60.2±0.8 | 70.0±0.3 |
| KGC [112] | w/o | 15.3±1.2 | 26.1±1.2 | 52.6±1.0 | 71.2±0.3 |
| | w/ | 19.2±0.6↑ | 33.0±0.4↑ | 58.4±0.9↑ | 72.7±0.1↑ |
| K-means | w/o | 23.9±0.5 | 37.7±0.5 | 47.9±0.4 | 59.9±0.4 |
| | w/ | 20.5±0.2↓ | 37.9±0.4↑ | 52.8±0.9↑ | 63.6±0.3↑ |
| ActiveFT [56] | w/o | 34.5±0.9 | 52.6±1.0 | 68.9±0.5 | 76.1±0.2 |
| | w/ | 35.2±0.7↑ | 53.2±0.7↑ | 69.2±0.4↑ | 76.3±0.4↑ |
| MMActiveFT | w/o | 33.5±1.5 | 51.6±0.9 | 68.8±0.4 | 76.4±0.3 |
| | w/ | 36.3±1.4↑ | 53.0±0.9↑ | 69.1±0.3↑ | 76.4±0.2↑ |
| MMCSAL(Ours) | w/o | 34.6±0.2 | 52.7±0.8 | 68.8±0.4 | 76.5±0.3 |
| | w/ | 36.7±1.3↑ | 53.7±0.3↑ | 69.7±0.3↑ | 76.7±0.2↑ |
| | | KineticsSound | | | |
| | | 1% | 2% | 5% | 10% |
| Random | - | 18.9±1.4 | 26.5±0.9 | 37.4±1.3 | 46.5±1.0 |
| BADGE [109] | - | - | 26.2±0.7 | 36.9±1.4 | 46.0±0.8 |
| BMMAL [110] | - | - | 25.6±1.3 | 37.2±0.7 | 46.6±0.8 |
| GCNAL [100] | - | - | 22.3±0.8 | 24.9±1.3 | 32.9±2.2 |
| ALFA-Mix [111] | - | - | 26.0±1.6 | 36.8±0.6 | 44.9±3.3 |
| KGC [112] | w/o | 14.6±1.1 | 19.4±1.5 | 29.0±1.2 | 38.4±1.4 |
| | w/ | 18.0±0.8↑ | 23.3±0.7↑ | 33.3±1.5↑ | 43.1±1.2↑ |
| K-means | w/o | 20.9±0.6 | 22.9±0.5 | 32.8±0.6 | 35.9±1.0 |
| | w/ | 16.2±0.4↓ | 23.9±0.8↑ | 30.0±0.7↓ | 38.4±0.7↑ |
| ActiveFT [56] | w/o | 21.9±0.9 | 28.5±0.9 | 38.2±0.7 | 47.0±0.7 |
| | w/ | 22.4±1.6↑ | 29.0±0.6↑ | 38.8±0.6↑ | 47.2±0.9↑ |
| MMActiveFT | w/o | 22.7±0.9 | 28.8±0.6 | 38.2±1.2 | 43.9±1.9 |
| | w/ | 23.0±1.5↑ | 29.0±1.5↑ | 38.6±1.0↑ | 47.1±1.0↓ |
| MMCSAL(Ours) | w/o | 22.6±1.0 | 28.5±1.1 | 38.8±0.8 | 47.4±1.1 |
| | w/ | 22.3±1.5↓ | 29.1±0.4↑ | 39.7±1.0↑ | 47.9±0.6↑ |
| | | VGGSound | | | |
| | | 1% | 2% | 5% | 10% |
| Random | - | 14.7±0.4 | 21.9±0.3 | 31.7±0.5 | 38.6±0.4 |
| BADGE [109] | - | - | 22.1±0.5 | 32.4±0.4 | 39.7±0.2 |
| BMMAL [110] | - | - | 22.0±0.2 | 32.9±0.2 | 40.0±0.3 |
| GCNAL [100] | - | - | 20.2±0.4 | 28.9±0.8 | 35.7±1.1 |
| ALFA-Mix [111] | - | - | 20.9±0.4 | 25.3±0.3 | 34.5±0.1 |
| KGC [112] | w/o | 12.0±0.4 | 17.4±0.4 | 28.4±0.2 | 36.7±0.1 |
| | w/ | 13.9±0.6↑ | 20.0±0.2↑ | 29.6±0.4↑ | 37.8±0.5↑ |
| K-means | w/o | 12.1±0.2 | 15.4±0.1 | 22.9±0.3 | 30.1±0.2 |
| | w/ | 12.2±0.3↑ | 15.8±0.2↑ | 24.0±0.2↑ | 30.3±0.1↑ |
| ActiveFT [56] | w/o | 18.7±0.2 | 24.9±0.4 | 33.1±0.5 | 39.3±0.1 |
| | w/ | 18.6±0.3↓ | 24.5±0.3↓ | 33.4±0.2↑ | 39.4±0.5↑ |
| MMActiveFT | w/o | 16.3±0.4 | 22.6±0.6 | 31.8±0.2 | 38.8±0.4 |
| | w/ | 19.4±0.3↑ | 25.5±0.2↑ | 33.9±0.3↑ | 39.8±0.2↑ |
| MMCSAL(Ours) | w/o | 18.8±0.4 | 25.1±0.4 | 33.9±0.4 | 39.8±0.3 |
| | w/ | 19.2±0.2↑ | 25.6±0.2↑ | 34.0±0.3↑ | 40.0±0.3↑ |

TABLE 4.2: The supervised cold-start AL experiment results. We report the mean and std of Top-1 accuracy with multiple runs. The best is marked **bold**, the second best is underlined. *w/o* and *w/* represent without and with uni-modal prototypes. Red ↑ and blue ↓ represent the performance is improved or degraded after integrating uni-modal prototypes.

| AL Method | Food101 | | | |
|----------------|-----------------|-----------------|-----------------|-----------------|
| | 1% | 2% | 5% | 10% |
| Random | 61.9±1.6 | 72.2±0.3 | 77.7±0.3 | 80.1±0.1 |
| BADGE [109] | - | 72.7±0.6 | 78.3±0.1 | 81.1±0.2 |
| BMMAL [110] | - | 72.4±0.1 | 78.4±0.1 | 81.1±0.1 |
| GCNAL [100] | - | 67.9±1.2 | 70.8±1.1 | 72.9±0.3 |
| ALFA-Mix [111] | - | 73.6±0.5 | 78.1±0.4 | 79.8±0.5 |
| KGC [112] | 59.4±1.2 | 69.3±0.8 | 77.5±0.4 | 81.0±0.2 |
| K-means | 44.5±0.6 | 62.6±0.4 | 70.8±0.1 | 75.1±0.2 |
| ActiveFT [56] | 69.4±0.6 | 74.2±0.8 | 78.5±0.1 | 80.8±0.5 |
| MMActiveFT | 69.8±0.4 | 74.0±0.1 | 78.6±0.1 | 80.9±0.3 |
| MMCSAL(Ours) | <u>69.7±1.0</u> | 74.4±0.5 | 78.9±0.2 | 81.2±0.2 |
| | KineticsSound | | | |
| | 1% | 2% | 5% | 10% |
| Random | 31.6±2.1 | 41.8±1.4 | 51.8±0.9 | 57.1±0.4 |
| BADGE [109] | - | 41.7±1.5 | 52.1±0.8 | 57.0±0.7 |
| BMMAL [110] | - | 41.8±1.2 | 51.7±0.6 | 57.0±0.4 |
| GCNAL [100] | - | 36.1±1.9 | 41.1±2.1 | 49.1±2.0 |
| ALFA-Mix [111] | - | 41.8±1.4 | 52.1±0.4 | 56.9±0.6 |
| KGC [112] | 31.6±0.7 | 40.7±1.3 | 50.8±0.9 | 56.8±0.3 |
| K-means | 25.0±0.9 | 34.7±1.1 | 43.3±0.8 | 48.0±0.6 |
| ActiveFT [56] | <u>37.1±1.1</u> | 45.0±0.7 | 52.9±0.6 | 57.6±0.9 |
| MMActiveFT | 37.2±1.6 | 45.7±1.3 | 53.2±0.7 | 57.8±0.3 |
| MMCSAL(Ours) | 36.6±1.8 | 45.7±1.1 | 53.8±0.7 | 57.9±0.6 |
| | VGGSound | | | |
| | 1% | 2% | 5% | 10% |
| Random | 29.8±0.4 | 35.8±0.3 | 42.2±0.3 | 45.7±0.4 |
| BADGE [109] | - | 37.1±0.4 | 43.3±0.2 | <u>47.2±0.1</u> |
| BMMAL [110] | - | 36.8±0.2 | 43.0±0.3 | 47.3±0.2 |
| GCNAL [100] | - | 34.7±0.2 | 40.3±0.2 | 43.9±0.2 |
| ALFA-Mix [111] | - | 35.2±0.5 | 40.9±0.4 | 44.4±0.5 |
| KGC [112] | 31.3±0.7 | 36.8±0.4 | 43.0±0.3 | 46.6±0.4 |
| K-means | 23.4±0.1 | 28.0±0.1 | 36.6±0.3 | 40.8±0.3 |
| ActiveFT [56] | <u>33.6±0.2</u> | <u>38.1±0.3</u> | <u>43.4±0.1</u> | 46.1±0.3 |
| MMActiveFT | 31.1±0.6 | 36.3±0.1 | 42.1±0.4 | 45.6±0.3 |
| MMCSAL(Ours) | 34.3±0.3 | 38.7±0.3 | 43.9±0.6 | 47.1±0.4 |

TABLE 4.3: The semi-supervised cold-start AL results. We report the mean and std of Top-1 accuracy with multiple runs.

ALFA-Mix [111] interpolates the features of unlabeled samples with those of labeled samples, and marks the unlabeled samples as informative ones if their model predictions change after mixing features. The K-means is applied to ensure the diversity of selected subset.

4.3.3 Experiment Setting

Multimodal SSL. We set the temperature τ in **Equation 4.4** as 0.07 for both cross-modal contrastive learning and uni-modal prototypical learning. The batch sizes are all set as 256 for all three datasets. The numbers of uni-modal prototypes are set to 500, 1000, 5000 for KineticsSound, Food101 and VGGSound, respectively, given their different dataset sizes.

Data Selection. We fix the diversity hyper-parameter λ_{div} to 1.0 and the temperature τ in **Equation 4.11** to 0.07 following the setting of ActiveFT [56]. For the alignment hyper-parameter λ_{align} , we set it to 1.0 for all experiments except the ablation study about this parameter in **Section 4.3.6**.

Other Settings. For Food101, we choose pretrained ResNet-101 [113] as the image backbone and pretrained Bert-base model [44] as the text backbone. For KineticsSound and VGGSound, we use ResNet2P1D-18 [102] as video backbone and ResNet-18 as the audio backbone, modifying the input channel from 3 to 1. We use AdamW [114] as the optimizer to train both multimodal SSL and downstream classification models. We repeat 10 runs for Food101 and KineticsSound and 5 runs for VGGSound.

4.3.4 Supervised Cold-start AL

We examine the performance of different AL strategies in a cold-start setting where only 1% to 10% labeling budget is available. **Table 4.2** shows the performance comparison using *Top-1 Accuracy* as the evaluation metric. The performance of warm-start AL methods with 1% labeling data is not shown because they require a random initial subset for model training, which aids in assessing the uncertainty and the diversity of data. The initially insufficient labeling budget leads to a poorly calibrated model, which fails to guide these warm-start AL methods in selecting informative samples. Consequently, these methods are not as competitive and may even perform worse than random data selection in a cold-start setting with an inadequate label budget. In cold-start AL, KGC and K-means are two diversity-based methods that underperform. KGC fails because its greedy selection prioritizes

outliers that are difficult for the models to learn, as evidenced in **Section 4.4.2**. K-means fails because it selects only the typical data samples, the centroids of clusters, without selecting samples with more uncertainty, thereby providing less novel information. As the table reflects, as more labeling budget becomes available, the performance gap between K-means and other methods widens, due to its tendency to select samples based solely on representativeness, neglecting novel knowledge. The multimodal version of ActiveFT, MMActiveFT, which replaces the uni-modal distribution gap with the cross- and uni-modal gap as shown in **Equation 4.9**, performs well even with an extremely low labeling budget. However, its performance diminishes with additional annotations. Therefore, simply using uni- and cross-modal distances as distribution gap is not robust for selecting multimodal data. MMCSAL, by independently enhancing cross-modal alignment within the selected subset, brings more complementary cross-modal information than ActiveFT and avoids the naive data selection method in MMActiveFT, resulting in a robust performance improvement.

Furthermore, to demonstrate the effectiveness of uni-modal prototypes for cold-start multimodal AL, we compare the supervised AL performance with and without uni-modal prototypes. **Table 4.2** shows the results. As discussed in **Sec 4.2.2**, our introduced uni-modal prototypes help bridge the modality gap and balance densities of uni-modal representations, making the representations more organized. Therefore, it estimates the distances between multimodal pairs more precisely and leads to an overall performance improvement for different cold-start AL methods.

4.3.5 Semi-supervised Cold-start AL

We conduct semi-supervised learning in cold-start AL where the initially selected labeled samples are used to fine-tune the pretrained models produced with MMSSL. The results are shown in **Table 4.3**. As expected, the semi-supervised learning boosts the downstream task performance by a large margin compared with supervised learning as the model sees more data. In semi-supervised learning, when labeling budget is 10%, the warm-start AL methods start to dominate the performance as they can train a well-calibrated model with sufficient data and guide the model to select more informative samples for future AL iterations. However, when available labeling budget is lower than 10%, warm-start AL methods are not

| | 0.0 | 0.25 | λ_{align} 0.50 | 1.0 | 2.0 |
|-----|---------------|-----------------|----------------------------------|-----------------|----------|
| | Food101 | | | | |
| 1% | 35.2±0.7 | 36.2±0.7 | 36.0±1.0 | 36.7±1.3 | 36.6±1.4 |
| 2% | 53.2±0.7 | 53.3±1.2 | 53.6±0.6 | 53.7±0.3 | 53.2±0.6 |
| 5% | 69.2±0.4 | 69.3±0.4 | 69.4±0.1 | 69.7±0.3 | 69.4±0.2 |
| 10% | 76.3±0.4 | 76.5±0.2 | 76.4±0.2 | 76.7±0.2 | 75.6±0.3 |
| | KineticsSound | | | | |
| 1% | 22.4±1.6 | 23.0±1.2 | 22.7±1.3 | 22.3±1.5 | 22.1±1.1 |
| 2% | 29.0±0.6 | 28.7±0.9 | 29.5±0.7 | 29.1±0.4 | 28.7±0.9 |
| 5% | 38.8±0.6 | 38.6±1.1 | 38.5±0.7 | 39.7±1.0 | 38.9±0.6 |
| 10% | 47.2±0.9 | 47.1±0.6 | 47.8±0.3 | 47.9±0.6 | 47.7±0.6 |
| | VGGSound | | | | |
| 1% | 18.6±0.3 | 19.2±0.3 | 19.3±0.2 | 19.2±0.2 | 18.7±0.2 |
| 2% | 24.5±0.3 | 25.4±0.3 | 25.4±0.5 | 25.6±0.2 | 24.9±0.2 |
| 5% | 33.4±0.2 | 33.9±0.2 | 33.6±0.3 | 34.0±0.3 | 33.5±0.3 |
| 10% | 39.4±0.5 | 39.5±0.2 | 40.0±0.4 | 40.0±0.3 | 38.4±0.3 |

TABLE 4.4: The ablation study of λ_{align} .

competitive with cold-start AL methods or even close to random data selection. This demonstrates the effectiveness of conducting cold-start AL strategies when labeling information is not sufficient for semi-supervised learning. Overall, our proposed method MMCSAL is still effective when labeling budget is constrained in semi-supervised learning, showing that enhancing modality alignment during data selection is a crucial step for multimodal active learning.

4.3.6 Ablation Study of λ_{align}

In supervised cold-start AL experiment, we notice that MMCSAL performs unwell when labeling budget is 1% on KineticsSound. We suspect that the reason is when labeling budget is low, the diversity of selected data samples is more important than other aspects such as informativeness [51]. We vary the hyper-parameter λ_{align} in **Equation 4.11** to control the contribution of cross-modal alignment in the selected sub-set. When λ_{align} equals to 0.0, it becomes original ActiveFT without any cross-modal alignment enhancement, and when λ_{align} equals to 1.0, it is the default setting for MMCSAL. As shown in **Tabel 4.4**, when λ_{align} is increased from 0.0 to 1.0, the supervised AL performance increases expect for the situations

where the labeling budget is extremely low for KineticsSound and VGGSound. If we keep increasing λ_{align} up to 2.0, the supervised AL performance decreases. It shows that enhancing cross-modal alignment for the selected subset does improve multimodal data quality for downstream multimodal tasks. While the diversity may be the priority when the labeling budget is extremely low, one may consider using a smaller λ_{align} to balance for diversity. Overall, we recommend to use 1.0 as the default value for λ_{align} .

4.4 Data Selection Preference of AL

We first explain how we derive the figures of data preference. We then provide additional results on the data selection preferences of cold-start active learning strategies on two other datasets, KineticsSound and VGGSound.

4.4.1 Rank the Samples by Confidence

We perform supervised multimodal classification training with the entire labeled dataset as in [55]. We calculate the confidence score as the average prediction probability across all training epochs:

$$\mu_i = \frac{1}{T} \sum_{t=1}^T p(y_i|x_i), \quad (4.12)$$

where T is the number of epochs and $p(y_i|x_i)$ is the prediction probability for the ground truth label y_i . We then rank all data samples in ascending order and use the normalized ranking scores r_i to indicate their confidence levels:

$$r_i = \frac{\text{rank}(\mu_i)}{N}, \quad (4.13)$$

where N is the number of all samples. A sample with a lower r_i is considered uncertain, while a sample with a higher r_i is considered confident. We group all samples into five groups with different ranges of normalized ranking scores. The bars in the data selection preference figures represent the numbers of selected samples by

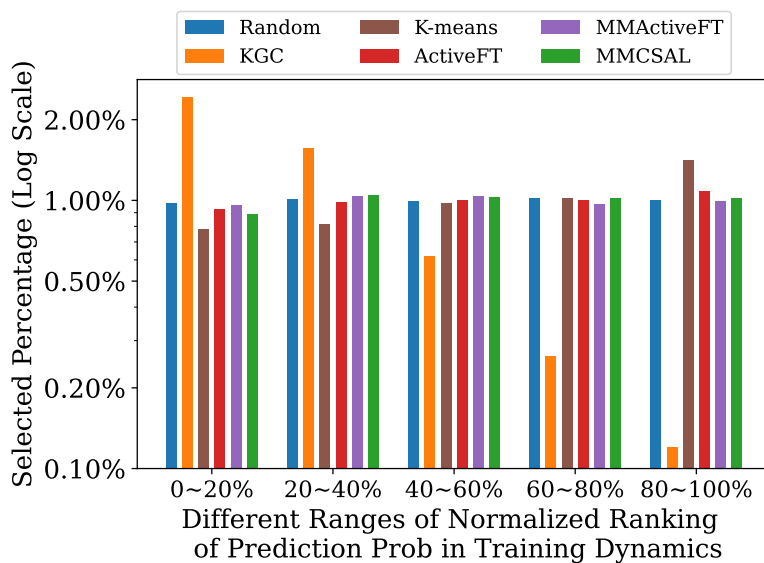


FIGURE 4.3: Preference for data selection of different AL strategies with 5% labeling budget on Food101.

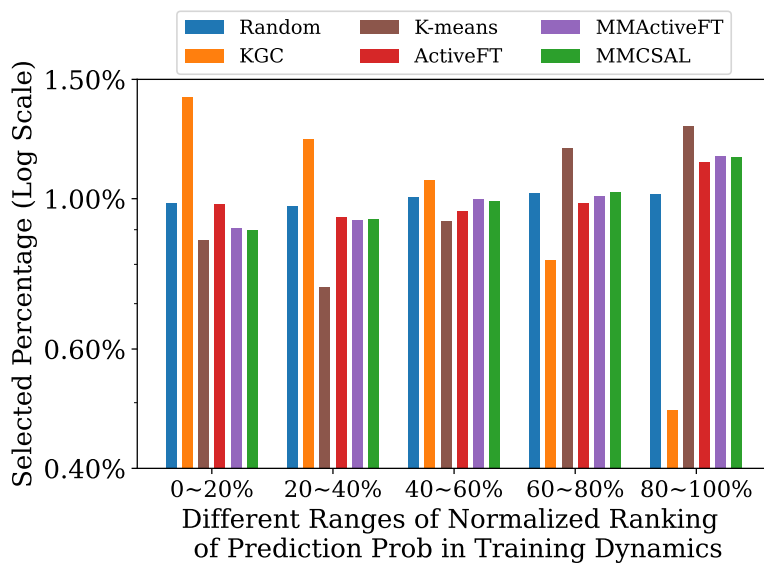


FIGURE 4.4: Preference for data selection of different AL strategies with 5% labeling budget on KineticsSound.

each cold-start active learning strategy, thereby showing the selection preference for uncertain and confident samples by different active learning strategies.

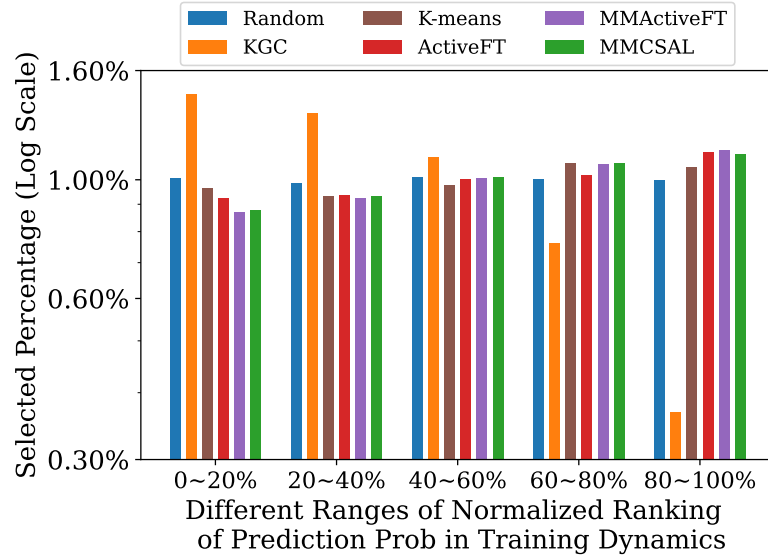


FIGURE 4.5: Preference for data selection of different AL strategies with 5% labeling budget on VGGSound.

4.4.2 Analyze the Data Preference of AL

In this analysis, we visualize the preference for data selection of cold-start AL strategies by counting the numbers of confident, ambiguous and uncertain samples in selected subset. We firstly rank all the samples by their average prediction probability over the ground truth class in an ascending order. This probability is calculated from the training dynamics of supervised multimodal classification [55]. A high-ranking sample is a confident one, while a low-ranking one is an uncertain one. The samples in the middle are the ambiguous samples. As shown in the **Figure 4.3**, KGC selects more samples that are uncertain due to its greedy data selection strategy. It explains its failure in cold-start active learning. K-means selects more samples that are confident given the fact that it selects the centroids of clusters which are often typical and easy-to-learn. Compared with ActiveFT, MMCSAL tends to select less samples that are the most and the least confident ones. The most confident samples are believed to be perfectly aligned but they contribute less on diversity as they are also more similar to other samples. The least confident samples are mostly poorly aligned and they contribute less on modality alignment. Since MMCSAL optimizes both the diversity and the modality alignment within the selected subset, these ambiguous samples may contribute more than these most and least confident samples. It shows that MMCSAL achieves a

more balanced data selection between diversity and modality alignment for multimodal data, leading to better cold-start AL performance.

We provide data selection preference on KineticsSound and VGGSound. As shown in **Figure 4.4** and **Figure 4.5**, KGC selects the most uncertain samples among these AL strategies. K-means selects more certain samples in KineticsSound, while it selects more uniformly on VGGSound due to the larger number of data samples. Our method, MMCSAL, selects fewer uncertain multimodal data pairs compared to ActiveFT, as these uncertain multimodal data samples are often poorly aligned. This indicates that MMCSAL prefers well-aligned multimodal samples to enhance multimodal learning. Compared with MMActiveFT, our method selects fewer samples that are most confident, thereby increasing diversity. This demonstrates that our method achieves a more balanced data selection between modality alignment and diversity.

4.5 Implementation Details

In this section, we provide more details about our implementation of multimodal self-supervised learning and both supervised and semi-supervised cold-start active learning to facilitate reproduction of our reported results.

4.5.1 Multimodal Self-supervised Learning

We perform multimodal self-supervised learning with a local batch size of 64 on each GPU card and a global batch size of 256 across four GPU cards. The unimodal representations are vectors with 512 dimensions. The multimodal representations are concatenated unimodal representations. The learning schedulers are cosine schedulers with linear warm-up. Note that the unimodal prototypical loss is calculated only after the warm-up period. We use one linear layer as a projector to map the outputs from the feature backbones into the unified 512-dimensional representations.

For data augmentation and loading, we follow the settings in BMMAL[110]. For Food101, we apply image data augmentations such as random resize crop, random horizontal flip, grayscale, and color jittering during training. The cropped images

with a size of 224 are used as visual inputs. We use Bert’s tokenizer to tokenize the textual food recipes and use the tokens as textual inputs. For KineticsSound and VGGSound, we first uniformly sample 10 frames from each video clip. We then randomly sample 3 frames and apply random resize crop and random horizontal flip as video data augmentations. The cropped 3 video frames with a size of 168 are used as video inputs. We resample audio clips to 16 kHz and randomly select 5-second long audio clips for audio data augmentation. We use the spectrum of the audio as audio inputs. The number of Fourier transform points for each frame in the audio is set to 512, the window size is set to 512, and the hop length is set to 159.

For Food101, we train for 25 epochs and warm up for 5 epochs. The learning rates are set to 1×10^{-5} for the text backbone, 1×10^{-4} for the image backbone, and 1×10^{-3} for projectors. For KineticsSound and VGGSound, we train for 45 epochs and warm up for 15 epochs. The learning rates are set to 1×10^{-3} for the video backbone, the audio backbone and the projectors. We use the same AdamW optimizer with a weight decay of 0.02 and betas of 0.9 and 0.95 across the three datasets.

4.5.2 Supervised Cold-start AL

We perform the same data augmentation as in MMSSL. For Food101, we train for 15 epochs and warm up for 5 epochs. The learning rates are set to 1×10^{-5} for the text backbone, 1×10^{-4} for the image backbone, and 1×10^{-3} for the projectors. For video-audio datasets, KineticsSound and VGGSound, we train for 45 epochs and warm up for 15 epochs. The learning rates are set to 1×10^{-3} for the video backbone, the audio backbone and the projectors. We use AdamW optimizer with a weight decay of 0.01 and betas of 0.9 and 0.999 for supervised training.

4.5.3 Semi-supervised Cold-start AL

We reduce the number of training epochs and the learning rates to preserve the knowledge learned from the pretraining stage. For Food101, we train for 10 epochs and warm up for 5 epochs. The learning rates are set to 2×10^{-6} for the text backbone, 2×10^{-5} for the image backbone, and 1×10^{-3} for projectors. For

KineticsSound, we train for 20 epochs and warm up for 10 epochs. The learning rates are set to 5×10^{-4} for the video backbone and the audio backbone and 1×10^{-3} for projectors. For VGGSound, we train for 20 epochs and warm up for 10 epochs. The learning rates are set to 1×10^{-4} for the video backbone and the audio backbone and 1×10^{-3} for projectors. We use the same optimizer as in supervised cold-start active learning.

4.5.4 Optimize the Parametric Selection Model

Following the implementation of ActiveFT [56], we use Adam as the optimizer to optimize the parametric selection model with learning objective in **Equation 4.11**. To make optimization easier, parameters are initialized using the representations of uniformly selected samples rather than being randomly initialized. We set the learning rate to 1×10^{-3} and use a cosine scheduler to adjust the rate. The model is optimized over 300 epochs. After optimization, we select the samples that have the smallest unimodal distances to the parameters as our labeled subset.

4.6 Chapter Summary

In this chapter, we propose a two-stage method to address the cold-start problem in multimodal AL. The introduced uni-modal prototypes bridge the modality gap created by cross-modal contrastive learning. With our MMSSL, the distance estimation among multimodal data pairs becomes more precise, and it benefits data selection. Moreover, we propose to increase the modality alignment for multimodal data pairs to provide more useful modality shared information for downstream multimodal classification tasks. MMCSAL can save labeling budgets for multimodal learning and we will explore more multimodal tasks as future works.

Chapter 5

Hallucination Reduction in Large Vision-language Models via Emphasizing Image-negative Tokens

5.1 Introduction

In Chapter 3 and Chapter 4, we focus on the labeled data efficiency of multimodal classification task. The proposed methods achieve outstanding performance on data selection in both warm-start and cold-start setting. While these methods are helpful on addressing joint-modal learning tasks, they may not be applicable on cross-modal tasks such as visual question answering and multimodal text generation. In recent years, large vision-language models (LVLMs) have successfully integrated vision capabilities into language models, opening new avenues for AI applications [6, 7, 115]. To train a usable LVLM usually requires millions of high-quality multimodal data pairs which is expensive to collect. In this chapter, we aim to improve the data efficiency of LVLM training by optimizing the usage of available training data to deliver a better version of LVLM.

One of the most challenging issues in deploying LVLMs in real-world scenarios is the problem of hallucination. Hallucination occurs when the models generate plausible but inaccurate content that is not grounded in the actual image or not

related to the provided text prompt [116]. This is distinct from standard object detection errors, which involve inaccuracies in object localization or classification. Hallucination in LVLMs often manifests as generating descriptions of objects or events that do not exist in the image or failing to adhere to the input text prompt. This issue of hallucination can be particularly harmful in downstream tasks such as medical diagnostics [117] and robotics [118], where the accurate extraction of image details is required. Therefore, addressing the hallucination problem is crucial for applying LVLMs in various practical applications.

As discussed in Chapter 2, several approaches have been proposed [74, 76, 78, 83] to mitigate hallucination in LVLMs. However, these approaches often require additional instructional data, which is expensive to collect, or incur extra computational costs during the inference stage. In this chapter, we aim to reduce multimodal hallucination in LVLMs during the training phase. The most recent training stage method, EOS [76], mitigates hallucination by training the model to stop generating redundant responses, which are often hallucinatory. However, this method tends to produce overly short answers, which may not be desirable.

In this chapter, we propose a method that does not require additional instructional data or increase computational costs during inference, while still maintaining appropriate response lengths. We begin by investigating the origins of hallucinations in LVLMs, focusing on the influence of images on the probability distribution of tokens in generated responses. We group the response tokens into three distinct categories based on how their probabilities change when exposed to a noisy image [78]. **Figure 5.1** illustrates the changes in token probabilities when a clean image input is replaced by a noisy one. Tokens whose probabilities decrease after seeing a distorted image are termed "image-negative tokens," as they are less likely to be generated with a clean image. Conversely, tokens whose probabilities increase after seeing a noisy image are termed "image-positive tokens," as their likelihood of being generated is closely connected to the image content. The rest of the tokens are image-invariant since their probabilities do not change significantly with image quality. We then analyze hallucinations across these different token types and find that most hallucinations occur within image-invariant and image-positive tokens. Further analysis using saliency scores [119] reveals that hallucinations in image-positive tokens likely stem from the image itself, possibly due to poor image

representations, while hallucinations in image-invariant tokens primarily originate from preceding text tokens, indicating a language bias.

We explore how different token types affect hallucination during the training stage. We propose scaling the loss of tokens according to their visual dependence. Unlike traditional supervised fine-tuning of LVLMs, which treats all tokens equally, our approach adjusts the loss by suppressing or amplifying certain tokens throughout training. We find that emphasizing image-positive tokens during training results in both increased hallucinations and richer responses with more visual details. In contrast, emphasizing image-negative tokens while suppressing image-positive and image-invariant tokens helps reduce hallucinations. Additionally, we implement a simple data filtering strategy to remove instructional samples based on their visual dependence. We find that removing a certain portion of instructions with high visual dependence effectively mitigates hallucinations.

By combining these strategies, our method not only reduces hallucinations but also ensures data-efficient training. The loss re-weighting approach enhances learning without requiring additional data, while the data filtering strategy reduces reliance on expensive instructional samples. Experiments conducted on three LVLMs demonstrate the effectiveness of our methods in reducing hallucinations while maintaining appropriate response lengths and minimizing training costs.

In summary, our approach provides a data-efficient solution to reduce hallucination in LVLM by optimizing the usage of training data. This chapter demonstrates how strategic adjustments during training can improve model reliability and efficiency, aligning with the broader theme of data-efficient multimodal learning.

5.2 Analysis of Hallucination in LVLMs

Given an input image v , the generative LVLM θ produces a response in an autoregressive manner. The probability of generating each token y_t at time step t is denoted as $p(y_t|y_{<t}, v)$, where $y_{<t}$ represents the preceding tokens, and p is the probability of the current token y_t conditioned on previous tokens $y_{<t}$ and visual input v . The images, as the conditional variants, are significant to hallucinations in LVLMs. In this section, we investigate how the image input affects the generation probability distribution and its relationship to hallucination.

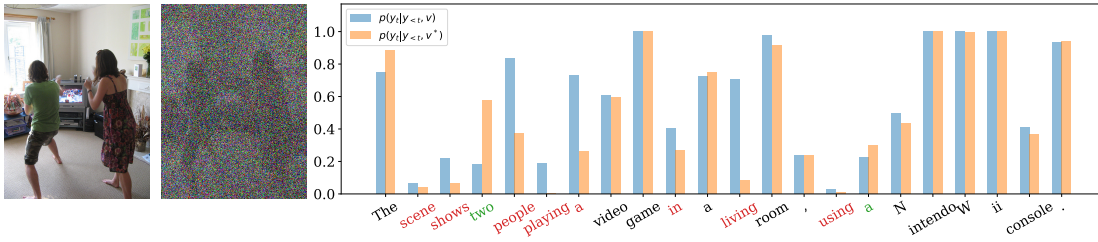


FIGURE 5.1: The left figure displays the input image and the middle figure is the noisy image with diffusion noise at time step 900. The model (fine-tuned LLaVA-v1.5-7b) is asked to generate a descriptive response given the image. The right figure shows the probability of each token y_t when provided with a clean image v and a noisy image v^* . The **image-positive**, **image-invariant** and **image-negative** tokens are marked as red, black and green.

5.2.1 Visual Dependence

To quantitatively assess visual dependence, we apply Visual Contrastive Decoding (VCD) [78], which generates a contrasting token output distribution by providing the model with a noisy image. The noisy image is created by applying Gaussian noise to the original image, similarly to the forward diffusion process [120]. The strength of the applied Gaussian noise is controlled by a time step T , which ranges from 0 to 1000, where $T = 0$ indicates clean image without any noise, and $T = 1000$ indicates complete Gaussian noise. We set the time step to 900, which retains minimal visual information while still allowing the model to perceive some aspects of the image without being entirely blinded. The noisy image is denoted as v^* , and we represent the visual dependence of each token using the relative difference in token probability:

$$d(y_t) = \frac{p(y_t|y_{<t}, v) - p(y_t|y_{<t}, v^*)}{\max[p(y_t|y_{<t}, v), p(y_t|y_{<t}, v^*)]}. \quad (5.1)$$

The visual dependence value $d(y_t)$ for a token y_t ranges between -1 and 1 . When $d(y_t)$ is significantly greater than 0 (within the range $[0.25, 1]$), it indicates that the probability of generating this token decreases with the noisy image, suggesting that the token relies on the image content. Conversely, when $d(y_t)$ is significantly less than 0 (within the range $[-1.0, -0.25]$), it means that this token is more likely to be generated even without image information, indicating a negative relationship with the visual content. As stated earlier, we refer to these two types of tokens as image-positive and image-negative tokens, respectively. Additionally, if a token

has a $d(y_t)$ value that is close to 0 (within the range $[-0.25, 0.25]$), we classify it as an image-invariant token, as it is not significantly affected by the distorted image. For a clearer illustration, we provide a concrete example in **Figure 5.1** with noise step set at 900. We show more examples in **Figure B.6, B.7** at our appendix B.

5.2.2 Hallucinations in Generated Text Tokens

After quantifying the visual dependence of each token and classifying them as image-invariant, image-positive, or image-negative, we analyze the prevalence of hallucination within these different types of tokens. To achieve this, we adopt fine-tuned LLaVA-v1.5-7b¹ to generate detailed descriptions for 500 randomly selected images from the COCO dataset [121]. The prompts used is: "Please describe this image in detail.". We then extract both the grounded and hallucinated objects from the generated responses and count the occurrences within the different groups of tokens based on their visual dependence. The results are presented at **Figure 5.2**.

Findings. As shown in **Figure 5.2**, image-invariant tokens contain the majority of both grounded and hallucinated objects. This is counter-intuitive, as we would expect detected or hallucinated object words to be more image-positive. On the other hand, there are more objects in image-positive tokens than in image-negative tokens, which aligns with our expectations. In the example from **Figure 5.1**, the Nintendo Wii console is difficult or ambiguous to detect from the image. The first token, "N", and the last token, "console", are image-related, while the other tokens are image-invariant, as they often appear together as a combination. Based on these observations, we conclude the following: (1) Hallucinations are more frequent in image-invariant tokens and in the most image-positive tokens. (2) Image-positive and invariant tokens are more likely to be associated with extracting information from the image content, whereas image-negative tokens are more often non-object terms.

The first finding relates to the co-occurrences phenomenon discussed in [77], where certain object pairs are frequently generated together in specific scenes. This leads the model to generate hallucinated objects associated with commonly co-occurring objects, even without directly seeing the image. The second finding suggests that

¹<https://huggingface.co/liuhaotian/llava-v1.5-7b>

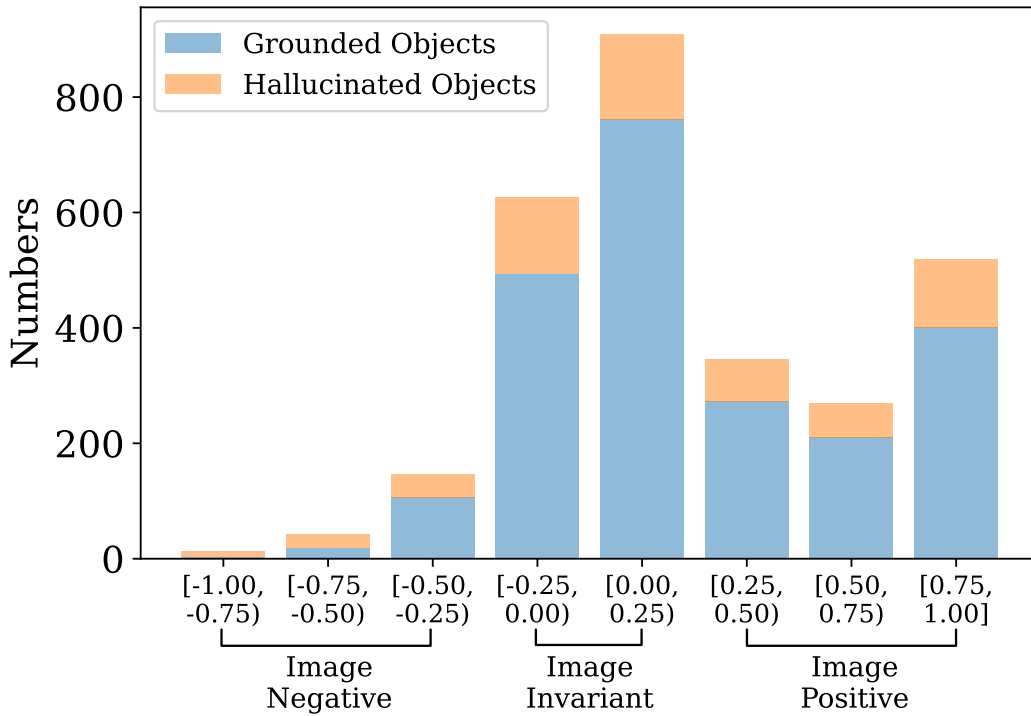


FIGURE 5.2: The number of grounded objects and hallucinated objects in different groups of tokens categorized by visual dependence. Tokens with lower visual dependence tend to be more image-negative, while tokens with higher visual dependence are more likely to be image-positive.

by emphasizing image-positive tokens during training, the model might be more inclined to focus on the actual image content. However, our experimental results in **Section 5.4** indicate that while emphasizing image-positive tokens does improve object recognition rates, it also leads to an increase in hallucination rates.

5.2.3 Information Flow across Tokens

Next, we aim to explore how different types of generated text tokens interact with image tokens and their preceding text tokens. For image-invariant tokens, we hypothesize that their hallucinations are primarily influenced by the language context rather than the image itself. For image-negative tokens, as illustrated in **Figure 5.1**, it is apparent that these tokens are more likely to be non-object words, such as prepositions, numerals and pronouns. To investigate these interactions, we utilize the saliency score [119], which measures the flow of information between any two tokens. The saliency score is an evaluation metric that captures the significance of this information flow. The saliency matrix I is defined as:

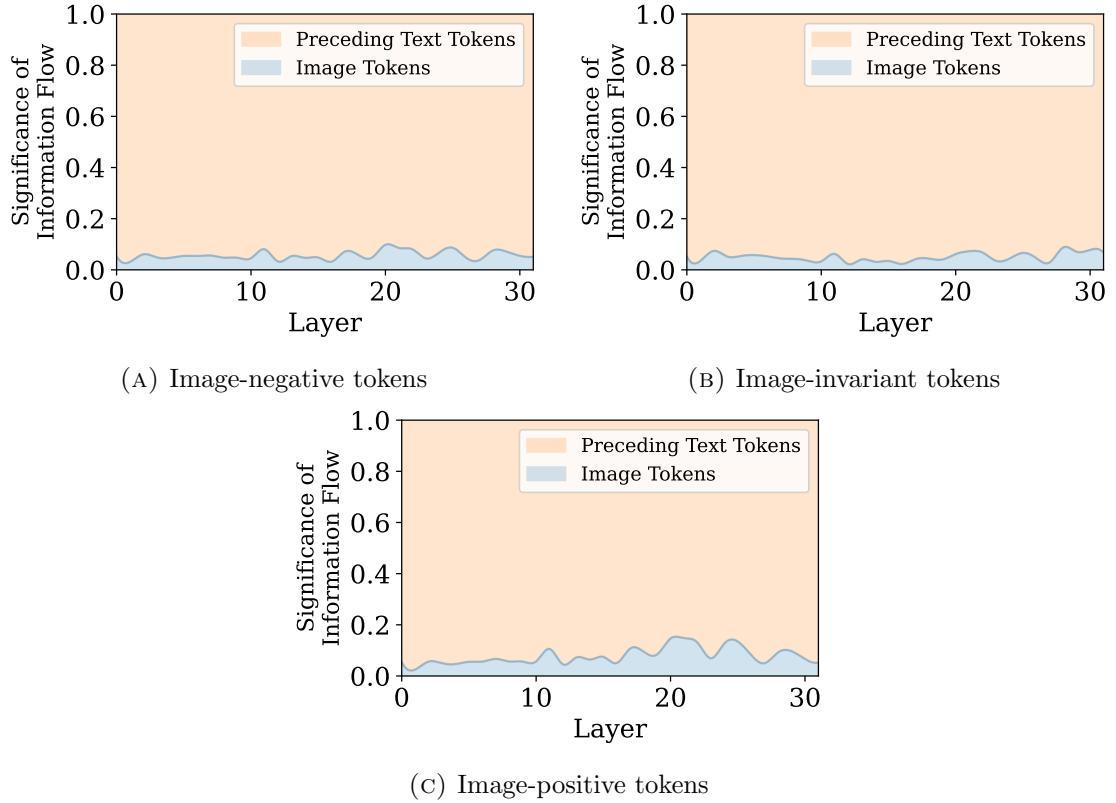


FIGURE 5.3: The significance of information flow per attention layer from image tokens and text tokens to different target tokens.

$$I = \left| A \odot \frac{\partial \mathcal{L}(x)}{\partial A} \right|, \quad (5.2)$$

where A is the attention matrix score of query and key, and \mathcal{L} is the cross-entropy loss of predicting the j -th token. I is a matrix with the same size as attention matrix where an element $I(i, j)$ represents the significance of the information flow from the j -th word to the i -th token. We adopt fine-tuned LLaVA-v1.5-7b model to evaluate 500 randomly selected image-text pairs from Detail23K dataset from LLaVA-instruction dataset [7]. We choose this dataset because it includes detailed image descriptions compared with COCO datasets used before. We select the top-10 most negative, positive and invariant text tokens from each image-text pair and compute their saliency scores. LLaVA-7b model has 32 attention layers and 32 attention heads. The saliency scores of each layer are averaged across all attention heads. We calculate the mean of saliency scores from image tokens and from their preceding tokens to the target tokens and show their proportions in **Figure 5.3**.

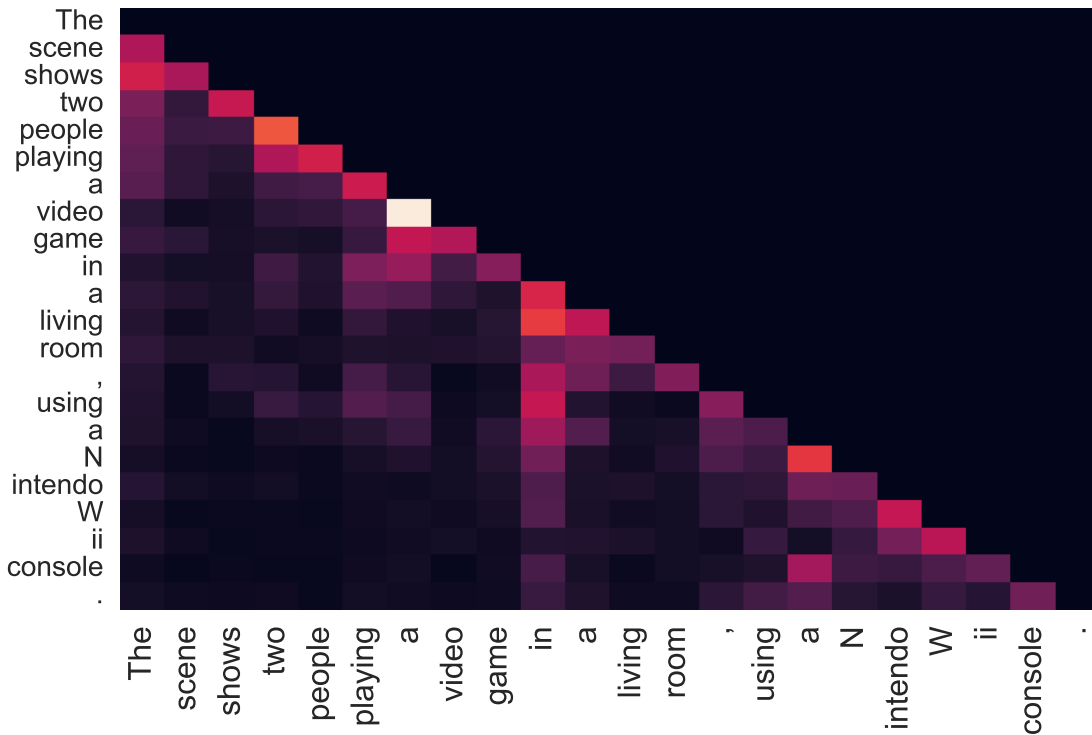


FIGURE 5.4: The saliency scores of each token in the attention matrix of LLaVA-v1.5-7b of example from Figure 1.

Findings. As shown in the figure, for all tokens, the primary source of information is their language context. The information flow from image tokens is greater for image-positive text tokens compared to image-invariant and image-negative tokens, indicating that image-invariant and image-negative tokens interact less with the image tokens. From these observations, we draw the following conclusion: Hallucinations in image-invariant tokens are more likely to stem from their preceding text tokens rather than from the image tokens. In other words, image-invariant tokens contribute to hallucinations differently than image-positive tokens. This also suggests that contrastive decoding methods, such as VCD [78], may not be effective for these image-invariant tokens, as they do not rely on image input and therefore cannot have their distributions adjusted by feeding a corrupted image to the model.

Additionally, we examine the saliency matrix exclusively among the generated text tokens, as shown in **Figure 5.4**. The figure reveals that non-object tokens interact more with other tokens than object tokens do. This relates to the patterns of summary tokens observed in OPERA [83], where these summary tokens, mostly non-object tokens, consolidate previously gathered knowledge and use it to guide

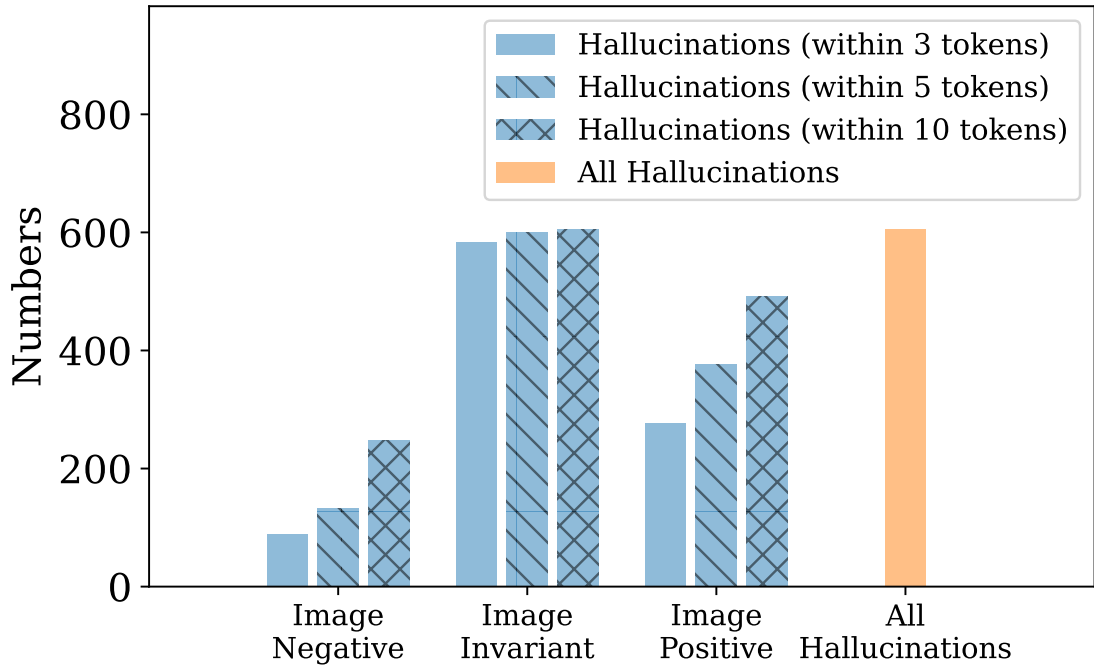


FIGURE 5.5: The hallucinated content are often found to be near the image-invariant tokens.

the generation of subsequent tokens. We then analyze the co-occurrence of hallucinations of different types of text tokens using 500 randomly selected images from the COCO dataset. The results are shown in the **Figure 5.5**. Over 95% of hallucinations occur within 3 tokens of an image-invariant token, and only approximately half of hallucinations appear within 3 tokens of an image-positive token. From these information, we can infer that the non-object image-invariant tokens also lead to hallucinations, and they exhibit behavior similar to the summary tokens described in OPERA.

5.3 Adjust Loss Weights During Training

Our analysis suggests that different caption tokens have varying impacts on hallucinations in LVLMs, where image-invariant tokens cause a significant influence on hallucinations. Therefore, we propose adjusting the training loss weights for these tokens based on their visual dependence, to explore how this adjustment influences hallucination in LVLMs. We design two re-weighting strategies: one that up-scales

the loss for image-negative tokens, and another that up-scales the loss for image-positive tokens. Each strategy targets different token types, allowing us to study their effects on hallucination. The re-weighting functions are defined as follows:

$$w_{\text{neg}}(y_t) = \begin{cases} -d(y_t) & \text{if } d(y_t) \leq 0, \\ 0 & \text{if } d(y_t) > 0, \end{cases} \quad (5.3)$$

$$w_{\text{pos}}(y_t) = \begin{cases} 0 & \text{if } d(y_t) \leq 0, \\ d(y_t) & \text{if } d(y_t) > 0, \end{cases} \quad (5.4)$$

where w_{neg} is applied when emphasizing image-negative tokens, while w_{pos} is used for emphasizing image-positive tokens. It is worth noting that, in both strategies, the weights of image-invariant tokens are down-scaled. To redistribute the loss weights without altering the overall magnitude of the original loss, we re-normalize the weights using softmax with a temperature parameter τ , controlling the sharpness of weight distributions, and derive the re-scaled loss objective:

$$\mathcal{L}(y) = -\frac{1}{T} \sum_{t=1}^T T \frac{e^{\tau w(y_t)}}{\sum_1^T e^{\tau w(y_t)}} \log p(y_t | y_{<t}, v). \quad (5.5)$$

With this loss re-scaling, for example, when using w_{neg} as $w(y_t)$, the loss weights become greater than 1 for image-negative tokens and less than 1 for image-positive tokens. In our experiments, we find that excessively down-scaling the loss weight of the EOS token could affect generation lengths. Therefore, we ensure that the loss weight of the EOS token is never reduced below 1 to avoid this issue. The computation of $d(y_t)$ is performed on-the-fly right before loss computation for every batch of data, and it is done in evaluation mode to save GPU memory. To allow the model to accurately assess visual dependence, we begin loss re-weighting after a certain portion of the training has been completed. By default, we start adjusting loss weights after half of the total training time. In our ablation studies, we evaluate the effects of different temperatures, starting times for adjustment, and various noise time steps used in calculation of visual dependence.

5.4 Experiments

In this section, we introduce the LVLMS, datasets, and benchmarks used in our experiments, along with the baseline methods. We then present our experimental results and provide analysis.

5.4.1 Experiment Setup

LVLMS. We select three LVLMS variants with different parameter sizes: PaliGemma-3b-224 [115], LLaVA-v1.5-7b and 13b [7] and Bunny-v1.1-8B [122], to evaluate our methods. The language model of PaliGemma-3b-224 is Gemma-2b [123] and its vision model is from SigLIP [124]. The language models of LLaVA-v1.5-7b and 13b are Vicuna-v1.5-7b and 13b [125] and their vision model is from CLIP [9]. The language model of Bunny-v1.1-8b model is LLaMa-3-8b-Instruct [126] and its vision model is from SigLIP [124]. By evaluating our method across these three variants of LVLMS, we can validate the effectiveness and the generality of our method.

Datasets. We use LLaVA-Detail 23k, LLaVA-Instruct 150k [7] for fine-tuning LLaVA models. Since PaliGemma-3b is a single-turn model, we use LLaVA-Detail 23k, which is also a single-turn dataset, to fine-tune it. For Bunny-v1.1-8b, we use the subset from its instruction fine-tune dataset SVIT-Detail 71k [127] to conduct training. The LLaVA-Detail 23k and SVIT-Detail 71k contain detailed image descriptions about images. The LLaVA-Instruct 150k contains the LLaVA-Detail 23k and more conversational and reasoning data samples.

Benchmarks. We evaluate models using two hallucination metrics: Caption Hallucination Assessment with Image Relevance (CHAIR) [116] and FaithScore [128]. CHAIR quantifies object hallucination by performing rule-based object matching in the generate responses. It identifies hallucinations by detecting objects not included in the ground-truth label set from COCO annotations. We report both the sentence-level score $CHAIR_S$, which indicates the percentage of captions that contain hallucinations, and the instance-level score $CHAIR_I$, which measures the frequency of hallucinated objects among all mentioned objects. Additionally, we report the Recall score to evaluate object recognition rates, representing the model’s

ability to generate richer responses. FaithScore evaluates the consistency of atomic facts extracted from the generated captions with the image content. This is done using a pipeline involving LLaMa 2 [129] in stage1, ChatGPT [130] in stage 2 and OFA [131] in stage 3 as the final evaluation steps. We report both the instance-level and sentence-level FaithScore.

5.4.2 Baselines

We select several baselines that effectively reduce hallucinations in LVLMS. **Over-trust penalty and retrospection-allocation (OPERA)** [83]: This inference-stage method identifies summary tokens that contribute to hallucinations and penalizes them. It further removes these tokens and re-samples the remaining ones to improve response accuracy. **Visual Contrastive Decoding (VCD)**: This decoding strategy mitigates the model’s language bias by contrasting the output distribution with an additional distribution generated from noisy images. **Selective EOS Supervision for Training (EOS)** [76]: This training-phase method excludes EOS (End of Sequence) token probabilities from the loss calculation for non-EOS tokens, allowing the model to terminate generation before producing redundant, often hallucinatory, content. We perform greedy sampling except for OPERA which requires beam search.

5.4.3 Main Results

We conduct experiments by fine-tuning the pre-trained LVLMS with LoRA [132] for 2 epochs with batch size of 128. The inference-based methods are applied to models fine-tuned using a standard maximum likelihood estimation loss function. For our method, we fix the temperature τ at 0.5 and apply the loss re-weighting only after 1 epoch to allow the model to warm-up and assess visual dependence more accurately. The noise time step is fixed to 900. **Table 5.1** presents the response length and the performance on CHAIR and FaithScore.

Emphasizing image-positive tokens increases both object recall and hallucination. As discussed in **Section 5.2.2**, image-positive tokens contribute to

| Model | Dataset | Method | Length | CHAIR _S ↓ | CHAIR _I ↓ | Recall ↑ | Faith ↑ | Faith _S ↑ |
|----------------|------------------------|------------------------------|-------------|----------------------|----------------------|-------------|-------------|----------------------|
| LLaVA-v1.5-7b | LLaVA-Details 23k | MLE | 102.1 | 54.2 | 15.9 | <u>78.3</u> | 86.1 | <u>71.0</u> |
| | | <i>EOS</i> | <i>62.6</i> | <i>30.2</i> | <i>10.9</i> | <i>69.2</i> | <i>89.9</i> | <i>74.7</i> |
| | | VCD-t=500 | 101.3 | 56.6 | 17.4 | 77.7 | 83.7 | 67.1 |
| | | OPERA | 96.5 | <u>52.6</u> | <u>15.7</u> | 75.0 | <u>85.9</u> | <u>71.0</u> |
| | | OPERA _{fast} | 102.8 | 56.2 | 17.4 | 75.4 | 84.3 | 68.6 |
| | | Ours w_{pos} | 102.5 | 56.8 | 16.5 | 78.6 | 85.5 | 70.2 |
| | | Ours w_{neg} | 101.8 | 51.2 | 15.5 | <u>78.4</u> | 85.8 | 72.3 |
| LLaVA-v1.5-13b | LLaVA-Details 23k | MLE | 105.2 | 54.6 | <u>15.3</u> | <u>77.3</u> | 86.6 | <u>71.5</u> |
| | | <i>EOS</i> | <i>74.9</i> | <i>39.4</i> | <i>13.1</i> | <i>71.9</i> | <i>88.8</i> | <i>74.5</i> |
| | | VCD _{t=200} | 104.4 | 54.8 | 16.2 | 75.7 | 84.9 | 69.5 |
| | | OPERA | 97.5 | 54.0 | 16.8 | 76.4 | 85.3 | 70.0 |
| | | OPERA _{fast} | 98.8 | 54.8 | 16.9 | 74.8 | 84.8 | 68.8 |
| | | Ours w_{pos} | 105.1 | <u>52.4</u> | 15.7 | 77.8 | 86.0 | 71.8 |
| | | Ours w_{neg} | 105.4 | 51.0 | 15.2 | 76.6 | <u>86.4</u> | 71.8 |
| LLaVA-v1.5-7b | LLaVA-Instruct 150k | MLE | 105.3 | 57.0 | 16.9 | <u>78.7</u> | 85.8 | <u>72.4</u> |
| | | <i>EOS</i> | <i>56.9</i> | <i>24.4</i> | <i>9.1</i> | <i>66.9</i> | <i>90.9</i> | <i>76.2</i> |
| | | VCD _{t=200} | 105.2 | 54.6 | 16.9 | 78.0 | 85.2 | 71.2 |
| | | OPERA | 97.2 | <u>51.2</u> | 14.9 | 76.4 | 86.1 | 72.3 |
| | | OPERA _{fast} | 102.3 | 54.6 | 15.6 | 77.3 | <u>86.2</u> | 72.6 |
| | | Ours w_{pos} | 99.6 | 52.2 | <u>14.8</u> | 79.3 | 86.0 | 71.8 |
| | | Ours w_{neg} | 99.8 | 48.8 | 13.7 | 77.6 | 86.4 | 71.8 |
| Pali-Gemma-3b | LLaVA-Details 23k | MLE | 100.7 | 51.0 | 15.3 | 76.3 | 88.4 | <u>74.0</u> |
| | | <i>EOS</i> | <i>42.2</i> | <i>15.8</i> | <i>6.5</i> | <i>56.1</i> | <i>93.6</i> | <i>80.4</i> |
| | | VCD _{t=500} | 98.9 | 52.8 | 16.1 | 77.3 | 87.8 | 71.9 |
| | | OPERA | 102.3 | 48.6 | <u>13.5</u> | <u>77.9</u> | 87.5 | 73.0 |
| | | OPERA _{fast} | 101.5 | <u>48.4</u> | 13.3 | 76.7 | 88.0 | 74.4 |
| | | Ours w_{pos} | 100.5 | 52.8 | 15.6 | 79.1 | 87.9 | 73.1 |
| | | Ours w_{neg} | 101.0 | 47.8 | 14.6 | 74.4 | <u>88.2</u> | 73.7 |
| Bunny-v1.1-8b | SVIT-Details 71k | MLE | 350.1 | 44.4 | 8.7 | 74.4 | 87.8 | 75.2 |
| | | EOS | 311.9 | 40.8 | <u>8.2</u> | <u>74.5</u> | 88.2 | 76.0 |
| | | VCD _{t=200} | 343.6 | 47.8 | 9.1 | 74.2 | 87.3 | 72.7 |
| | | OPERA | 334.1 | 40.2 | 7.7 | 74.0 | 89.2 | 75.1 |
| | | OPERA _{fast} | 347.6 | 42.4 | 8.8 | 73.9 | <u>88.8</u> | 74.2 |
| | | Ours w_{pos} | 343.8 | 44.0 | 8.5 | 75.9 | 88.6 | <u>75.4</u> |
| | | Ours w_{neg} | 353.4 | <u>40.4</u> | 8.5 | 73.5 | <u>88.8</u> | 75.1 |

TABLE 5.1: Main results of three variants of LVLMS: LLaVA-v1.5 (7b and 13b), PaliGemma-3b and Bunny-v1.1-8b on CHAIR and FaithScore (Faith). *The rows in italics* are excluded from the comparison because their object recognition recall is significantly lower than the others, making the comparison unfair. The best results are marked as **bold** and the second best are underlined.

both object detection and hallucination. Therefore, up-scaling the loss for image-positive tokens enables the model to generate richer responses, but it also increases hallucination. In scenarios where hallucination is more acceptable, and users require responses with richer visual information [82], this fine-tuning approach might be a good choice.

Emphasizing image-negative tokens reduces hallucination. This improvement in mitigating hallucination can be attributed to two factors. First, image-negative tokens are typically non-object words and are often located farther from where hallucinations tend to occur, as shown in **Figure 5.5**. Second, by simultaneously downscaling the loss weights for image-invariant and image-positive tokens, which are more likely to contain hallucinations, we effectively reduce the model’s focus on noisy content. As a result, by directing the model’s attention away from these tokens, we help prevent it from learning and generating hallucinated information.

VCD performs poorly and EOS generates short responses. We test VCD algorithm using three noise time steps $T = 200, 500, 900$ and report the best one. We find that it performs not as good as expected. One possible reason is that VCD performs not well on image-invariant tokens as they are not affected by the image conditions, as discussed in **Section 5.2.3**. We also test EOS and find it is highly effective at reducing hallucinations, but it comes at the cost of shorter response lengths and lower object recall scores. Naturally, shorter responses contain fewer details and content, making them less likely to produce hallucinations. As a result, we exclude EOS from comparisons with other methods when its recall score is significantly lower.

In Appendix B, we present qualitative results comparing image descriptions generated by LVLMs trained with vanilla loss and those trained with our re-weighted loss, which emphasizes image-negative tokens. Hallucinated text content is manually marked in red. These results demonstrate that LVLMs trained with our re-weighted loss produce fewer instances of hallucinated text compared to those trained with vanilla loss.

| Method | Length | CHAIR _S ↓ | CHAIR _I ↓ | Recall ↑ |
|-----------------------------------|-------------|----------------------|----------------------|-------------|
| LLaVA-v1.5-7b + LLaVA-Detail 23k | | | | |
| - | 101.4 | 51.4 | 14.0 | 78.7 |
| MLE | 100.4 | 48.8 | 13.8 | 77.9 |
| EOS | 82.9 | 41.0 | 12.1 | 75.0 |
| VCD _{t=500} | 101.5 | 52.8 | 14.8 | <u>78.6</u> |
| OPERA | 93.1 | 42.8 | 13.1 | 76.9 |
| OPERA _{fast} | 97.2 | 48.2 | 13.7 | 75.7 |
| Ours w_{neg} | 97.9 | <u>42.4</u> | <u>12.6</u> | 76.5 |
| LLaVA-v1.5-13b + LLaVA-Detail 23k | | | | |
| - | 102.2 | 48.4 | 13.4 | 77.4 |
| MLE | 101.0 | 48.2 | 14.2 | 77.8 |
| EOS | 85.7 | 41.8 | 12.9 | 75.1 |
| VCD _{t=200} | 102.0 | 51.8 | 14.3 | <u>77.5</u> |
| OPERA | 102.4 | 43.3 | <u>12.7</u> | 76.8 |
| OPERA _{fast} | 98.7 | 47.5 | 13.6 | 76.9 |
| Ours w_{neg} | 99.1 | <u>42.4</u> | 12.2 | 76.5 |
| PaliGemma-3b + LLaVA-Detail 23k | | | | |
| - | <i>51.0</i> | <i>18.0</i> | <i>7.0</i> | <i>58.1</i> |
| MLE | 66.1 | 26.6 | 11.9 | 69.1 |
| <i>EOS</i> | <i>33.4</i> | <i>13.2</i> | <i>11.1</i> | <i>57.8</i> |
| VCD _{t=900} | 65.6 | 29.6 | 10.6 | 68.6 |
| OPERA | 76.2 | <u>26.2</u> | 8.8 | 67.8 |
| OPERA _{fast} | 82.7 | 24.2 | <u>8.7</u> | <u>69.0</u> |
| Ours w_{neg} | 71.3 | <u>26.2</u> | 7.9 | 68.6 |
| Bunny-v1.1-8b + SVIT-Detail 71k | | | | |
| - | 246.5 | 40.6 | <u>7.6</u> | 74.3 |
| MLE | 300.8 | 40.8 | 7.5 | 75.6 |
| EOS | 272.5 | 42.4 | 8.5 | 74.8 |
| VCD _{t=200} | 300.5 | 44.2 | 8.9 | 75.2 |
| OPERA | 293.4 | <u>37.8</u> | 7.7 | 74.5 |
| OPERA _{fast} | 302.8 | 39.4 | 7.5 | <u>75.4</u> |
| Ours w_{neg} | 302.7 | 37.2 | 7.8 | 73.8 |

TABLE 5.2: The experimental results of continue fine-tuning of instructional tuned models of three variants of LVLMS: LLaVA-v1.5 (7b and 13b), PaliGemma-3b and Bunny-v1.1-8b on CHAIR benchmark.

5.4.4 Further Training Fine-tuned LVLMS

We conduct another experiment where we further train the already fine-tuned LVLMS with LoRA for an additional epoch with batch size of 128. Given the short training duration, we use a larger temperature of $\tau = 1$. However, for PaliGemma, which has not been previously trained on the LLaVA-Detail 23k dataset, we set the temperature to 0.5 to avoid overly aggressive loss adjustment. Re-weighting of the loss begins after half an epoch of training. The performance results are shown in **Table 5.2**. In the table, the method labeled with '-' represents the original instructional fine-tuned base model. As shown in the table, our method, which suppresses image-invariant and image-positive tokens during training, achieves an overall improvement in reducing hallucinations. However, as with most existing hallucination reduction methods, the recall score decreases, as the model is less encouraged to identify more objects to minimize the risk of potential hallucinations.

5.4.5 Filtering Data by Visual Dependence

We design a data filtering strategy by ranking data according to their total visual dependence, calculated as $\sum_{i=1}^T d(y_t)$. The experiment is conducted using LLaVA-Instruct 150k dataset, with visual dependence evaluated by the fine-tuned LLaVA-v1.5-7b model and the noise step is set at 900. After filtering the dataset, we fine-tune pretrained LLaVA-v1.5-7b from scratch using filtered dataset for 2 epochs with LoRA, employing the vanilla loss function. **Table 5.3** presents the results. Notably, our filtering strategy does not affect the overall training data length or the response length. On the one hand, removing the top 10% of data with the highest visual dependence significantly reduces hallucinations. However, this strategy also lowers the recall score, as image-positive data, which contain both valuable visual information and noisy, hallucinated information, are removed. On the other hand, removing data with the lowest visual dependence also helps control hallucinations without compromising recall scores, since these data contain fewer visual details compared to those with high visual dependence.

| Strategy | Data Len. | Response Len. | CHAIR _S | CHAIR _I | Recall |
|-------------|-----------|---------------|--------------------|--------------------|-------------|
| No Filter | 172.8 | 105.3 | 57.0 | 16.9 | 78.7 |
| Lowest 10% | 172.9 | 99.6 | 49.2 | 14.2 | 78.3 |
| Lowest 20% | 172.4 | 100.2 | <u>46.8</u> | <u>13.3</u> | 78.4 |
| Highest 10% | 176.1 | 100.8 | 44.4 | 12.9 | 76.9 |
| Highest 20% | 177.8 | 102.5 | 48.4 | 14.5 | 75.2 |
| Random 10% | 172.7 | 101.4 | 52.8 | 15.3 | <u>78.6</u> |
| Random 20% | 172.8 | 99.3 | 50.4 | 14.7 | 78.3 |

TABLE 5.3: The experimental results of data filtering. Data Len. represents the average data sentence length of filtered data, and Response Len. represents the average response length.

5.4.6 The Distribution of Visual Dependence

We remove data by ordering them by their visual dependence value evaluated by an instructional fine-tuned model. **Figure 5.6** shows the metrics distribution of LLaVA-Instruct 150k dataset. The metrics is the sum of visual dependence of all tokens from each instruction $\sum_{t=1}^T d(y_t)$ evaluated by instructional fine-tuned LLaVA-v1.5-7b. A higher value indicates that the model’s prediction probability will decrease, in other words, the model is less confident with its original response, after seeing a noisy image. As illustrated in the figure, most of instructions have positive values of visual dependence. By filtering out data with lowest value or with highest value, the hallucination performance of the model is affected in different ways.

5.4.7 Ablation Studies of Hyperparameters

In addition to the main results of fine-tuning the pre-trained large vision-language models (LVLMs), we investigate how various hyperparameters influence hallucination performance, including temperature τ in **Equation 5.5**, the starting time for loss adjustment, and the noise time steps used for calculating visual dependence. The results for each hyperparameter are detailed in **Table 5.4, 5.5, 5.6**, respectively. These experiments incorporate loss re-weighting, emphasizing image-negative tokens to reduce hallucinations. The default settings are τ of 0.5, adjustment starting after 50% of the total training time, and a noise step T of 900.

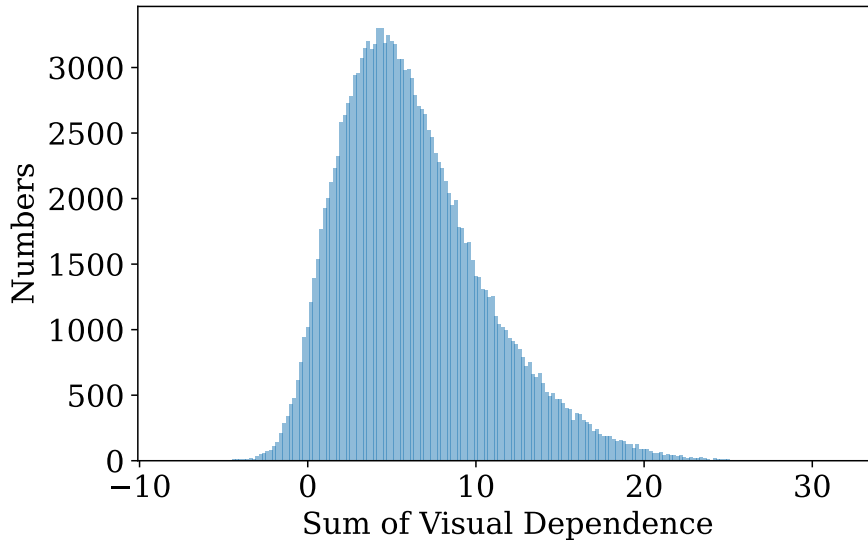


FIGURE 5.6: The distribution of the sum of visual dependence of each sample from LLaVA-Instruct 150k.

Temperature τ . When the temperature τ is set to 0, it becomes the vanilla loss without any re-weighting. As τ becomes larger, the distribution of loss weights becomes sharper, meaning that the weights are more concentrated to emphasize only few image-negative tokens. As shown in **Table 5.4**, the recall score drops with higher temperatures because the image-invariant and image-positive tokens are more heavily suppressed.

Start Time for Weight Adjustment. Since our loss adjustment approach requires assessing the visual dependence of each token, applying this adjustment too early can lead to inaccurate assessments and negatively impact performance. As shown in **Table 5.5**, starting the loss adjustment after the model has completed half of its total training time yields better results than starting at the beginning.

Noise Step T . Different noise steps T introduce varying levels of visual uncertainty when computing visual dependence, thereby affecting the calculation. A lower noise step generally produces visual dependence values closer to 0, whereas a higher noise step results in values with a larger magnitude. As **Table 5.6** shows, a T value at 200 provides the best recall performance, and the LLaVA model is less sensitive to T compared to the PaliGemma model.

| Temp τ | Length | CHAIR _S ↓ | CHAIR _I ↓ | Recall ↑ |
|-----------------------------------|--------|----------------------|----------------------|-------------|
| LLaVA-v1.5-7b + LLaVA-Details 23k | | | | |
| 0.25 | 101.9 | 52.2 | 15.9 | 77.9 |
| 0.50 | 101.8 | 51.2 | 15.5 | 78.4 |
| 1.00 | 100.7 | 52.0 | 15.8 | 77.0 |
| 1.50 | 101.1 | 51.8 | 16.1 | 75.9 |
| PaliGemma-3b + LLaVA-Details 23k | | | | |
| 0.25 | 100.8 | 49.8 | 15.1 | 76.4 |
| 0.50 | 101.0 | 47.8 | 14.6 | 74.4 |
| 1.00 | 100.4 | 46.8 | 14.1 | 74.5 |
| 1.50 | 100.4 | 44.4 | 13.8 | 74.0 |

TABLE 5.4: The performance on CHAIR with different temperature τ in fine-tuning pre-trained LLaVA-v1.5-7b and PaliGemma-3b with LLaVA-Details 23k dataset.

| Start Time | Len | CHAIR _S | CHAIR _I | Recall |
|-----------------------------------|-------|--------------------|--------------------|-------------|
| LLaVA-v1.5-7b + LLaVA-Details 23k | | | | |
| 0% | 102.1 | 51.4 | 15.4 | 77.7 |
| 25% | 102.3 | 52.2 | 15.3 | 77.6 |
| 50% | 101.8 | 51.2 | 15.5 | 78.4 |
| 75% | 101.5 | 53.4 | 16.1 | 77.6 |
| PaliGemma-3b + LLaVA-Details 23k | | | | |
| 0% | 100.7 | 48.4 | 14.9 | 76.0 |
| 25% | 101.3 | 50.8 | 15.0 | 75.6 |
| 50% | 101.0 | 47.8 | 14.6 | 74.4 |
| 75% | 101.0 | 47.6 | 14.8 | 75.4 |

TABLE 5.5: The performance on CHAIR with different start times of loss re-weighting in fine-tuning pre-trained LLaVA-v1.5-7b and PaliGemma-3b with LLaVA-Details 23k dataset. For example, 0% means that starting loss re-weighting from the beginning.

Based on these results, for a better trade-off between hallucination and recall score, we choose a temperature τ of 0.5, adjustment starting after 50% of the total training time, and a noise step T of 900 as our default settings. However, these parameters need to be tuned when the experiment settings are changed.

| Noise Step T | Len | CHAIR _S ↓ | CHAIR _I ↓ | Recall ↑ |
|-----------------------------------|-------|----------------------|----------------------|-------------|
| LLaVA-v1.5-7b + LLaVA-Details 23k | | | | |
| 200 | 101.9 | 51.2 | 15.1 | 78.9 |
| 500 | 102.0 | 50.6 | 15.2 | 78.4 |
| 900 | 101.8 | 51.2 | 15.5 | 78.4 |
| PaliGemma-3b + LLaVA-Details 23k | | | | |
| 200 | 100.7 | 50.8 | 15.1 | 77.0 |
| 500 | 100.8 | 50.6 | 15.3 | 76.1 |
| 900 | 101.0 | 47.8 | 14.6 | 74.4 |

TABLE 5.6: The performance on CHAIR with different noise step T in fine-tuning pre-trained LLaVA-v1.5-7b and PaliGemma-3b with LLaVA-Details 23k dataset.

5.4.8 Performance on MME and POPE

In addition to CHAIR and FaithScore benchmarks for image captioning, we report performance on two yes/no visual question answering benchmarks: MME [133] and POPE [134]. We evaluate the fine-tuned LLaVA-v1.5-7b model based on pre-trained checkpoints with LLaVA-Instruct 150k dataset. Moreover, we select a subset from LLaVA-Mix 665k dataset [7] consist of AOKVQA (66k), RefCOCO (48k), LLaVA-Conversation (58k), VQAv2 (83k) and OKVQA (9k), to save computational cost. This subset has 264k samples in total. We fine-tune LLaVA-v1.5-7b with this dataset on top of pre-trained checkpoints and evaluate its performance. The results are shown in **Table 5.7**. Overall, our method is better when fine-tuned with LLaVA-Instruct 150k. However, when fine-tuned with LLaVA-Mix subset 264k, our method performs close to other baselines. One possible explanation is that the responses in LLaVA-Instruct 150k to answer a yes/no question are usually in a multi-turn manner, which are longer than those in LLaVA-Mix subset 264k, which are usually few tokens long, therefore loss re-weighting is not working effectively on these overly short sequences.

5.4.9 Experiment Settings and Computational Cost

The experiments are conducted with 4 NVIDIA A100-40GB GPU cards. Due to heavy computational cost of training LVLMS, all experiments are single-run with random seed at 42. For reproducibility, we provide the detailed experiment settings

| Dataset | Method | MME _P ↑ | MME _C ↑ | POPE _{Acc} ↑ | POPE _{F1} ↑ |
|------------------------------|-----------------------|--------------------|--------------------|-----------------------|----------------------|
| LLaVA- Instruct 150k | MLE | 756.9 | 261.8 | 59.9 | 71.2 |
| | EOS | 840.8 | 269.6 | 62.8 | 72.6 |
| | Ours w_{neg} | 849.9 | 271.4 | 64.2 | 73.4 |
| LLaVA- Mix Subset 264k | MLE | 1377.4 | 268.2 | 85.4 | 84.9 |
| | EOS | 1369.1 | 267.9 | 85.9 | 85.3 |
| | Ours w_{neg} | 1386.6 | 282.1 | 85.9 | 85.3 |

TABLE 5.7: The performance of LLaVA-v1.5-7b on MME and POPE of vanilla MLE loss, selective EOS supervision and our approach of emphasizing image-negative tokens. MME_P and MME_C represent the perception and cognition scores. POPE_{Acc} and POPE_{F1} are the average accuracy and f1score across three POPE subsets: random, popular and adversarial.

| Hyper-parameters | LLaVA-v1.5-7b | LLaVA-v1.5-13b | PaliGemma-3b | Bunny-v1.1-8b |
|-----------------------------|---------------|----------------|--------------|---------------|
| Random seed | 42 | 42 | 42 | 42 |
| LoRA rank | 128 | 128 | 64 | 128 |
| LoRA alpha | 256 | 256 | 128 | 256 |
| LoRA dropout | 0.05 | 0.05 | 0.05 | 0.05 |
| Epochs | 2 | 2 | 2 | 2 |
| Warmup ratio | 0.03 | 0.03 | 0.03 | 0.03 |
| Per device train batch size | 16 | 16 | 16 | 8 |
| Gradient accumulation step | 2 | 2 | 2 | 2 |
| GPU devices | 4 | 4 | 4 | 4 |
| Global batch size | 128 | 128 | 128 | 64 |
| Deepspeed stage | zero2 | zero3 | zero2 | zero3 |
| Precision | BFloat16 | BFloat16 | BFloat16 | BFloat16 |
| Optimizer | AdamW | AdamW | AdamW | AdamW |
| Weight decay | 0 | 0 | 0 | 0 |
| Base learning rate | 2e-4 | 2e-4 | 1e-4 | 2e-4 |
| MM projector learning rate | 2e-5 | 2e-5 | 1e-5 | 2e-5 |

TABLE 5.8: The hyper-parameters for experiment of fine-tuning LVLMs from scratch.

for fine-tuning the pre-trained LVLMs in **Table 5.8**, including LoRA parameters, batch size, optimizer and etc. We use greedy decoding strategy for our method.

Our method introduces around 20% extra training cost as we conduct evaluation of visual dependence during training. We present the computational cost of using the vanilla loss and our re-weighted loss in **Table 5.9**. Note that the visual dependence and loss weights are computed in evaluation mode, thus our method does not consume extra GPU memory.

| Experiment | Model | Dataset | Epochs | Vanilla (Hours) | Ours (Hours) |
|--------------------|----------------|---------------------|--------|-----------------|--------------|
| From Scratch | LLaVA-v1.5-7b | LLaVA-Details-23k | 2 | 0.9 | 1.1 |
| | LLaVA-v1.5-7b | LLaVA-Instruct-150k | 2 | 9.7 | 11.2 |
| | LLaVA-v1.5-13b | LLaVA-Details-23k | 2 | 1.7 | 2.0 |
| | PaliGemma-3b | LLaVA-Details-23k | 2 | 0.5 | 0.6 |
| | Bunny-v1.1-8b | SVIT-Details-71k | 2 | 6.8 | 8.0 |
| Continue Fine-tune | LLaVA-v1.5-7b | LLaVA-Details-23k | 1 | 0.5 | 0.6 |
| | LLaVA-v1.5-13b | LLaVA-Details-23k | 1 | 0.8 | 1.0 |
| | PaliGemma-3b | LLaVA-Details-23k | 1 | 0.2 | 0.3 |
| | Bunny-v1.1-8b | SVIT-Details-71k | 1 | 3.5 | 4.0 |

TABLE 5.9: Computational cost for each experiment.

5.5 Chapter Summary

In this chapter, we investigate the relationship of hallucination between visual dependence of generated tokens. Our findings suggest that hallucinations mostly occur around image-invariant tokens and these tokens are mostly affected by the language context during generation. Moreover, image-positive tokens contribute to both hallucinations and extraction of visual information. Based on these findings, we develop a loss re-weighting approach to optimize data usage of training dataset and achieve a lower hallucination rate. Moreover, we design a data filtering strategy to mitigate hallucinations by removing data samples which are with more potentially hallucinated content.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

Deep multimodal learning has driven advancements in various fields, including healthcare, robotics, and multimedia chatbots. However, one of the biggest challenges in this area is the extensive amount of data required to train deep multimodal networks effectively. Additionally, efficiently utilizing this training data to produce well-trained models is a complex task. Current data-efficient algorithms are often not optimized for multimodal learning, particularly in terms of selecting balanced multimodal data, enhancing the quality of multimodal representations, and reducing hallucinations.

In this thesis, we have developed several data-efficient multimodal learning algorithms to address these challenges. In Chapter 3, we introduced a balanced multimodal active learning algorithm that facilitates fair data selection between weak and dominant modalities. In Chapter 4, we tackled the cold-start active learning problem by developing a novel two-stage cold-start multimodal active learning algorithm designed to enhance the representation quality and alignment of selected data. In Chapter 5, we proposed a visual dependence-aware algorithm that adjusts training weights for each token during training to reduce hallucinations in large vision-language models.

The first contribution is introducing a new method for multimodal active learning that selects data while considering modality balance. We found that conventional

active learning methods often fail to select balanced multimodal data pairs, leading to a bias toward the dominant modality. Our method estimates modality contributions using Shapley value estimation and applies gradient embedding modulation based on these contributions, enabling fairer data selection and allowing for the training of multimodal networks with less annotated data.

The second contribution is proposing a two-stage method to address the cold-start problem in multimodal active learning. We observed that contrastive-based multimodal self-supervised learning can create a modality gap that affects distance measurements among multimodal data pairs. To address this, we introduced unimodal prototypes to reduce the modality gap during self-supervised learning and added a regularization term to maximize alignment during data selection. Our method demonstrated superior performance under cold-start conditions in both supervised and semi-supervised learning scenarios.

The third contribution is developing a method to reduce the hallucination rate in large vision-language models by analyzing how tokens with different visual dependencies contribute to hallucinations. We proposed adjusting the training weights for different tokens so that the model focuses more on tokens less likely to cause hallucinations. Additionally, we suggested filtering out training data prone to hallucinations, allowing the model to be trained on cleaner data. Our approach achieved lower hallucination rates in both fine-tuning from scratch and continued fine-tuning settings.

In summary, our contributions enhance data efficiency in various multimodal applications, including multimodal classification, multimodal active learning, and multimodal large language models. The balanced multimodal active learning method helps select fairer data pairs, and this algorithm can be extended to improve multimodal learning performance by filtering out unbalanced data as a preprocessing step. The cold-start active learning approach is particularly useful when building multimodal datasets from scratch without any available annotations. Finally, our hallucination reduction method can be applied to train various multimodal agents, significantly mitigating their hallucination rates.

6.2 Future Work

In this section, we discuss future directions for data efficient multimodal learning.

Data Curation for Large-Scale Multimodal Self-Supervised Learning.

Training effective multimodal encoders typically requires millions of data pairs, often sourced from the web, which poses challenges in terms of data collection and computational costs. Current contrastive learning methods rely on large datasets for model convergence. However, by implementing careful data curation techniques and generating more positive and negative pairs from the same dataset, it may be possible to reduce data requirements without compromising model performance.

Data Generation for Cost-Effective Dataset Enrichment. Another promising direction is using data generation techniques to enrich training datasets at a lower cost. As image and text generation models continue to advance, their outputs are becoming increasingly indistinguishable from human-generated data. This machine-generated data offers a more controlled and cost-effective alternative to traditional data collection methods, such as web crawling or manual annotation. However, a significant challenge is the lack of diversity in generated data, which can negatively impact the model’s ability to generalize to out-of-distribution scenarios.

Proxy Models for Efficient Data Filtering and Selection. In the context of large vision-language models, data selection can be prohibitively expensive due to the immense size of these models, which often have tens of billions of parameters. Data filtering, however, is a simpler task compared to data generation. Therefore, developing a smaller, more efficient proxy model to handle data filtering or selection presents a cost-effective trade-off. This approach could significantly enhance data efficiency when training larger models.

Appendix A

Appendix of Chapter 3

In this appendix, we supplement the computational cost analysis of our proposed method in Chapter 3, how we address the issue of storing large gradient embedding in large-scale active learning experiments in Section 3.3.5 and the implementation details of NL-gate in Section 3.3.6 .

Computational Complexity. Computing the Shapley values of each unimodal feature requires to perform inference 2^M times in total, where M is the number of modalities. In our two-modality learning case, we need to perform inference four times with different combination of unimodal features to obtain the Shapley values, which is acceptable. Then, given the computed gradient embedding of N unlabeled samples, the sampling time complexity of BMMAL is $\mathcal{O}(NBDK)$, where B is the query budget of each AL round, D is the size of weight matrix of the last linear classifier and K is the number of classes.

Split the Large Unlabeled Data Pool. In large-scale AL experiments, the gradient embedding produced by all unlabeled data samples could be too large to be stored in the memory. To address this issue, we split the unlabeled data pool into S smaller pools to save memory space, where S is the split size. After splitting, we query $\frac{N}{S}$ unlabeled samples from each smaller pool and aggregate them to form the final query set. The space complexity of BMMAL is correspondingly reduced by S times. Moreover, the sampling time complexity becomes $\mathcal{O}(\frac{N}{S}\frac{B}{S}DKS) = \mathcal{O}(\frac{1}{S}NBDK)$, which is also reduced by S times compared with

original time complexity. We use split size of eight in the large-scale AL experiment with the VGGSound-full dataset. Although splitting might affect the AL performance, we observe that both BMMAL and BADGE still perform better than random data selection. It indicates that splitting the unlabeled data pool is acceptable in large-scale AL.

Implementation of NL-gate. NL-gate [103] is a mid-fusion mechanism that behaves similar to multi-head attention. We implement it in the video classification task, where Resnet-18 is used as the audio backbone and Resnet2P1D-18 is used as the video backbone. Note that both Resnet-18 and Resnet2P1D-18 have four blocks. We extract the middle 2D audio features from the third block of Resnet-18 and the middle 3D video features from the third block of Resnet2P1D-18 as inputs to the NL-gate.

We show the implementation of NL-gate in **Figure A.1**. The 3D video feature is average pooled over the spatial channels into a 1D video feature. It is then tiled over the frequency channel into a 2D video feature that has the same size as the 2D audio feature. The concatenation of the 2D video feature and the 2D audio feature is used as key and value in NL-gate. The original 3D video feature is used as query in NL-gate. After audio and video features are mixed, they will be processed with a random initialized module with the same layout as the fourth Block of Resnet2P1D-18 to produce the final feature. To compute the marginal unimodal contribution, we choose to compute the Shapley values of the features generated by the last shared convolution layers ($Conv_V$ and $Conv_A$) before the NL-gate fusion module.

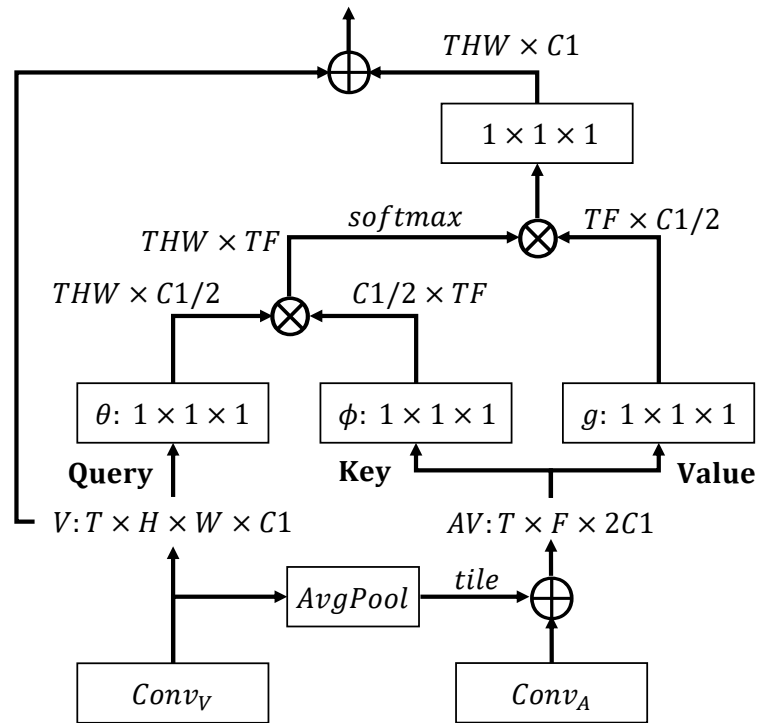


FIGURE A.1: The implementation of NL-gate. We use the 3D video feature as query and the 2D concatenated audio and video feature as key and value.

Appendix B

Appendix of Chapter 5

In this appendix, we demonstrate qualitative results of experiments from Chapter 5 of generated image descriptions using the fine-tuned model from scratch with vanilla loss and our re-weighted loss by emphasizing image-negative tokens. **Figures B.1, B.2, B.3, B.4, B.5** show the results. Moreover, we show more examples of probability distribution of predicted tokens with clean image and with noisy image of fine-tuned LLaVA-v1.5-7b, which is trained from scratch with LoRA on LLaVA-Details 23k dataset. The results are shown in **Figure B.6, B.7**.

**Vanilla Loss:**

The image features a beautiful scene of cherry blossoms in full bloom, creating a picturesque atmosphere. A traffic light is situated in the middle of the scene, surrounded by the blossoming trees. The traffic light is red, indicating that the cars should stop and wait.

There are several cars in the image, with one car on the left side of the scene, another car on the right side, and two more cars further back in the scene. The cars are patiently waiting for the traffic light to change, allowing them to continue their journey through the lovely cherry blossom-filled area.

Ours w_{neg} :

The image features a beautiful scene of a tree filled with white flowers, likely cherry blossoms, in full bloom. The tree is located near a traffic light, which is currently displaying a red light. The traffic light is positioned at the top left corner of the image. The combination of the flowering tree and the red traffic light creates a striking contrast and a picturesque view.

Vanilla Loss:

The image features a snow-covered slope with a man and a young girl skiing. The man is standing on skis in the foreground, while the young girl is skiing behind him. They are both wearing skiing gear and enjoying their time on the snowy hill.

There are several other people in the background, some of whom are also wearing skis. The scene captures the excitement and fun of skiing as a group of people engage in the winter sport.

Ours w_{neg} :

The image features a snowy hill with a man and a child standing on it. The man is wearing a black jacket and is positioned towards the left side of the hill, while the child is standing closer to the right side. Both of them are wearing skis, ready to enjoy the snowy terrain.

There are also two sets of skis visible in the scene, one set belonging to the man and the other set belonging to the child. The skis are placed on the snow, indicating that they are preparing to ski down the hill.

**Vanilla Loss:**

The image features a person holding a large sandwich in their hands, with a bite taken out of it. The sandwich is placed on a dining table, and there is a cup nearby, possibly containing a soft drink. Another person is visible in the background, partially obscured by the main subject.

In the scene, there are two additional sandwiches on the table, one located near the left edge and the other closer to the right side. A fork can also be seen on the table, likely used to eat the sandwiches.

Ours w_{neg} :

The image features a person holding a large, juicy hamburger in their hand, ready to take a bite. The hamburger is placed on a bun and appears to be quite messy. The person is wearing a ring on their finger.

In the background, there is a dining table with a Pepsi cup on it, suggesting that the person is enjoying a meal in a casual setting. Another person can be seen in the background, partially visible on the right side of the image.

Vanilla Loss:

The image features a round piece of dough, likely a pizza dough, sitting on a piece of parchment paper on a table. The dough is uncooked and has a circular shape, **with a hole in the center**. The dough is placed on a piece of wax paper, which is covering the table.

In the background, there is a bowl, possibly containing ingredients for the dough or other food items. The scene suggests that the dough is being prepared for baking or further processing.

Ours w_{neg} :

The image features a round piece of dough, likely a pizza crust, sitting on a table. The dough is in the process of being shaped and is still in the middle of being rolled out. It is placed on a piece of parchment paper, which is spread across the table. The dough appears to be in the process of being stretched and shaped, possibly for making a pizza or another type of dough-based dish.

**Vanilla Loss:**

The image features three women dressed in costumes, standing in front of a fireplace. They are posing for a picture, with one of the women on the left side, another in the middle, and the third on the right side of the frame. The women are wearing hats, adding to their festive appearance.

A dog is also present in the scene, standing near the leftmost woman. **A handbag can be seen placed on the floor, close to the center of the image.** The room appears to be decorated for Christmas, with a Christmas tree visible in the background.

Ours w_{neg} :

The image features three women dressed in costumes, standing in front of a Christmas tree. They are posing for a picture, with one of the women standing closer to the left side of the tree, another in the middle, and the third on the right side. The women are wearing hats, adding to their festive appearance.

A dog is also present in the scene, located near the left side of the image. The dog appears to be well-behaved and enjoying the company of the women. The overall atmosphere of the image is cheerful and celebratory, with the women and the dog all dressed up for the holiday season.

FIGURE B.1: Qualitative results of LLaVA-v1.5-7b trained with vanilla loss baseline and ours approach with w_{neg} on LLaVA-Instruct 150k dataset. Hallucinations are marked as red.

**Vanilla Loss:**

The image features a large, colorful teddy bear sitting in a display case, surrounded by various sweets. The teddy bear is positioned in the center of the display, drawing attention to itself. There are multiple candy items scattered throughout the scene, including a few donuts and a sports ball. A person can be seen in the background, possibly admiring the teddy bear or browsing the candy display. The scene appears to be set in a store or a candy shop, where the teddy bear serves as an eye-catching centerpiece for the sweets.

Ours w_{neg} :

The image features a large, colorful teddy bear sitting in a display case, surrounded by various candy items. The teddy bear is positioned in the center of the display, drawing attention to itself. The display case is filled with different types of candy, including multiple donuts and a few cakes. There are two people in the scene, one on the left side and another on the right side. They appear to be observing the teddy bear and the candy display, possibly admiring the arrangement or considering making a purchase.

Vanilla Loss:

The image features a large orange and white cat sitting on a window sill, looking out of the window. The cat is positioned near the center of the scene, with its body stretched out and its head peeking out of the window. In the background, there are two cars visible, one on the left side and another on the right side of the image. A person can also be seen in the background, standing near the right side of the scene. The scene appears to be a busy street with the cat observing the surroundings from the comfort of its window.

Ours w_{neg} :

The image features a cat sitting on a window sill, looking out of the window with curiosity. The cat is positioned near the center of the scene, and the window is located on the left side of the image. In the background, there is a car parked outside, partially visible on the right side of the image. The car appears to be a small vehicle, and it is likely that the cat is observing the car or something else happening outside.

**Vanilla Loss:**

The image features a street scene with a traffic light prominently visible in the foreground. The traffic light is situated on a pole, and its red light is illuminated. In the background, there is a street sign that reads "Santa Monica Blvd."

There are several cars parked along the street, with one car closer to the left side of the image and two others further down the road. Additionally, there is a person standing near the middle of the scene, possibly waiting to cross the street or walking along the sidewalk.

Ours w_{neg} :

The image features a street scene with a traffic light hanging above the road. The traffic light is red, indicating that vehicles should stop. The street sign above the traffic light reads "Santa Monica Blvd." There are several houses in the background, with one house prominently visible on the left side of the image. The scene also includes a few cars parked or driving along the street, with one car closer to the left side of the image and two other cars further down the road.

Vanilla Loss:

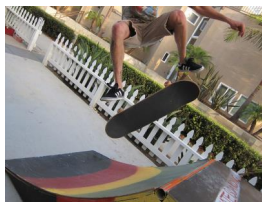
The image features a group of zebras in a sandy enclosure, likely a zoo or wildlife park. There are five zebras in total, with some standing closer to the foreground and others further back. They are grazing and enjoying the sunshine in their enclosure.

In addition to the zebras, there are two birds visible in the scene. One bird is located near the top left corner of the image, while the other is positioned towards the right side. The birds seem to be observing the zebras as they graze.

Ours w_{neg} :

The image features a group of zebras standing in a dirt field, likely in a zoo enclosure. There are five zebras in total, with some of them standing closer to the fence and others further away. The zebras are positioned in various directions, creating a dynamic scene.

The field is surrounded by a fence, which separates the zebras from the visitors. The zebras appear to be grazing or simply enjoying their time in the enclosure.

**Vanilla Loss:**

The image features a man skillfully performing a trick on his skateboard, jumping over a rail with the skateboard under his feet. He is in mid-air, showcasing his talent and athleticism.

In the background, there are two potted plants visible, one on the left side and another on the right side of the scene. Additionally, there is a person standing further back, possibly watching the skateboarder's performance or waiting for their turn to skate.

Ours w_{neg} :

The image captures a man skillfully performing a trick on his skateboard, jumping over a ramp. He is in mid-air, showcasing his talent and control over the skateboard. The skateboard is positioned beneath him, indicating the successful execution of the trick.

In the background, there are two potted plants, one on the left side and the other on the right side of the scene. These plants add a touch of greenery to the environment, creating a pleasant atmosphere for the skateboarder's performance.

FIGURE B.2: Qualitative results of LLaVA-v1.5-7b trained with vanilla loss baseline and ours approach with w_{neg} on LLaVA-Instruct 150k dataset. Hallucinations are marked as red.

**Vanilla Loss:**

The image features a city street with a snow-covered sidewalk and several cars parked along the side. There are at least five cars visible in the scene, with some parked closer to the foreground and others further back. The street is lined with tall buildings, giving the scene a classic urban atmosphere.

In addition to the cars, there is a person standing near the middle of the scene, possibly observing the parked cars or waiting for someone. The overall scene captures the essence of a typical city street during winter.

Ours w_{neg} :

The image features a city street with a large Coca-Cola sign prominently displayed on one of the buildings. The street is lined with several cars parked along the side, including a few parked closer to the sign. The cars are of various sizes and are positioned at different distances from the sign.

The street is also lined with several buildings, some of which have a snow-covered appearance. The combination of the snow-covered buildings and the Coca-Cola sign create a unique and interesting scene in the city.

Vanilla Loss:

The image captures a soccer game in progress, with several players on the field. One player is running with the ball, trying to get ahead of the other players. The ball is located in the center of the field, and the players are actively engaged in the game.

There are multiple people visible in the scene, including the players and possibly some spectators. Some players are closer to the foreground, while others are further back in the field. A bench can be seen in the background, likely for resting or waiting for their turn to play.

Ours w_{neg} :

The image captures a group of men playing soccer on a field. There are several players on the field, with one player in the foreground running after the soccer ball. Another player is close to the ball, while a third player is farther away from the action.

In total, there are eight people visible in the scene, with some of them closer to the foreground and others further back. The soccer ball is located in the middle of the field, and the players are actively engaged in the game.

**Vanilla Loss:**

The image depicts a group of people sitting around a long dining table in a restaurant. There are at least 12 people in the scene, with some sitting closer to the table and others further away. The table is covered with various items, including multiple cups, bottles, and bowls.

There are also a few books scattered around the table, possibly indicating that the group is engaged in a discussion or reading material. The atmosphere appears to be social and relaxed, as the people are enjoying their time together.

Ours w_{neg} :

The image depicts a group of people sitting around a long dining table, enjoying a meal together. There are at least 12 people in the scene, with some sitting closer to the table and others further away. The table is covered with various items, including multiple cups, bottles, and bowls.

The dining area is well-equipped with chairs and a couch, providing ample seating for the guests. The table is adorned with a checkered tablecloth, adding a touch of elegance to the setting. The people are engaged in conversation and enjoying their time together, creating a warm and inviting atmosphere.

Vanilla Loss:

The image captures a skateboarder performing a trick on a ramp, with the skateboard visible in mid-air. The skateboarder is in the process of jumping over the ramp, showcasing their skill and talent.

In the background, there are several potted plants, with one located near the left side of the image and another on the right side of the image. Additionally, there are two benches in the scene, one on the left side and another on the right side of the image. A person can be seen standing near the left side of the image, possibly observing the skateboarder's performance or waiting for their turn to ride the ramp.

Ours w_{neg} :

The image captures a skateboarder performing a trick on a ramp, showcasing their skills. The skateboarder is in mid-air, with their skateboard visible underneath them. The ramp is located in a park, and the skateboarder is positioned towards the left side of the ramp.

In the background, there are several potted plants scattered around, adding a touch of greenery to the scene. Additionally, there are two benches in the park, one located near the left side of the ramp and the other further to the right.

**Vanilla Loss:**

The image features a wooden dining table with a black tray filled with a variety of food items. The tray is divided into four sections, each containing different types of food. There are several pieces of meat, such as meatloaf, and a variety of vegetables, including broccoli and carrots.

In addition to the food, there are two bowls placed on the table, one on the left side and the other on the right side. A fork can be seen on the left side of the table, ready to be used for enjoying the meal. The table setting appears to be set for a meal with a diverse assortment of food options.

Ours w_{neg} :

The image features a wooden dining table with a black tray filled with a variety of food items. The tray is divided into four sections, each containing different types of food. There are several pieces of broccoli spread across the tray, with some placed near the top, middle, and bottom sections.

In addition to the broccoli, there are also pieces of meat and rice, making the tray a well-rounded meal. A bowl is also present on the table, located near the top right corner. A fork can be seen on the table, positioned near the top left corner.

FIGURE B.3: Qualitative results of PaliGemma-3b trained with vanilla loss baseline and ours approach with w_{neg} on LLaVA-Details 23k dataset. Hallucinations are marked as red.

**Vanilla Loss:**

The image captures a vibrant scene of a field, where a white pickup truck is parked. The truck is quite large, taking up a significant portion of the image. It has a white color and is detailed with a black stripe. The truck's features are clearly visible, including its windows, tires, and doors. The windows are black, and **the truck has two doors**. The tires are black and silver, and one of them is prominently displayed in the front. **The truck also has a black handle on the door.**

In the background, there is a mountain range that adds depth to the image. The mountains are not the main focus but provide a scenic backdrop. The sky above is clear and blue, with no clouds.

What makes this image particularly lively is the presence of kites in the sky. There are several kites, each with different colors and designs. One of the kites is red and black, while another is yellow and black. There is also a green and yellow kite, and a black and yellow kite. These kites are flying high in the sky, adding a sense of movement and color to the scene.

The field where the truck is parked is covered in dry, brown grass. The grass is not lush and seems to be in a state of dryness. The field extends to the horizon, and the truck is parked on it.

Overall, the image portrays a day out in a field with a white pickup truck as the central object, surrounded by the natural beauty of the mountains in the background and the colorful kites flying in the clear blue sky.

Ours w_{pos} :

The image captures a vibrant scene of a white pickup truck parked in a field of dry, brown grass. The truck is quite large, taking up a significant portion of the image, and it's detailed with various features. The truck's tires are black and silver, with one of them having a silver hubcap. The wheels are also black, and the truck's windows are white. **The truck's door is open, and there's a person sitting in the driver's seat. The truck also has a red and white sticker on it.**

In the sky above the truck, there are several kites flying. One of the kites is red and black, while another is yellow and black. There's also a green and yellow kite, as well as a blue and yellow one. The kites are scattered across the sky, with some being closer to the top of the image and others being lower.

The field in which the truck is parked is vast and stretches across the bottom of the image. The grass is dry and brown, **indicating that it might be a hot day or the field hasn't seen rain recently.**

In the background, there are hills and mountains visible. **The hills are green, suggesting that they are covered in vegetation, while the mountains are brown, possibly indicating that they are rocky or barren. There's also a white building in the distance, which adds some architectural interest to the scene.**

The sky is clear and blue, which provides a beautiful backdrop for the kites flying. The combination of the white truck, the colorful kites, and the natural landscape creates a dynamic and lively image.

Ours w_{neg} :

The image captures a vibrant scene of a white pickup truck parked in a field of dry, brown grass. **The truck is a Toyota**, as indicated by the logo on the truck door. **It's a four-door model, with a black door handle and a black side mirror.** The truck's tires are black, and the hubcaps are silver. The truck's windows are rolled up, and **there's a person sitting in the driver's seat, wearing a black shirt.**

The field in which the truck is parked is vast and stretches out to the horizon. **There are patches of green grass scattered throughout the field, and there are also some bushes in the distance.** The sky above is clear and blue, with no clouds.

What makes this scene particularly lively is the presence of kites flying in the sky. There are several kites, including a red kite, a green kite, and a yellow kite. **One of the kites is shaped like a dragon, and another one is shaped like a lizard.** The kites are flying high in the sky, adding a sense of movement and color to the image.

In the background, there are hills and mountains, which add depth to the image. The mountains are far in the distance, and they appear to be quite tall.

Overall, the image portrays a serene and picturesque setting with a white Toyota pickup truck parked in a field, surrounded by dry grass and flying kites in the clear blue sky. The presence of the truck and the person inside it suggests that this might be a recreational or leisure activity, possibly a day out in the countryside.

FIGURE B.4: Qualitative results of Bunny-v1.1-8b trained with vanilla loss baseline and ours approach with w_{pos} and w_{neg} on SVIT-Details 71k dataset. Hallucinations are marked as red, the correct details that are not discovered in baseline are marked green.



Vanilla Loss:

The image depicts a cozy indoor setting, likely a living room or a study, with a prominent wooden table that takes up most of the space. The table is adorned with a variety of objects, including a stack of books, a glass vase, a wine glass, and a candle holder.

The stack of books is quite noticeable, with three books neatly arranged on top of each other. The books are of different colors, including blue, yellow, and black, and they are all hardcover. The titles of the books are not visible, but one of them is identified as "The Centurys Man" and another as "These Days". The books are placed on the table, and there is a book with a yellow cover on top of the stack.

Beside the stack of books, there is a glass vase filled with water. The vase is clear and has a green plant with red flowers inside it. The flowers are vibrant and add a touch of color to the scene. The vase is placed on the table, and the plant is visible through the clear glass.

There is also a wine glass on the table, which is empty. The wine glass is placed near the vase and is made of clear glass.

A candle holder is also present on the table, and it contains a white candle. The candle is placed inside the holder, which is likely made of glass or another material.

In addition to the objects on the table, there is a chair visible in the image. The chair is made of wood and has a brown color. It is positioned near the table, possibly for someone to sit and read the books or enjoy the ambiance created by the candle and the flowers.

The background of the image is a white wall, which contrasts with the brown color of the table and the chair. The wall is clean and provides a simple backdrop for the objects on the table.

Overall, the image conveys a sense of tranquility and intellectual pursuit, with the combination of books, flowers, and candlelight creating a serene atmosphere.

Ours w_{pos} :

The image depicts a cozy indoor setting, likely a living room or a study. The room is dominated by a large wooden table, which spans almost the entire width of the image. The table is adorned with various objects, including a stack of books, a glass vase, and a wine glass.

The stack of books is prominently placed on the table. There are four books in total, each with a distinct cover. The top book is blue, while the second book from the top is yellow. The third book is red, and the bottom book is black. The titles of the books are visible, with one of them being "The Gentility Man". The books are neatly stacked, with the blue book on top and the black book at the bottom.

Beside the stack of books, there is a clear glass vase filled with water. The vase is placed on the table and contains a bouquet of red tulips. The tulips are vibrant and fresh, with green stems and leaves. The vase is transparent, allowing the water inside to be visible.

In addition to the vase, there is a wine glass on the table. The glass is clear and contains a white candle.

The candle is lit, casting a warm glow. The wine glass is placed near the vase, and both are positioned on the table.

Behind the table, there is a wooden chair. The chair is brown and has a backrest. It is positioned in such a way that it is partially visible behind the table.

The background of the image is a white wall, which provides a neutral backdrop to the scene. The wall extends from the left side of the image to the right, and it is clean and uncluttered.

Overall, the image conveys a sense of tranquility and intellectual pursuits, with the stack of books suggesting a love for reading, and the vase of tulips and lit candle adding a touch of elegance and warmth to the setting.

Ours w_{neg} :

The image depicts a cozy indoor scene, likely a reading corner or a study room. The main focus is a wooden table, which is dark brown in color. The table is quite large, occupying a significant portion of the image. On the table, there are three stacks of books, each with a different color cover. The books are neatly arranged, with the top book being blue, the middle one being yellow, and the bottom one being red. The titles of the books are not visible, but the colors of their covers are distinct.

Beside the stacks of books, there is a clear glass vase filled with water. The vase is placed on the table and contains a bouquet of red tulips. The tulips are vibrant and fresh, with their green stems visible through the clear glass of the vase. The vase is positioned near the books, adding a touch of nature to the scene.

In addition to the books and vase, there is a wine glass on the table. The wine glass is empty and has a slender stem. It is placed near the books, possibly indicating that someone might be enjoying a glass of wine while reading.

On the table, there is also a piece of paper with a red and blue design. The paper is placed near the books and adds a splash of color to the scene.

The background of the image is a white wall, which provides a clean and simple backdrop to the table and its contents. The wall is plain and does not have any decorations or patterns.

Overall, the image conveys a sense of tranquility and intellectual pursuit, with the stacks of books, the vase of tulips, and the wine glass suggesting a space for reading, relaxation, and contemplation.

FIGURE B.5: Qualitative results of Bunny-v1.1-8b trained with vanilla loss and ours approach with w_{pos} and w_{neg} on SVIT-Details 71k dataset. Hallucinations are marked as red, the correct details that are not discovered in baseline are marked green.

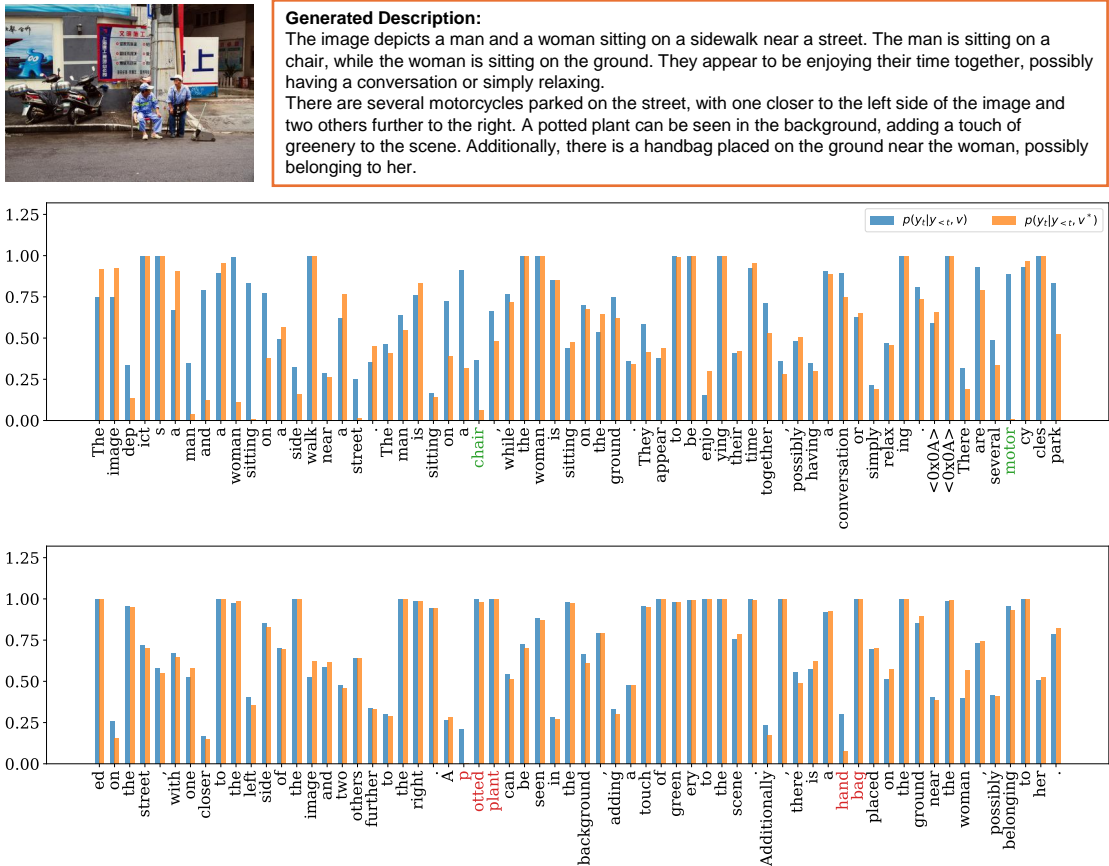


FIGURE B.6: The top left sub-figure is the input image, and the top right sub-figure is the generated image description from LLaVA-v1.5-7b fine-tuned from scratch with LLaVA-Details 23k dataset. The bottom bar charts display the probability of each token with clean image input v and with noisy image input v^* . The hallucinated objects are marked red, the grounded objects are marked green.

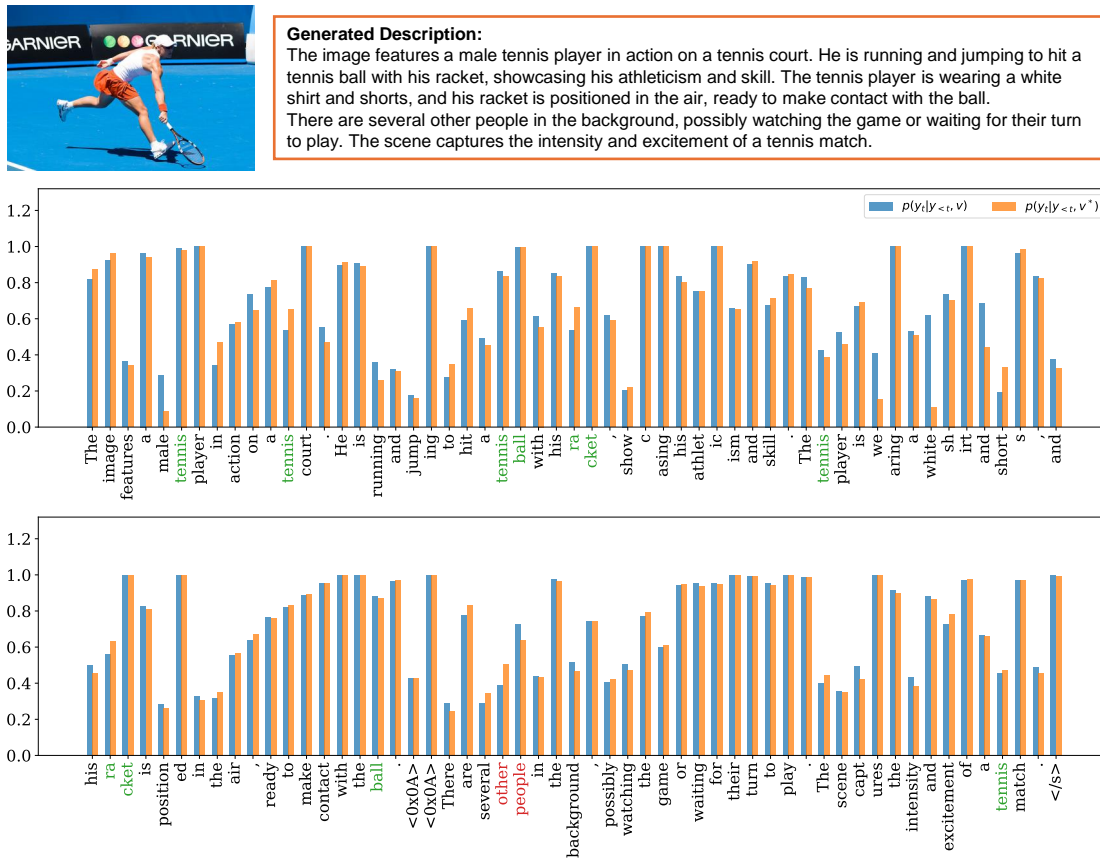


FIGURE B.7: The top left sub-figure is the input image, and the top right sub-figure is the generated image description from LLaVA-v1.5-7b fine-tuned from scratch with LLaVA-Details 23k dataset. The bottom bar charts display the probability of each token with clean image input v and with noisy image input v^* . The hallucinated objects are marked red, the grounded objects are marked green.

List of Publications

Conference Proceedings

- **Meng Shen**, Yizheng Huang, Jianxiong Yin, Heqing Zou, Deepu Rajan, and Simon See, “Towards Balanced Active Learning for Multimodal Classification”. In Proceedings of the 31st ACM International Conference on Multimedia (MM '23).
- **Meng Shen**, Yake Wei, Jianxiong Yin, Deepu Rajan, Di Hu, and Simon See, “Enhancing Modality Representation and Alignment for Multimodal Cold-start Active Learning”. In Proceedings of the 6th ACM International Conference on Multimedia in Asia (MM Asia '24).

Bibliography

- [1] S. M. S. A. Abdullah, S. Y. A. Ameen, M. A. Sadeeq, and S. Zeebaree, “Multimodal emotion recognition using deep learning,” *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 73–79, 2021. 1
- [2] E. Cambria, D. Das, S. Bandyopadhyay, and A. Feraco, “Affective computing and sentiment analysis,” *A practical guide to sentiment analysis*, pp. 1–10, 2017. 1
- [3] S.-C. Huang, L. Shen, M. P. Lungren, and S. Yeung, “Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3942–3951, 2021. 1
- [4] M. Moor, Q. Huang, S. Wu, M. Yasunaga, Y. Dalmia, J. Leskovec, C. Zakka, E. P. Reis, and P. Rajpurkar, “Med-flamingo: a multimodal medical few-shot learner,” in *Machine Learning for Health (ML4H)*, pp. 353–367, PMLR, 2023.
- [5] Y. Zhang, N. He, J. Yang, Y. Li, D. Wei, Y. Huang, Y. Zhang, Z. He, and Y. Zheng, “mm-former: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 107–117, Springer, 2022. 1
- [6] OpenAI, “GPT-4 technical report,” *CoRR*, vol. abs/2303.08774, 2023. 1, 63
- [7] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 1, 63, 69, 73, 82
- [8] T. Bai, H. Liang, B. Wan, L. Yang, B. Li, Y. Wang, B. Cui, C. He, B. Yuan, and W. Zhang, “A survey of multimodal large language model from a data-centric perspective,” *arXiv preprint arXiv:2405.16640*, 2024. 1
- [9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning transferable visual models from natural language supervision,” in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, vol. 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, PMLR, 2021. 13, 14, 44, 73
- [10] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, “CLAP learning audio concepts from natural language supervision,” in *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pp. 1–5, IEEE, 2023. 1, 14
- [11] P. Ren, Y. Xiao, X. Chang, P.-Y. Huang, Z. Li, B. B. Gupta, X. Chen, and X. Wang, “A survey of deep active learning,” *ACM computing surveys (CSUR)*, vol. 54, no. 9, pp. 1–40, 2021. 2

- [12] R. Takezoe, X. Liu, S. Mao, M. T. Chen, Z. Feng, S. Zhang, X. Wang, *et al.*, “Deep active learning for computer vision: Past and future,” *APSIPA Transactions on Signal and Information Processing*, vol. 12, no. 1, 2023. 2
- [13] Y. Zong, O. Mac Aodha, and T. Hospedales, “Self-supervised multimodal learning: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2
- [14] Z. Bai, P. Wang, T. Xiao, T. He, Z. Han, Z. Zhang, and M. Z. Shou, “Hallucination of multimodal large language models: A survey,” *arXiv preprint arXiv:2404.18930*, 2024. 2
- [15] H. Liu, W. Xue, Y. Chen, D. Chen, X. Zhao, K. Wang, L. Hou, R. Li, and W. Peng, “A survey on hallucination in large vision-language models,” *arXiv preprint arXiv:2402.00253*, 2024. 2
- [16] T. Baltrusaitis, C. Ahuja, and L. Morency, “Multimodal machine learning: A survey and taxonomy,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, 2019. 5, 7, 17, 19
- [17] W. Guo, J. Wang, and S. Wang, “Deep multimodal representation learning: A survey,” *Ieee Access*, vol. 7, pp. 63373–63394, 2019. 5, 7
- [18] L. Yu, J. Chen, A. Sinha, M. Wang, Y. Chen, T. L. Berg, and N. Zhang, “Commercemm: Large-scale commerce multimodal representation learning with omni retrieval,” in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4433–4442, 2022. 5
- [19] Z. Zhang, A. Zhang, M. Li, H. Zhao, G. Karypis, and A. Smola, “Multimodal chain-of-thought reasoning in language models,” *arXiv preprint arXiv:2302.00923*, 2023. 5
- [20] J. Y. Koh, D. Fried, and R. R. Salakhutdinov, “Generating images with multimodal language models,” *Advances in Neural Information Processing Systems*, vol. 36, 2024. 5
- [21] Y. Yang, K.-T. Wang, D.-C. Zhan, H. Xiong, and Y. Jiang, “Comprehensive semi-supervised multi-modal learning,” in *IJCAI*, pp. 4092–4098, 2019. 7
- [22] A. Katharopoulos and F. Fleuret, “Not all samples are created equal: Deep learning with importance sampling,” in *International conference on machine learning*, pp. 2525–2534, PMLR, 2018. 8
- [23] M. Ren, W. Zeng, B. Yang, and R. Urtasun, “Learning to reweight examples for robust deep learning,” in *International conference on machine learning*, pp. 4334–4343, PMLR, 2018. 8
- [24] B. Settles, *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning, Morgan & Claypool Publishers, 2012. 10, 26
- [25] X. Wang, D. Kumar, N. Thome, M. Cord, and F. Precioso, “Recipe recognition with large multimodal food dataset,” in *2015 IEEE International Conference on Multimedia & Expo Workshops, ICME Workshops 2015, Turin, Italy, June 29 - July 3, 2015*, pp. 1–6, IEEE Computer Society, 2015. 10, 26, 51
- [26] D. Roth and K. Small, “Margin-based active learning for structured output spaces,” in *Machine Learning: ECML 2006, 17th European Conference on Machine Learning, Berlin, Germany, September 18-22, 2006, Proceedings*, vol. 4212 of *Lecture Notes in Computer Science*, pp. 413–424, Springer, 2006. 10

- [27] W. H. Beluch, T. Genewein, A. Nürnberger, and J. M. Köhler, “The power of ensembles for active learning in image classification,” in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 9368–9377, Computer Vision Foundation / IEEE Computer Society, 2018. 10, 17
- [28] Y. Gal and Z. Ghahramani, “Dropout as a bayesian approximation: Representing model uncertainty in deep learning,” in *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, vol. 48 of *JMLR Workshop and Conference Proceedings*, pp. 1050–1059, JMLR.org, 2016. 10
- [29] A. Parvaneh, E. Abbasnejad, D. Teney, R. Haffari, A. van den Hengel, and J. Q. Shi, “Active learning by feature mixing,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 12227–12236, IEEE, 2022. 10, 11
- [30] M. Ducoffe and F. Precioso, “Adversarial active learning for deep networks: a margin based approach,” *CoRR*, vol. abs/1802.09841, 2018. 10, 26
- [31] S. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, “Deepfool: A simple and accurate method to fool deep neural networks,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 2574–2582, IEEE Computer Society, 2016. 10
- [32] H. T. Nguyen and A. W. M. Smeulders, “Active learning using pre-clustering,” in *Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Alberta, Canada, July 4-8, 2004*, vol. 69 of *ACM International Conference Proceeding Series*, ACM, 2004. 11
- [33] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009. 11
- [34] O. Sener and S. Savarese, “Active learning for convolutional neural networks: A core-set approach,” in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, OpenReview.net, 2018. 11, 17, 26
- [35] E. Biyik, K. Wang, N. Anari, and D. Sadigh, “Batch active learning using determinantal point processes,” *CoRR*, vol. abs/1906.07975, 2019. 11
- [36] S. Sinha, S. Ebrahimi, and T. Darrell, “Variational adversarial active learning,” in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 5971–5980, IEEE, 2019. 11
- [37] T. J. Viering, J. H. Krijthe, and M. Loog, “Nuclear discrepancy for single-shot batch active learning,” *Mach. Learn.*, vol. 108, no. 8-9, pp. 1561–1599, 2019. 11
- [38] J. Su, Y. Tsai, K. Sohn, B. Liu, S. Maji, and M. Chandraker, “Active adversarial domain adaptation,” in *IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020*, pp. 728–737, IEEE, 2020. 11
- [39] C. Shui, F. Zhou, C. Gagné, and B. Wang, “Deep active learning: Unified and principled method for query and training,” in *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, vol. 108 of *Proceedings of Machine Learning Research*, pp. 1308–1318, PMLR, 2020. 11
- [40] G. Hacohen, A. Dekel, and D. Weinshall, “Active learning on a budget: Opposite strategies suit high and low budgets,” in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, vol. 162 of *Proceedings of Machine Learning Research*, pp. 8175–8195, PMLR, 2022. 11

- [41] J. T. Ash, C. Zhang, A. Krishnamurthy, J. Langford, and A. Agarwal, “Deep batch active learning by diverse, uncertain gradient lower bounds,” in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, OpenReview.net, 2020. 11, 18, 19, 25, 26, 32
- [42] D. Arthur and S. Vassilvitskii, “k-means++: the advantages of careful seeding,” in *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007, New Orleans, Louisiana, USA, January 7-9, 2007*, pp. 1027–1035, SIAM, 2007. 11
- [43] M. Yuan, H. Lin, and J. L. Boyd-Graber, “Cold-start active learning through self-supervised language modeling,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 7935–7948, Association for Computational Linguistics, 2020. 11, 40, 41
- [44] J. Devlin, M. Chang, K. Lee, and K. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Association for Computational Linguistics, 2019. 11, 27, 54
- [45] A. L. Chandra, S. V. Desai, C. Devaguptapu, and V. N. Balasubramanian, “On initial pools for deep active learning,” in *NeurIPS 2020 Workshop on Pre-registration in Machine Learning, 11 December 2020, Virtual Event*, vol. 148 of *Proceedings of Machine Learning Research*, pp. 14–32, PMLR, 2020. 11
- [46] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, “A simple framework for contrastive learning of visual representations,” in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, vol. 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607, PMLR, 2020. 11, 12
- [47] W. V. Gansbeke, S. Vandenhende, S. Georgoulis, M. Proesmans, and L. V. Gool, “SCAN: learning to classify images without labels,” in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part X*, vol. 12355 of *Lecture Notes in Computer Science*, pp. 268–285, Springer, 2020. 11, 12
- [48] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. 11
- [49] J. S. K. Yi, M. Seo, J. Park, and D. Choi, “PT4AL: using self-supervised pretext tasks for active learning,” in *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXVI*, vol. 13686 of *Lecture Notes in Computer Science*, pp. 596–612, Springer, 2022. 12, 40, 41
- [50] X. Chen and K. He, “Exploring simple siamese representation learning,” in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 15750–15758, Computer Vision Foundation / IEEE, 2021. 12
- [51] G. Hachohen, A. Dekel, and D. Weinshall, “Active learning on a budget: Opposite strategies suit high and low budgets,” in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, vol. 162 of *Proceedings of Machine Learning Research*, pp. 8175–8195, PMLR, 2022. 12, 40, 41, 56
- [52] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, “Emerging properties in self-supervised vision transformers,” in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 9630–9640, IEEE, 2021. 12

- [53] L. Chen, Y. Bai, S. Huang, Y. Lu, B. Wen, A. L. Yuille, and Z. Zhou, “Making your first choice: To address cold start problem in medical active learning,” in *Medical Imaging with Deep Learning, MIDL 2023, 10-12 July 2023, Nashville, TN, USA*, vol. 227 of *Proceedings of Machine Learning Research*, pp. 496–525, PMLR, 2023. 12, 40, 41
- [54] X. Chen, H. Fan, R. B. Girshick, and K. He, “Improved baselines with momentum contrastive learning,” *CoRR*, vol. abs/2003.04297, 2020. 12, 44
- [55] S. Swayamdipta, R. Schwartz, N. Lourie, Y. Wang, H. Hajishirzi, N. A. Smith, and Y. Choi, “Dataset cartography: Mapping and diagnosing datasets with training dynamics,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020* (B. Webber, T. Cohn, Y. He, and Y. Liu, eds.), pp. 9275–9293, Association for Computational Linguistics, 2020. 12, 57, 59
- [56] Y. Xie, H. Lu, J. Yan, X. Yang, M. Tomizuka, and W. Zhan, “Active finetuning: Exploiting annotation budget in the pretraining-finetuning paradigm,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 23715–23724, IEEE, 2023. 12, 41, 46, 47, 51, 52, 53, 54, 62
- [57] W. Wang, D. Tran, and M. Feiszli, “What makes training multi-modal classification networks hard?,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 12692–12702, Computer Vision Foundation / IEEE, 2020. 13, 18
- [58] Y. Huang, J. Lin, C. Zhou, H. Yang, and L. Huang, “Modality competition: What makes joint training of multi-modal network fail in deep learning? (provably),” in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, vol. 162 of *Proceedings of Machine Learning Research*, pp. 9226–9259, PMLR, 2022. 13
- [59] N. Wu, S. Jastrzebski, K. Cho, and K. J. Geras, “Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks,” in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, vol. 162 of *Proceedings of Machine Learning Research*, pp. 24043–24055, PMLR, 2022. 13
- [60] X. Peng, Y. Wei, A. Deng, D. Wang, and D. Hu, “Balanced multimodal learning via on-the-fly gradient modulation,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 8228–8237, IEEE, 2022. 13, 18
- [61] Y. Zong, O. M. Aodha, and T. M. Hospedales, “Self-supervised multimodal learning: A survey,” *CoRR*, vol. abs/2304.01008, 2023. 13
- [62] H. Xu, G. Ghosh, P. Huang, D. Okhonko, A. Aghajanyan, F. Metze, L. Zettlemoyer, and C. Feichtenhofer, “Videoclip: Contrastive pre-training for zero-shot video-text understanding,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pp. 6787–6800, Association for Computational Linguistics, 2021. 14
- [63] M. Zolfaghari, Y. Zhu, P. V. Gehler, and T. Brox, “Crossclr: Cross-modal contrastive learning for multi-modal video representations,” in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 1430–1439, IEEE, 2021. 14
- [64] S. Ma, Z. Zeng, D. McDuff, and Y. Song, “Active contrastive learning of audio-visual video representations,” in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, OpenReview.net, 2021. 14

- [65] S. Jenni, A. Black, and J. P. Collomosse, “Audio-visual contrastive learning with temporal self-supervision,” in *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pp. 7996–8004, AAAI Press, 2023. 14
- [66] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, “Unsupervised learning of visual features by contrasting cluster assignments,” in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 14
- [67] J. Li, P. Zhou, C. Xiong, and S. C. H. Hoi, “Prototypical contrastive learning of unsupervised representations,” in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, OpenReview.net, 2021. 14, 44, 45
- [68] H. Alwassel, D. Mahajan, B. Korbar, L. Torresani, B. Ghanem, and D. Tran, “Self-supervised learning by cross-modal audio-video clustering,” in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 14
- [69] B. Shi, W. Hsu, K. Lakhota, and A. Mohamed, “Learning audio-visual speech representation by masked multimodal cluster prediction,” in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, OpenReview.net, 2022. 14
- [70] H. Bao, W. Wang, L. Dong, and F. Wei, “Vl-beit: Generative vision-language pretraining,” *CoRR*, vol. abs/2206.01127, 2022. 14
- [71] J. Li, R. R. Selvaraju, A. Gotmare, S. R. Joty, C. Xiong, and S. C. Hoi, “Align before fuse: Vision and language representation learning with momentum distillation,” in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 9694–9705, 2021. 14
- [72] W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Y. Zou, “Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning,” in *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. 14, 44, 46
- [73] Z. Sun, S. Shen, S. Cao, H. Liu, C. Li, Y. Shen, C. Gan, L.-Y. Gui, Y.-X. Wang, Y. Yang, *et al.*, “Aligning large multimodal models with factually augmented rlhf,” *arXiv preprint arXiv:2309.14525*, 2023. 15
- [74] F. Liu, K. Lin, L. Li, J. Wang, Y. Yacoob, and L. Wang, “Mitigating hallucination in large multi-modal models via robust instruction tuning,” in *The Twelfth International Conference on Learning Representations*, 2023. 15, 64
- [75] A. Favero, L. Zancato, M. Trager, S. Choudhary, P. Perera, A. Achille, A. Swaminathan, and S. Soatto, “Multi-modal hallucination control by visual information grounding,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14303–14312, 2024. 15
- [76] Z. Yue, L. Zhang, and Q. Jin, “Less is more: Mitigating multimodal hallucination from an eos decision perspective,” *arXiv preprint arXiv:2402.14545*, 2024. 15, 64, 74

- [77] Y. Zhou, C. Cui, J. Yoon, L. Zhang, Z. Deng, C. Finn, M. Bansal, and H. Yao, “Analyzing and mitigating object hallucination in large vision-language models,” *arXiv preprint arXiv:2310.00754*, 2023. 15, 67
- [78] S. Leng, H. Zhang, G. Chen, X. Li, S. Lu, C. Miao, and L. Bing, “Mitigating object hallucinations in large vision-language models through visual contrastive decoding,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024. 15, 64, 66, 70
- [79] Y. Han, L. Nie, J. Yin, J. Wu, and Y. Yan, “Visual perturbation-aware collaborative learning for overcoming the language prior problem,” *arXiv preprint arXiv:2207.11850*, 2022. 15
- [80] S. Tong, Z. Liu, Y. Zhai, Y. Ma, Y. LeCun, and S. Xie, “Eyes wide shut? exploring the visual shortcomings of multimodal llms,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024. 15
- [81] D. Jiang, Y. Liu, S. Liu, X. Zhang, J. Li, H. Xiong, and Q. Tian, “From clip to dino: Visual encoders shout in multi-modal large language models,” *arXiv preprint arXiv:2310.08825*, 2023. 15
- [82] B. Zhai, S. Yang, X. Zhao, C. Xu, S. Shen, D. Zhao, K. Keutzer, M. Li, T. Yan, and X. Fan, “Halle-switch: Rethinking and controlling object existence hallucinations in large vision language models for detailed caption,” *arXiv preprint arXiv:2310.01779*, 2023. 15, 76
- [83] Q. Huang, X. Dong, P. Zhang, B. Wang, C. He, J. Wang, D. Lin, W. Zhang, and N. 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*, pp. 13418–13427, 2024. 15, 64, 70, 74
- [84] W. Guo, J. Wang, and S. Wang, “Deep multimodal representation learning: A survey,” *IEEE Access*, vol. 7, pp. 63373–63394, 2019. 17
- [85] Z. Liu, H. Ding, H. Zhong, W. Li, J. Dai, and C. He, “Influence selection for active learning,” in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 9254–9263, IEEE, 2021. 17
- [86] T. Yuan, F. Wan, M. Fu, J. Liu, S. Xu, X. Ji, and Q. Ye, “Multiple instance active learning for object detection,” in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 5330–5339, Computer Vision Foundation / IEEE, 2021. 17
- [87] M. Yuan, H. Lin, and J. L. Boyd-Graber, “Cold-start active learning through self-supervised language modeling,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 7935–7948, Association for Computational Linguistics, 2020. 17
- [88] K. Margatina, G. Vernikos, L. Barrault, and N. Aletras, “Active learning by acquiring contrastive examples,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pp. 650–663, Association for Computational Linguistics, 2021. 17
- [89] P. Ren, Y. Xiao, X. Chang, P. Huang, Z. Li, B. B. Gupta, X. Chen, and X. Wang, “A survey of deep active learning,” *ACM Comput. Surv.*, vol. 54, no. 9, pp. 180:1–180:40, 2022. 17

- [90] A. Kirsch, J. van Amersfoort, and Y. Gal, “Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning,” in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 7024–7035, 2019. 21
- [91] S. Frank, E. Bugliarello, and D. Elliott, “Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 9847–9857, Nov. 2021. 23
- [92] I. Gat, I. Schwartz, and A. G. Schwing, “Perceptual score: What data modalities does your model perceive?,” in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual* (M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, eds.), pp. 21630–21643, 2021. 23
- [93] J. Hessel and L. Lee, “Does my multimodal model learn cross-modal interactions? it’s harder to tell than you might think!,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020* (B. Webber, T. Cohn, Y. He, and Y. Liu, eds.), pp. 861–877, Association for Computational Linguistics, 2020. 23
- [94] L. S. Shapley, “A value for n-person games,” *Contributions to the Theory of Games*, vol. 2, 1953. 23
- [95] S. M. Lundberg and S. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA* (I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, eds.), pp. 4765–4774, 2017. 23
- [96] R. Arandjelovic and A. Zisserman, “Look, listen and learn,” in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 609–617, IEEE Computer Society, 2017. 26, 50
- [97] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, “The kinetics human action video dataset,” *CoRR*, vol. abs/1705.06950, 2017. 26, 50
- [98] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, “Vggsound: A large-scale audio-visual dataset,” in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pp. 721–725, IEEE, 2020. 26, 51
- [99] Y. Gal, R. Islam, and Z. Ghahramani, “Deep bayesian active learning with image data,” in *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, vol. 70 of *Proceedings of Machine Learning Research*, pp. 1183–1192, PMLR, 2017. 26
- [100] R. Caramalau, B. Bhattarai, and T. Kim, “Sequential graph convolutional network for active learning,” in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 9583–9592, Computer Vision Foundation / IEEE, 2021. 27, 51, 52, 53
- [101] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, OpenReview.net, 2019. 27

- [102] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, “A closer look at spatiotemporal convolutions for action recognition,” in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 6450–6459, Computer Vision Foundation / IEEE Computer Society, 2018. 27, 54
- [103] X. Wang, R. B. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 7794–7803, Computer Vision Foundation / IEEE Computer Society, 2018. 27, 90
- [104] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017. 27
- [105] P. Ren, Y. Xiao, X. Chang, P. Huang, Z. Li, B. B. Gupta, X. Chen, and X. Wang, “A survey of deep active learning,” *ACM Comput. Surv.*, vol. 54, no. 9, pp. 180:1–180:40, 2022. 39
- [106] A. van den Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *CoRR*, vol. abs/1807.03748, 2018. 44
- [107] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 9726–9735, Computer Vision Foundation / IEEE, 2020. 44
- [108] Y. Fan, W. Xu, H. Wang, J. Wang, and S. Guo, “PMR: prototypical modal rebalance for multimodal learning,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 20029–20038, IEEE, 2023. 44
- [109] J. T. Ash, C. Zhang, A. Krishnamurthy, J. Langford, and A. Agarwal, “Deep batch active learning by diverse, uncertain gradient lower bounds,” in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, OpenReview.net, 2020. 51, 52, 53
- [110] M. Shen, Y. Huang, J. Yin, H. Zou, D. Rajan, and S. See, “Towards balanced active learning for multimodal classification,” in *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pp. 3434–3445, ACM, 2023. 51, 52, 53, 60
- [111] A. Parvaneh, E. Abbasnejad, D. Teney, R. Haffari, A. van den Hengel, and J. Q. Shi, “Active learning by feature mixing,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 12227–12236, IEEE, 2022. 52, 53
- [112] O. Sener and S. Savarese, “Active learning for convolutional neural networks: A core-set approach,” in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, OpenReview.net, 2018. 51, 52, 53
- [113] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778, IEEE Computer Society, 2016. 54
- [114] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, OpenReview.net, 2019. 54

- [115] L. Beyer, A. Steiner, A. S. Pinto, A. Kolesnikov, X. Wang, D. Salz, M. Neumann, I. Al-abdulmohsin, M. Tschannen, E. Bugliarello, *et al.*, “Paligemma: A versatile 3b vlm for transfer,” *arXiv preprint arXiv:2407.07726*, 2024. 63, 73
- [116] A. Rohrbach, L. A. Hendricks, K. Burns, T. Darrell, and K. Saenko, “Object hallucination in image captioning,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pp. 4035–4045, Association for Computational Linguistics, 2018. 64, 73
- [117] J. Chen, D. Yang, T. Wu, Y. Jiang, X. Hou, M. Li, S. Wang, D. Xiao, K. Li, and L. Zhang, “Detecting and evaluating medical hallucinations in large vision language models,” *CoRR*, vol. abs/2406.10185, 2024. 64
- [118] W. Huang, C. Wang, R. Zhang, Y. Li, J. Wu, and L. Fei-Fei, “Voxposer: Composable 3d value maps for robotic manipulation with language models,” in *Conference on Robot Learning, CoRL 2023, 6-9 November 2023, Atlanta, GA, USA*, vol. 229 of *Proceedings of Machine Learning Research*, pp. 540–562, PMLR, 2023. 64
- [119] L. Wang, L. Li, D. Dai, D. Chen, H. Zhou, F. Meng, J. Zhou, and X. Sun, “Label words are anchors: An information flow perspective for understanding in-context learning,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 9840–9855, Association for Computational Linguistics, 2023. 64, 68
- [120] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 66
- [121] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755, Springer, 2014. 67
- [122] M. He, Y. Liu, B. Wu, J. Yuan, Y. Wang, T. Huang, and B. Zhao, “Efficient multimodal learning from data-centric perspective,” *CoRR*, vol. abs/2402.11530, 2024. 73
- [123] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, *et al.*, “Gemma: Open models based on gemini research and technology,” *arXiv preprint arXiv:2403.08295*, 2024. 73
- [124] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, “Sigmoid loss for language image pre-training,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023. 73
- [125] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, “Judging llm-as-a-judge with mt-bench and chatbot arena,” 2023. 73
- [126] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, *et al.*, “The llama 3 herd of models,” *arXiv preprint arXiv:2407.21783*, 2024. 73
- [127] B. Zhao, B. Wu, M. He, and T. Huang, “Svit: Scaling up visual instruction tuning,” *arXiv preprint arXiv:2307.04087*, 2023. 73
- [128] L. Jing, R. Li, Y. Chen, M. Jia, and X. Du, “FAITHSCORE: evaluating hallucinations in large vision-language models,” *CoRR*, vol. abs/2311.01477, 2023. 73

-
- [129] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, *et al.*, “Llama 2: Open foundation and fine-tuned chat models,” *arXiv preprint arXiv:2307.09288*, 2023. 74
- [130] OpenAI, “Introducing chatgpt,” 2023. 74
- [131] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, and H. Yang, “Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework,” in *International conference on machine learning*, pp. 23318–23340, PMLR, 2022. 74
- [132] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “Lora: Low-rank adaptation of large language models,” *arXiv preprint arXiv:2106.09685*, 2021. 74
- [133] C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, Z. Qiu, W. Lin, J. Yang, X. Zheng, K. Li, X. Sun, and R. Ji, “MME: A comprehensive evaluation benchmark for multimodal large language models,” *CoRR*, vol. abs/2306.13394, 2023. 82
- [134] Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, and J. Wen, “Evaluating object hallucination in large vision-language models,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 292–305, Association for Computational Linguistics, 2023. 82