THE HONG KONG
POLYTECHNIC UNIVERSITY
香港理工大學
Pao Yue-kong Library
包玉剛圖書館

Copyright Undertaking

This thesis is protected by copyright, with all rights reserved.

By reading and using the thesis, the reader understands and agrees to the following terms:

1. The reader will abide by the rules and legal ordinances governing copyright regarding the use of the thesis.
2. The reader will use the thesis for the purpose of research or private study only and not for distribution or further reproduction or any other purpose.
3. The reader agrees to indemnify and hold the University harmless from and against any loss, damage, cost, liability or expenses arising from copyright infringement or unauthorized usage.

IMPORTANT

If you have reasons to believe that any materials in this thesis are deemed not suitable to be distributed in this form, or a copyright owner having difficulty with the material being included in our database, please contact lbsys@polyu.edu.hk providing details. The Library will look into your claim and consider taking remedial action upon receipt of the written requests.

Pao Yue-kong Library, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

http://www.lib.polyu.edu.hk

TOWARDS ROBUST MULTIMODAL LEARNING IN THE OPEN WORLD

FUSHUO HUO

PhD

The Hong Kong Polytechnic University

2025

The Hong Kong Polytechnic University
Department of Computing

Towards Robust Multimodal Learning in the Open World

Fushuo HUO

A thesis submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy
April 2025

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgment has been made in the text.

Signature: ____________________

Name of Student: Fushuo HUO

Abstract

The rapid evolution of machine learning has propelled neural networks to unprecedented success across diverse domains. In particular, multimodal learning has emerged as a transformative paradigm, leveraging complementary information from heterogeneous data streams (e.g., text, vision, audio) to advance contextual reasoning and intelligent decision-making. Despite these advancements, current neural network-based models often fall short in open-world environments characterized by inherent unpredictability, where unpredictable environmental composition dynamics, incomplete modality inputs, and spurious distributions relations critically undermine system reliability. While humans naturally adapt to such dynamic, ambiguous scenarios, artificial intelligence systems exhibit stark limitations in robustness, particularly when processing multimodal signals under real-world complexity. This study investigates the fundamental challenge of multimodal learning robustness in open-world settings, aiming to bridge the gap between controlled experimental performance and practical deployment requirements. Here, we study the multimodal learning robustness in the open world settings:

(1). Humans can extrapolate new concepts from previously learned multi-modal knowledge. This ability is known as compositional generalization, while neural networks have deficiencies in compositional generalization robustness, struggling to reliably handle unseen compositions due to rigid feature representations and over-reliance on training data biases. (2). Humans can seamlessly infer unimodal inputs based on

memorized contextual multimodal information, with robust inference in the absence of modality. However, neural networks hardly achieve satisfactory results when inferring unimodal inputs, based on integrated multimodal information. (3). With the development of large language models (LLMs), large-scale multimodal large language models (MLLMs), especially large vision language models (LVLMs), have demonstrated expressing comprehensive abilities, approaching or even surpassing human abilities. However, most LVLMs are derived from LLMs by instruction tuning on multimodal datasets. LVLMs usually have the strong language modality prior or statistical bias to LLMs, which is one of the main reasons that arises the significant challenge problem known as 'hallucination', even when queried by simple questions.

In summary, we study above three problems to improve *class-level and modality-level multimodal robustness* in terms of composition gneralization robustness (i.e., class-level), modality missing robustness (i.e., modality-level), and modality prior robustness (i.e., modality-level). Concretely, In Chapter 3, we propose a novel Progressive Cross-primitive Compatibility (ProCC) network, mimicking the human learning progress of recognizing the multimodal compositions to improve the modality composition ability. In Chapter 4, we propose the customized crossmodal knowledge distillation (C^{2}KD) to inherit multimodal knowledge during the pre-training period, and enhance the inference robustness when missing some modalities. In Chapter 5, we propose the train-free decoding strategy to alleviate language modality prior of LVLMs to mitigate the hallucination issues while not compromising general abilities of foundation models. Extensive experimental evaluations and ablation studies show the performance advantages of our works with provable advances in robustness abilities for multiple modalities.

Publications arising from the thesis

- **Fushuo Huo**, Wenchao Xu, Song Guo, Jingcai Guo, Haozhao Wang, Ziming Liu, Xiaocheng Lu, “ProCC: Progressive Cross-Primitive Compatibility for Open-World Compositional Zero-Shot Learning”, Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2024.

- **Fushuo Huo**, Wenchao Xu, Jingcai Guo, Haozhao Wang, Song Guo, “C^2KD: Bridging the Modality Gap for Cross-Modal Knowledge Distillation”, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

- **Fushuo Huo**, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, Peilin Zhao, “Self-Introspective Decoding: Alleviating Hallucinations for Large Vision-Language Models”, International Conference on Learning Representations (ICLR), 2025.

- **Fushuo Huo**, Wenchao Xu, Song Guo, Jingcai Guo, Haozhao Wang, Yunfeng Fan, “Non-exemplar Online Class-Incremental Continual Learning via Dual-Prototype Self-Augment and Refinement”, Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2024.

- **Fushuo Huo**, Ziming Liu, Jingcai Guo, Wenchao Xu, Song Guo, “UTDNet: A Unified Triplet Decoder Network for Multimodal Salient Object Detection”, Neural Networks (NN), 2024.

- Yunfeng Fan, Wenchao Xu, Haozhao Wang, **Fushuo Huo**, Jinyu Chen, and Song Guo, "Overcome Modal Bias in Multi-modal Federated Learning via Balanced Modality Selection", European Conference on Computer Vision (ECCV), 2024.

- Ziming Liu, Song Guo, Xiaocheng Lu, Jingcai Guo, Jiewei Zhang, Yue Zeng, and **Fushuo Huo**, "$(ML)^2$P-Encoder: On Exploration of Channel-Class Correlation for Multi-Label Zero-Shot Learning", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

- Jingcai Guo, Song Guo, Qihua Zhou, Ziming Liu, Xiaocheng Lu, and **Fushuo Huo**, "Graph knows unknowns: Reformulate zero-shot learning as sample-level graph recognition", Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2023.

- Ziming Liu, Song Guo, Jingcai Guo, Yuanyuan Xu, and **Fushuo Huo**, "Towards Unbiased Multi-Label Zero-Shot Learning with Pyramid and Semantic Attention", IEEE Transactions on Multimedia (IEEE TMM), 2022.

Acknowledgments

It is highly privileged for me to take this opportunity to express my sincere gratitude to those who have helped and supported me on my way to pursuing my Ph.D. degree.

First of all, I am immensely grateful to my supervisors, Prof. Song Guo and Prof. Wenchao Xu, for their invaluable guidance, expertise, and patience. Their mentorship and insightful feedback have been instrumental in shaping the direction of my research. I would also like to express my sincere gratitude to Prof. Bin Xiao, Prof. Jingcai Guo, and Dr. Haozhao Wang for their valuable input and constructive suggestions, which have greatly spurred me on the academic journey. Thanks to all my groupmates in the Pervasive Intelligence Lab (PEILAB). I will never forget the happy and hard times we undergo together. I am also immensely grateful to my visiting university's supervisors, Prof. Dacheng Tao and Prof. Baosheng Yu. They give me the opportunity to study at the Nanyang Technological University and inspire me a lot in many ongoing topics. Meanwhile, thanks to all the groupmates in the Generative AI Lab at the Nanyang Technological University. I will never forget the happy sports and nervous deadline times. I would also like to express my gratitude to Dr. Peilin Zhao, Dr. Zhong Zhang, and other friends at Tencent for spending a fulfilling and happy time there. Also, I would like to thank Prof. WANG Cong and Prof. GONG Shimin for serving on my Ph.D. defense as external committees, and thank Prof. LUO Xiapu Daniel for serving on my Ph.D. defense as the BoE chair. Lastly, thanks to all the anonymous reviewers, whether the scoring was ACCEPT or

REJECTION, I learned a lot from the submission and rebuttal periods. Thanks for all the ACCEPT and REJECTION!

I would like to express my heartfelt gratitude to my parents, who have struggled to maintain the family and have wholeheartedly supported me in pursuing my master's and doctoral degrees. Their resilience and persistence inspire me to overcome various setbacks on the academic journey. I am truly fortunate to have a mother in my life who, though physically fragile, possesses an iron will. Her unwavering support has profoundly shaped the person I am today. I would also like to thank my lovely girlfriend, Qian Zhang, whose smile has soothed my soul in the face of adversity. Thanks for my family's patience, understanding, and support. Without their support, I could not achieve this goal and accomplishment.

Table of Contents

Abstract **i**

Publications arising from the thesis **iii**

Acknowledgments **v**

List of Figures **xi**

List of Tables **xvi**

1 Introduction **1**

1.1 Overview 2

1.2 Overall Challenges 3

1.3 Research Framework 4

1.4 Thesis Contribution 6

1.5 Thesis Organization 8

2 Background **11**

2.1 Composition Generalization Robustness 11

2.1.1 Problem Formulation 11

2.1.2 Composition Zero Shot Learning 12

2.2 Knowledge Distillation over Cross Modality 14

2.2.1 Unimodal Knowledge Distillation 14

2.2.2 Cross-modal Knowledge Distillation 15

2.3 Multimodal Hallucination 16

2.3.1 Multimodal Large Language Models 16

2.3.2 Decoding Strategy in LLMs 17

2.3.3 Hallucination in Foundation Models 18

3 ProCC: Progressive Cross-Primitive Compatibility for Composition Generalization 19

3.1 Challenges and Motivations 19

3.2 Progressive Cross-primitive Compatibility (ProCC) 22

3.2.1 Revisit Visual Product 23

3.2.2 Cross-primitive Compatibility Module 24

3.2.3 Visual Explanation 26

3.2.4 Progressive Learning Strategy 27

3.3 Experimental Results 31

3.3.1 Datasets and Evaluation Metrics 31

3.3.2 Baselines and Implementation Details 31

3.3.3 Open-World CZSL (OW-CZSL) Results 32

3.3.4 Partial-supervision CZSL (pCZSL) Results 33

3.4 Ablation Analysis 35

3.5 Chapter Summary 36

4 C^{2}KD: Bridging the Modality Gap for Cross-Modal Knowledge Distillation 38

4.1 Challenges and Motivations 38

4.2 Cross-Modal KD Effectiveness Analysis 42

4.3 Customized Cross-modal Knowledge Distillation (C^{2}KD) 44

4.3.1 Formulation of C^{2}KD 47

4.3.2 Analysis of Cross-modal Knowledge distillation 48

4.4 Experimental Results 51

4.4.1 Multimodal Classification 51

4.4.2 Multimodal Semantic Segmentation 54

4.5 Ablation and Sensitivity Analysis 55

4.6 Discussion 60

4.7 Chapter Summary 62

5 SID: Self-Introspective Decoding for Modality Prior to Alleviate Hallucinations for Multimodal Large Models 63

5.1 Challenges and Motivations 63

5.2 Preliminary and Discussions 66

5.2.1 Paradigm of LVLMs Generation 66

5.2.2 Re-thinking Contrastive Decoding in LVLMs 67

5.3 Self-Introspective Decoding (SID) 70

5.3.1 Understanding the Self-Introspective Pre-trained LVLMs . . . 70

5.3.2 Context and Text-aware Token Selection (CT^{2}S) Strategy . . 75

5.4 Experimental Results 77

5.4.1 Experimental Settings 77

5.4.2 Evaluation Results 78

5.5 Ablation Analyses 86

5.6 Discussion 93

5.7 Chapter Summary 95

6 Conclusion and Suggestions for Future Research **96**

6.1 Conclusion 96

6.2 Suggestions for Future Research 98

6.2.1 Multimodal Test-time Adaptation 98

6.2.2 Task-Aware Adaptation of Multimodal LLMs 99

6.2.3 Multimodal Agent as Experts 100

List of Figures

1.1 **Research framework of this thesis.** We organize the positioning of this thesis within the field of robust multimodal learning in the open world. We classify the challenges into class-level and modality-level robustness and illustrate the contributions we focus on for each chapter. 5

2.1 **Intuitive presentation** of cross-modal knowledge distillation. 15

3.1 **The overall concept of our method.** Following the principle of 'forest before trees' [55], human feedforward hierarchy underlies implicit processing for initial vision at a glance (i.e., green rectangle), and feedback connections add details to explicit vision with scrutiny (i.e., red rectangle). As for composition generalization learning, humans first (Ⅰ) learn to recognize overall objects, then (Ⅱ) gradually identify the scrutiny attribute of objects, i.e., state, and finally (Ⅲ) reasonably compose the object and state primitives. Inspired by this, we aim to progressively recognize the object and state primitives and guide the network to exploit discriminative information conditioned on learned knowledge via the CPC module. 20

3.2 **The framework of ProCC.** Features from the encoder (ω) are respectively fed to the object and state (φ_o and φ_s) classifiers, where the Cross-Primitive Compatibility (CPC) aims to model the cross-primitive interactions. Progressive learning strategy is proposed to gradually modulate primitive compatibility, especially for pCZSL. For detailed training procedure, please refers to **Algorithm 2**. Class Activation Maps (CAM) of input samples are illustrated to show visual attention. 22

3.3 **The detailed framework of the object-state Cross-Primitive Compatibility (CPC$_{o\to s}$).** Features from the object classifier (φ_{o-1} and φ_{o-2}) are encoded by learnable Cross-Primitive Memory (CPM) units. Then respectively interact with state features (φ_{s-1} and φ_{s-2}) to achieve compatibility of state features conditioned on objects. . . . 24

3.4 **Visualizations of class activation maps** of ProCC with and without CPC modules on the testing dataset of MIT-States. The discriminative regions are marked with red rectangles. 26

3.5 **Confusion matrices** about prediction probabilities of states conditioned on objects (w/ CPC) or not (w/o CPC). 29

4.1 **The Modality Gap of CMKD. (a)** Top-1 accuracy (followed by average prediction probability of target classes) of each modality. Both modalities utilize ResNet-18 as the backbone. **(b)** Up: Example of three-class classification. Down: Kendall Rank Correlation [81] of soft labels across modalities in VGGSound. A: audio; V: visual; RN: ResNet. 41

4.2 **Evolution of our Customized Cross-modal Knowledge Distillation (C^2KD) method.** (a) Traditional KD [54] with output logits from the fixed teacher. (b) We (partially) tune the teacher with the bidirectional distillation to provide customized teacher knowledge. (c) To bridge the modality gap of CMKD, On-the-Fly Selection Distillation (OFSD) is proposed to filter out samples with distorted rank correlations and perform KD on non-target classes. (d) Additionally, we introduce proxy teacher and proxy student as bridges to progressively transfer receptive cross-modal knowledge. 46

4.3 **Training dynamics analysis.** The solid lines correspond to the test accuracy, and the dotted lines indicate the average number of samples with KRC $< \omega$ each data batch during the training process. Here we set $\{\omega = 0, \text{batchsize} = 64\}$. 49

4.4 **Comparisons of different distance metrics.** The X-axis represents the value of ω. 50

4.5 **Generalizability of each module.** We conduct experiments on VGGsound dataset in terms of DIST [62] and NKD [188]. 55

4.6 **The CKA score of intermediate features on AVE.** 55

4.7 **Analysis of λ_1, λ_2, λ_3.** We conduct experiments on the AVE [158] dataset with ResNet-18 as the multimodal backbones. 59

5.1 **Contrastive Decoding strategies:** (a) Visual Contrastive Decoding (VCD) [94] **manually** distort vision inputs. (b) Instruction Contrastive Decoding (ICD) [170, 84] also **manually** design noisy instruction (negative prompt). We ablate other modules like the vision encoder and tokenizer for clarity. t: `‘Please describe this image in detail.’`; sys.: system prompt. g: generated text tokens. 65

5.2 **Overview of Self-Introspective Decoding (SID).** CT^{2}S: Context and Text-aware Token Selection strategy. LLaVA-1.5 7B is utilized as an example to visualize visual tokens with low and high scores (Equation 5.5). 70

5.3 **Instance Illustration of Different Disturbance Results.** Examples are from MSCOCO inferred by LLaVA-1.5 7B with $i = 3$ and Top-k=50. Hallucinations are marked in red. 71

5.4 **Instance Illustration of Different Disturbance Results.** Examples are from MSCOCO [109] inferred by LLaVA-1.5 7B with $i = 3$ and Top-k=50. Hallucinations are marked in red. 72

5.5 **Visualization Results** of the **least** important vision tokens on discrimination tasks informed by preceding vision and text tokens. LLaVA-1.5 7B with Layer $i = 3$ is utilized. 73

5.6 **Visualization Results of *Adaptively* Selecting** the **least** important vision tokens on open-end generative tasks informed by preceding vision and text tokens. LLaVA-1.5 7B with Layer $i = 3$ is utilized. . . 74

5.7 Analyses of varying i and preserved ratios in CT^{2}S. VE: vision encoder; i: i-th decoder layer. 76

5.8 **GPT-4 assisted benchmark** [206]. **Hallucination** (SHR), **fluency** (1&2-gram), and **detailness** (WPI and SPI) aspects are compared. Larger areas mean better performances. VCD and ours adopt the same sampling decoding. Please zoom in for details. 81

5.9 **Prompts of GPT-4 for evaluations.** 86

5.10 **Prompts of GPT4-V for evaluations.** 87

5.11 **Case Study from LLaVA-Bench-in-the-Wild** on LLaVA-1.5 7B. Hallucinations are marked in red. 88

5.12 **Case Study from LLaVA-Bench-in-the-Wild** on LLaVA-1.5 7B. Hallucinations are marked in red. 89

5.13 **Case Study from LLaVA-Bench-in-the-Wild** on LLaVA-1.5 7B. Hallucinations are marked in red. 90

5.14 **Hyperparameter Sensitivity of α and β** with POPE metric (under the sampling decoding). 91

5.15 **Results of different decoding strategies.** 91

List of Tables

3.1 **Quantitative comparisons in the OW-CZSL setting.** We report the best seen (S), best unseen (U) accuracy, HM, AUC on the test and validation sub-datasets. The best and second-best results are bold and underlined. 30

3.2 **Quantitative comparisons in the pCZSL setting.** We report the seen (S), unseen (U) accuracy, and best harmonic mean (HM) on the test and validation sub-datasets. The best and second-best results are bold and underlined. 33

3.3 **Ablation studies** for both OW-CZSL and pCZSL. 34

4.1 **Performances of traditional KD in CMKD.** The results of distilled student modality infer only on the student modality. **A→V**: Audio teacher modality distills visual student modality; **V→A**: Visual teacher modality distills audio student modality. 39

4.2 **Efficacy Analysis on modality imbalance and soft label misalignment.** *proba.* represents average prediction probability of target class. DKD is with defaulted $\{\alpha = 1, \beta = 8\}$ (Eq. 4.2). 45

4.3 **Comparison results on Visual-Audio and Image-Text datasets.** The metric is the top-1 accuracy (%). Ours‡ means fully updating the teacher model, and Ours† means partially finetuning the top 2 layers. The best is in **bold**, and the second is underlined. 52

4.4 **Comparison results on RGB-Depth semantic segmentation dataset.** The metric denotes the mean Intersection over Union (mIoU: %). Ours+[62] means we replace KL loss with the advanced DIST loss. RN18: ResNet-18; MNV2: MobileNetV2. 56

4.5 **Ablation studies on each module.** (a), (b), (c), and (d) represent the evolution steps of C^{2}KD (Figure 4.2). (e) indicates the results without KD. The metric is the top-1 accuracy (%). 57

4.6 **Comparison results of different Self-KD methods.** 58

4.7 **Comparison results on Visual-Audio and Image-Text datasets.** The metric is the top-1 accuracy (%). RN18: ResNet-18; RN50: ResNet-50; SNV2: ShuffleNet V2 [122]. 59

4.8 **Analysis of the structure of the proxy.** We conduct experiments on the AVE [158] and VGGsound [18] datasets with ResNet-18 as the multimodal backbones. 60

4.9 **Comparison results of different multimodal teachers.** The *italic* numbers mean teachers' accuracy. S: student; T: teacher. 61

5.1 **Efficacy Analyses on CD strategies** on MSCOCO dataset. The *Random* setting means objects absent from the image are chosen randomly, while the *Adversarial* setting prioritizes co-occurring objects which are not present in the image. 68

5.2 **Efficacy Analyses on Vision Token Attention Scores** with POPE metric on MSCOCO dataset and CHAIR metric. We select the Top-100 and Least-100 important vision tokens out of a total of 576 vision tokens of LLaVA-1.5 7B, based on Equation 4.5 (i=3). **VD**: Visual Disturbance; **ID**: Instruction Disturbance. 74

5.3 **Analyses of Contrastive Decoding Mechanisms** on the POPE metric. Hyperparameters are consistent with CD settings. 77

5.4 **Results on the CHAIR metric.** * and * denote adopting the same sampling and greedy decoding strategies, respectively. 79

5.5 **Average results on the POPE metric.** * and * denote adopting the same sampling and greedy decoding strategies, respectively. Results are from the original papers or re-implemented based on official codes. 82

5.6 **GPT4-V assisted hallucination evaluations** [61, 190]. VCD and ours adopt the same sampling decoding strategy. C: correctness; D: detailedness 85

5.7 **LVLM benchmark evaluations.** DoLa, ICD, VCD, and SID employ the same greedy decoding. 85

5.8 **Efficiency Comparisons** on NVIDIA V100. $_{10\%}$ and $_{40\%}$ mean tokens preserved ratios. 91

5.9 **Results on Larger-scale Backbones.** Sampling decoding is adopted and results average of three running times. 92

5.10 **Comparisons with Visual Enhancing Strategy (VDGD).** * denotes employing greedy decoding strategy. 93

5.11 **Analyses of Different Token Selection Strategies** with POPE on MSCOCO dataset and CHAIR metrics. We select the high importance scores (Equation 5.5) of vision tokens (-**Top**) and random vision tokens (-**Random**) for contrastive decoding. Experiments are conducted on LLaVA-1.5 7B. 94

Chapter 1

Introduction

The pursuit of robust multimodal learning systems capable of operating in open-world environments represents a pivotal challenge at the frontier of artificial intelligence (AI). As artificial intelligent systems transition from controlled laboratory settings to real-world deployment in domains ranging from autonomous driving to AI-assisted healthcare, their ability to process heterogeneous data streams (e.g., visual, textual, auditory) with human-like adaptability becomes mission-critical. Although multimodal learning has shown remarkable success in leveraging complementary cross-modal information for tasks like visual question answering and multimodal sentiment analysis, current approaches remain fundamentally constrained by three existential limitations when confronted with open-world dynamics: (1) brittleness to novel concept compositions from known modality primitives, (2) catastrophic performance degradation under partial modality availability, and (3) systemic hallucinations induced by imbalanced modality priors in large multimodal foundation models. In this chapter, we first introduce the overview of our research problems in section 1.1. Next, we describe the challenges of this research topic in Section 1.2. Then, we present the sketch of our research framework in Section 1.3. After that, we present the main contributions of this thesis in Section 1.4. Finally, we give the overall organization of the thesis in Section 1.5

1.1 Overview

Multimodal learning represents a transformative paradigm in artificial intelligence that aims to process and integrate information from diverse data modalities (e.g., text, images, audio, video, sensor data) to mimic human-like perception and decision-making [9, 177]. Unlike unimodal systems that operate on isolated data types, multimodal learning leverages the complementary strengths of heterogeneous signals to enhance contextual understanding, improve inference accuracy, and enable robust performance in real-world scenarios. This interdisciplinary field sits at the intersection of computer vision, natural language processing, and signal processing, driven by the recognition that human cognition inherently synthesizes multisensory inputs for holistic reasoning.

The concept of open-world settings represents a paradigm shift in machine learning, moving beyond the constraints of traditional closed-world assumptions where models operate within predefined, static environments with fully observed data distributions. In contrast, open-world settings reflect the inherent complexity and unpredictability of real-world scenarios, where systems must contend with dynamic data streams, unseen concept compositions, partial or corrupted inputs, and evolving contextual relationships [135, 80]. This framework is particularly critical for multimodal learning systems, as real-world applications, from autonomous robotics [30] to healthcare diagnostics [17, 169], demand adaptability to novel or even poor situations that defy the tidy boundaries of laboratory-trained models.

The robustness of multimodal learning systems from controlled laboratory environments to open-world deployment exposes fundamental limitations in current methodologies. Three interrelated challenges, compositional generalization robustness (class level), modality missing fragility (modality level), and uncontrollable hallucinations (modality level), emerge as critical barriers to reliable open-world multimodal intelligence. These three issues are interconnected facets of a core problem: current multimodal systems lack robustness mechanisms to dynamically balance modality-specific evidence with cross-modal causal relationships in open environments.

1.2 Overall Challenges

The transition from controlled experimental settings to open-world deployment exposes multimodal learning systems to a spectrum of challenges that defy traditional algorithmic assumptions. These challenges stem from the inherent unpredictability, heterogeneity, and dynamic nature of real-world environments, demanding paradigm shifts in model design and evaluation. Below, we dissect the core challenges: (1). Human-like reasoning requires extrapolating to novel compositions of known concepts when adapting to new knowledge, while neural networks might not easily generalize to new compositions composed by multimodal primitives (i.e., objects and attributes). However, humans can extrapolate new concepts from previously learned modality primitives. For instance, if the people are taught what the fried chip and toasted bread are, most of them can recognize the fried bread immediately. This ability is known as compositional generalization [4], which is one of the ultimate targets for artificial intelligence. In this report, such a task is formulated as Compositional Zero-Shot Learning (CZSL). Concretely, the training set contains images with corresponding multi-modal descriptions (primitives), i.e., state and object. The model is expected to recognize unseen compositions based on known primitives, which is non-trivial because object and state are in the multi-modal formations and semantically tangled, i.e., objects in different states often have different appearances, and states can vary greatly conditioned on different objects. The major challenge behind the CZSL lies in how to model the interactions between state and object primitives and extrapolate seen compositions to unseen ones. (2). Real-world systems must operate under partial, asynchronous, or corrupted modality inputs, a stark contrast to curated datasets with aligned, complete data. For instance, autonomous vehicles may lose LiDAR signals during heavy rain while relying solely on cameras, or healthcare algorithms might face missing modality information during emergency diagnostics. This thesis focuses on developing the multimodal knowledge transfer (i.e., crossmodal knowledge distillation (CMKD)) to distill the heterogeneous modality information to another modality, which are not systematically explored. Consequently, the multimodal learning system might not degrade much when some modality is missing or even compared to all modalities available. (3). Recently, in the era of foundation models, with the development of scaling law theory [77] and Graphics Processing Unit (GPU) devices, multimodal learning has evolved into the paradigm of "pre-training on massive datasets and fine-tuning in downstream fields". The advent of Multimodal Large Language Models (MLLMs), especially Large Vision-Language Models (LVLMs), derived from their Large Language Model (LLM) foundations, introduce a critical challenge in open-world multimodal learning: inherent modality bias, where most LVLMs are derived from LLMs with strong language modality prior [94, 170] that dominates cross-modal reasoning, undermining robustness in dynamic, unpredictable environments. For example, given a picture of a black rotting banana, LVLMs will usually recognize the picture as "yellow" when asked for its color. The sticky modality-level prior stems from LVLMs pretraining strategies. This thesis aims to alleviate the modality prior to enabling LVLMs to generate trustworthy answers.

1.3 Research Framework

Our thesis aims to solve the above challenges and propose new frameworks for robust multimodal learning. The structural outline of the thesis is shown in Figure 1.1.

As shown in Figure 1.1, we categorize our works into two base categories in the robust

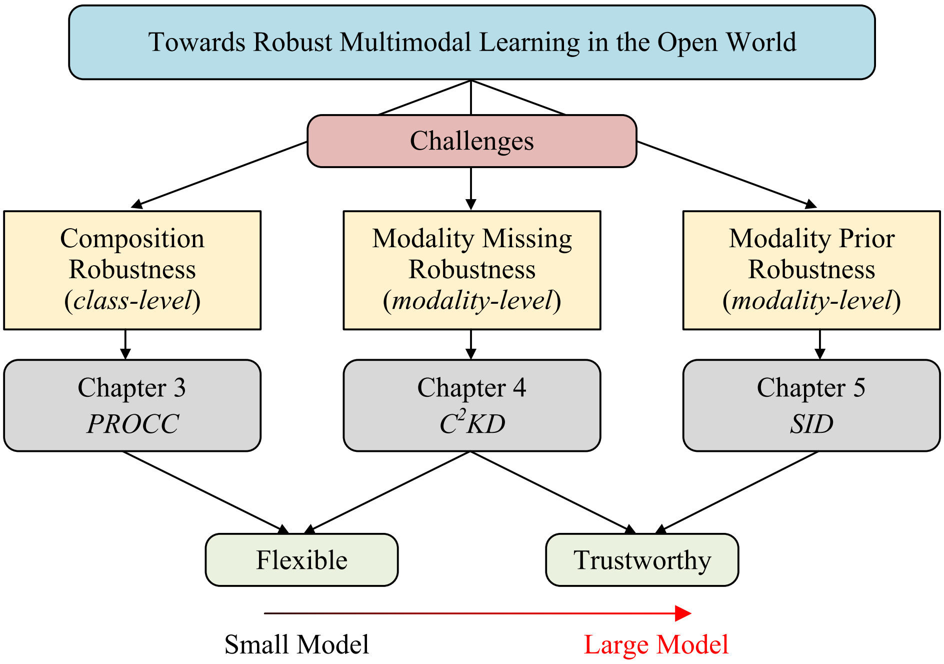

Towards Robust Multimodal Learning in the Open World
Challenges
Composition Robustness (*class-level*)
Modality Missing Robustness (*modality-level*)
Modality Prior Robustness (*modality-level*)
Chapter 3 *PROCC*
Chapter 4 *C²KD*
Chapter 5 *SID*
Flexible
Trustworthy
Small Model
Large Model

Figure 1.1: **Research framework of this thesis.** We organize the positioning of this thesis within the field of robust multimodal learning in the open world. We classify the challenges into class-level and modality-level robustness and illustrate the contributions we focus on for each chapter.

multimodal learning, i.e., *class-level* and *modality-level* robustness. Furthermore, the evolution of the multimodal learning robustness from *small-scale* to *large-scale* model.

Concretely, current approaches are fundamentally constrained by three existential limitations when faced with open-world dynamics: (1) brittleness to novel concept compositions from known modality primitives, (2) catastrophic performance degradation when only partial modalities are available, and (3) systemic hallucinations caused by imbalanced modality priors in large multimodal foundation models. In this thesis, we begin by providing an overview of our research background in Chapter 2. Chapter 3 introduces the Progressive Cross-Primitive Compatibility (ProCC) network, a framework inspired by human cognitive processes for learning multimodal compositions. By modeling cross-primitive dependencies through a curriculum-driven approach, ProCC enhances compositional generalization, enabling robust recognition of unseen object-state pairs in open-world scenarios. Chapter 4 addresses modality-missing robustness via Customized Crossmodal Knowledge Distillation ($\mathrm{C^2KD}$), which preserves cross-modal synergies during pretraining and transfers them to unimodal inference. This ensures consistent performance even when critical modalities are absent, bridging the gap between multimodal training and real-world deployment. In Chapter 5, we tackle hallucination biases in large vision-language models (LVLMs) by proposing a training-free decoding strategy, Self-Introspective Decoding (SID), which suppresses language-prior dominance without fine-tuning or compromising foundational capabilities. SID aligns model outputs with visual evidence while maintaining generative fluency.

1.4 Thesis Contribution

We briefly summarize the contribution of this thesis as follows:

1. Enhancing Modality Composition Generalization Robustness. The first contribution addresses the challenge of recognizing novel compositions of state-object modalities in open-world scenarios (OW-CZSL), where no prior knowledge of valid compositions exists. To tackle this, the thesis introduces the Progressive Cross-Primitive Compatibility (ProCC) framework. By mimicking the human learning process, ProCC employs the Cross-Primitive Compatibility (CPC) module to explicitly model interactions between state and object features using trainable memory units, eliminating reliance on external knowledge. A progressive training paradigm further refines these interactions in an easy-to-hard manner, effectively handling partial supervision (pCZSL) where labels are incomplete. This approach achieves state-of-the-art performance across benchmarks, significantly improving generalization to unseen compositions while filtering invalid cross-modal correlations, thus enhancing *class-level* robustness in dynamic environments.

2. Ensuring Robustness Under Modality Missing. The second contribution targets the problem of modality imbalance and misalignment in cross-modal knowledge distillation (CMKD), which often degrades performance when modalities are missing during inference. The proposed Customized Crossmodal Knowledge Distillation (C^2KD) method bridges these gaps through the dual strategy: On-the-Fly Selection Distillation (OFSD) filters misaligned samples using Kendall Rank Correlation (KRC) metric, while bidirectional distillation between teacher-student proxies preserves cross-modal knowledge. By dynamically adapting to modality gaps, C^2KD ensures the inheritable crossmodal knowledge during the CMKD. Consequently, C^2KD outperforms traditional knowledge distillation methods on audio-visual, image-text, and RGB-depth tasks. This innovation ensures reliable performance in real-world applications, such as sensor failures or resource-constrained settings, by maintaining robustness even when critical modalities are absent, thereby addressing *modality-level* reliability.

3. Balancing Modality Priors to Mitigate Hallucinations. The third contribution confronts the hallucination problem in Large Vision-Language Models (LVLMs), where over-reliance on language priors leads to factually inconsistent outputs. The Self-Introspective Decoding (SID) strategy introduces the Context and Text-aware Token Selection (CT^2S) mechanism to adaptively prune low-importance vision tokens in early decoder layers, amplifying vision-text association errors for contrastive suppression. This train-free approach reduces hallucinations by 12–20% on metrics like POPE and CHAIR while cutting inference costs by 30% compared to methods like VCD [94] and ICD [170]. Crucially, SID preserves LVLMs' general abilities, as evidenced by strong MME and MMBench scores. By rebalancing modality priors without compromising functionality, SID advances *modality-level* robustness, ensuring trustworthy outputs in open-world deployment.

In summary, these contributions form a cohesive framework for robust multimodal learning: ProCC ensures generalization to unseen compositions, C^2KD safeguards against missing modalities, and SID mitigates modality priors. This triad equips multimodal learning systems to operate reliably in dynamic, real-world environments where compositions are novel, inputs are incomplete, and modality dominance varies, marking a significant stride toward trustworthy, open-world multimodal learning artificial intelligence.

1.5 Thesis Organization

Thesis Organization Summary This thesis is structured into six chapters that systematically address the critical challenges of robust multimodal learning in open-world environments, progressing from problem formulation to methodological innovation and validation:

Chapter 1 establishes the research foundation, delineating the core challenges of compositional generalization robustness, modality missing robustness, and hallucination mitigation in multimodal models. It outlines the overarching research framework, key contributions, and thesis roadmap.

Chapter 2 (Background) provides a comprehensive review of foundational concepts, including: composition problem formulation, Compositional Zero-Shot Learning (CZSL), and its open-world limitations; unimodal knowledge distillation and Cross-modal knowledge distillation (CMKD); multimodal large language models, decoding strategy in LLMs, and hallucinations issues in Large Vision-Language Models (LVLMs).

Chapter 3 (ProCC: Progressive Cross-Primitive Compatibility) introduces a novel framework to enhance compositional generalization robustness. Firstly, in Chapter 3.1, detailed challenges and motivations are illustrated. Then, we propose the Progressive Cross-Primitive Compatibility network aligns visual and semantic primitives through curriculum learning in Chapter 3.2. Chapter 3.3 validates ProCC in three widely-use datasets including UT-Zappos, MIT-States, and C-GQA under various settings. Detailed ablation studies confirm the relations of each modules of ProCC.

Chapter 4 (C^2KD: Customized Cross-modal Knowledge Distillation) tackles modality missing inference via a customized crossmodal distillation framework. Firstly, in Chapter 4.1, detailed challenges of knowledge distillation across modalities and motivations to develop crossmodal knowledge distillation methods are illustrated. Then, in Chpater 4.2, we comprehensively revisit traditional knowledge distillation effectiveness in cross-modal scenario. We follow this up with a solution named Customized Cross-modal Knowledge Distillation (C^2KD) in Chapter 4.3. Extensive experiments of audio-visual, image-text, and RGB-depth modalities in terms of classification and segmentation tasks are performed in Chapter 4.4, Experimental results of C^2KD significantly outperform existing KD methods, demonstrating robustness in transferring knowledge even from low- to high-accuracy modalities while mitigating training instability and performance degradation caused by modality gaps. Ablation and sensitivity analysis as well as discussion are in Chapter 4.5 and 4.6, respectively.

Chapter 5 (SID: Self-Introspective Decoding) addresses hallucination robustness in LVLMs through token-level adaptive pruning to amplify the fine-grained hallucinations then contrastively to alleviate the hallucinations. Firstly, Chapter 5.1 shows the detailed challenges of hallucination issues of LVLMs and analysis of previous contrastive decoding strategies. We then show the paradigm of LVLMs generation and comprehensively analyze contrastive decoding in LVLMs in Chapter 5.2. In Chapter 5.3, we propose the Self-Introspective Decoding (SID) strategy to dynamically suppress the priors of LLM that conflict visual evidence, achieving a significant reduction in hallucinations. Chapter 5.4 illustrates the detailed experimental results on CHAIR, POPE, MME, MMBench, GPT-4 assisted benchmark. and GPT4-V assisted evaluation to validate that SID effectively alleviate the hallucination issues while preserving general abilities of LVLMs.

Chapter 6 (Conclusion and Future Work) synthesizes the thesis contributions in

Chapter 6.1 and proposes future research directions in Chapter 6.2, including multimodal test-time adaptation (i.e., Chapter 6.2.1), task-aware adaptation of multimodal LLMs (i.e., Chapter 6.2.2), and developing multimodal agent as experts (i.e., Chapter 6.2.3).

Chapter 2

Background

Following the research framework we presented in Figure 1.1 of Chapter 1, we will discuss previous and contemporary methodologies for building robust multimodal learning systems, including background of composition generalization robustness in Section 2.1, unimodal and crossmodal knowledge distillation background in Section 2.2, and hallucination-related background in Section 2.3.

2.1 Composition Generalization Robustness

2.1.1 Problem Formulation

Compositional Zero-Shot Learning (CZSL) aims to recognize the composition of two primitives, i.e., an state ($e.g., tiny$) and an object (e.g., *dog*). Given S and O as two sets of states and objects, spanning all classes, we compose a set of possible state-object pairs, i.e., $C = S \times O = \{(s, o) \, | s \in S, o \in O\}$. Formally, given a training set $D^s = \{(i, c) | i \in I^s, c \in C^s\}$, where I^s is an training image set, and C^s is the corresponding state-object labels. The close world CZSL follows the generalized ZSL [175] that the test sample comes from either seen (C^s) or unseen (C^u) composition ($C^s \cup C^u$). For the **Open-World CZSL (OW-CZSL) setting** [124], there assumes no prior on the set of testing compositions. It means the model must consider the full compositional space (C), which is much larger than $C^s \cup C^u$. Consequently, the unseen compositions are $C^u_{ow} = C \backslash C^s$. OW-CZSL introduces a more practical setting while bringing more challenging problems: 1) It is hard to generalize from small seen compositions to large unseen compositions. 2) There are a large number of less feasible compositions in the full composition space (C), confusing the prediction models. [79] recently proposes a new practical setting, i.e., only training with one of the state and object annotations, named **partial-supervision CZSL (pCZSL)**. Formally, for the training set C^s, The relation of the partial label of state and object primitives can be formulated as: $\{(s, u)\} \cup \{(u, o)\} = C^s$, where u indicates unlabeled primitives. Consequently, the test set in pCZSL has the full output composition space (C) like OW-CZSL, while the training set in pCZSL does not have the composition knowledge about any state-object pairs.

2.1.2 Composition Zero Shot Learning

Different from typical zero-shot learning [175, 73, 104], which aims to utilize attributed vectors or inherent semantic descriptions to recognize unseen instances, **Compositional Zero-shot Learning** (**CZSL**) aims to recognize the state and object primitive (or modality) from the images, and even the state-object compositions are not ever seen in the training datasets. Unlike humans, which can extrapolate new concepts from previously learned knowledge, For instance, if the people are taught what the fried chip and toasted bread are, most of them can recognize the fried bread immediately, neural networks lack the compositional generalization ability. The main challenge of CZSL is modeling the relation and affordance of states and objects modalities, generalizing this capability to unseen compositions. Existing methods mainly deal with CZSL in two ways. The first way is inspired by Biederman's Recognition-ByComponents theory [11] and Hoffman's part theory [56]. For instance, Misra et al. [129] learn a transformation between individual classifiers of states and objects. Other representative methods learn hierarchical decomposition and composition of the state and object primitives [185, 51, 60], model objects to be symmetric under attribute transformations [107], and learn independent prototypical representations of visual primitives then propagated prototype via a compositional graph [146]. The second way tries to learn the joint representation of the state-object compositions from given images. Specially, SymNet [107] enforces symmetries in the representation of objects given their state transformations. Graph network is also employed in [130] to enforce the compositional information transfer from seen to unseen compositions. AoP [131] regards attribute as the operator and models each state as a linear transformation of objects. CANet [167] learns conditional attributes to enhance embedding space. LAP [82] exploits the self-attention mechanism to embed related compositions closer and unrelated far away. Differently, causality-based methods [5, 186] explore decomposable objects and state representations.

Above methods perform well on the close-world CZSL, while suffering from severe degradation for the open-world setting [124, 125, 79], where the output space has not imposed any limit. Mancini et al. [124] compute feasibility scores (i.e., cosine similarity) between visual features and compositional embeddings to reduce the output space. Then they further inject the feasibility scores both at the loss level and within the graph connections [125]. Karthik et al. [79] follows the Visual Product [129] and predicts state and object primitives independently with non-linear feature extractors. To refine the relation between independent primitives, Conceptnet [155] is introduced as the external knowledge. We revisit the Visual Product and achieve cross-primitive compatibility in an easy-hard learning manner, avoiding the external knowledge in [79] and cumbersome word embeddings in [124, 125].

2.2 Knowledge Distillation over Cross Modality

2.2.1 Unimodal Knowledge Distillation

Unimodal Knowledge Distillation (KD) transfers the knowledge of a pretrained teacher to a student by minimizing the discrepancies between output logits or intermediate features between student and teacher. Previous KD methods primarily concentrate on inheriting knowledge from the large-capacity teacher. Pioneering work [54] regularizes Kullback–Leibler (KL) divergence between student and teacher soft labels. CRD [159] develops contrastive-based objectives for knowledge transferring. SCKD [212] automatically adjusts the KD process according to the distillation gradient similarity. Yang et al. [183] utilize the teacher's pre-trained classifier to regularize the student's penultimate layer feature. Zhu et al. [211] identify and discard the undistillable classes from the large teacher model based on the validation set. DKD [204] decouple KD into target class and non-target class knowledge distillation to balance learning effectiveness and flexibility. Review [23] proposes the review mechanism to utilize knowledge of teacher's multi-level features. RKD [134] regularizes the student with distance-wise and angle-wise structural relations to replace KL loss. DIST [62] further proposes a novel correlation-based loss to capture the inter-class and intra-class relations. L2D [187] extends relation-based distillation into multi-label classification. These KD methods focus on unimodal KD and learn to inherit knowledge from a fixed teacher. However, for CMKD, the modality gap impedes knowledge transfers across modalities. We *argue* that teacher modality should be optimized with feedback supervision of student modality to produce *receptive knowledge*. Previous online knowledge distillation methods [203, 90, 28, 101] update teacher model to adapt student in unimodal scenarios. Specifically, DML [203] simply applies KD losses mutually, treating each other as teachers. ONE [90] further exploits gated ensemble logits of multiple training networks. AFD [28] proposes online feature alignments via adversarial training. The recently proposed SHAKE [101] bridges offline and online KD by transferring

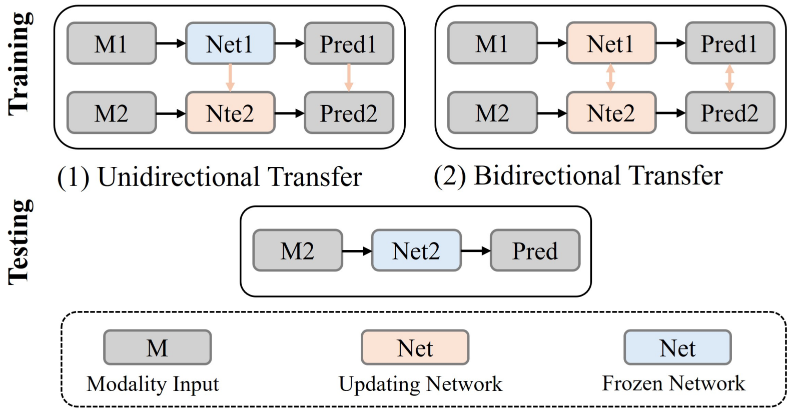


Figure 2.1: **Intuitive presentation** of cross-modal knowledge distillation.

knowledge through extra shadow heads.

2.2.2 Cross-modal Knowledge Distillation

With the rising prevalence of machine learning [67, 68, 65, 70, 119, 95, 118, 49] and mulimodal sensors [71, 72, 69, 66, 36], traditional KD methods have been extended to achieve knowledge transfer across multimodal data, thereby enhancing downstream tasks [31, 50, 143, 180, 202, 93, 198]. Figure 2.1 illustrates the protocols for cross-modal knowledge distillation, where multiple modalities are utilized and cross-modal knowledge is transferred during the training phase. In the inference phase, only one modality is available, but it benefits from the inherited multimodal knowledge acquired during CMKD, which is critical for the modality missing situations. *However*, previous methods typically utilize *high-accuracy* or *well-labeled* modality as the teacher to transfer knowledge to low-accuracy or unlabeled modality [31]. For example, [50] leverage a large labeled modality as the supervisory signal for a new unlabeled paired modality. [143] transfers knowledge among the missing and available modalities via GANs. [180] adapts a multimodal network to the unlabeled modality by inheriting knowledge from the well-trained unimodal teacher. [93] proposes a decomposed cross-modal distillation method to enhance RGB-based detector by transferring knowledge of the optical flow modality. [198] distills ImageNet pre-trained visual modality to audio modality for indoor dense prediction. Recently, Xue et al. [179] first perform an in-depth investigation on CMKD and propose the modality focusing hypothesis (MFH), suggesting that modality-general decisive features are crucial determinants of CMKD efficacy. [179] contributes to MFH but doesn't develop unified solutions. In this chapter, we further quantitatively analyze the challenges of CMKD (the modality gap, i.e., *modality imbalance* and *soft label misalignment*) and propose effective solutions to address these issues.

2.3 Multimodal Hallucination

2.3.1 Multimodal Large Language Models

Motivated by the success of Large Language Models (LLMs) [160, 7, 26, 157, 161, 126], recent studies have extended LLMs to multimodal regions and provided Large Vision-Language Models (LVLMs) [112, 209, 19, 189, 97, 8, 99, 29, 111, 10, 25, 96] powered by pre-trained LLMs. LVLMs understand and generate diverse content in a more comprehensive way by integrating user instruction and vision inputs. LLaVA [112] connects open-set vision encoder with LLMs (i.e., Vicuna [26]) by instruction tuning with elaborated language-image instruction-following data. Then, LLaVA-1.5 [111] develops the vision-language connector that is data-efficient and powerful for better multimodal understanding. Shikra [19] further incorporates grounding data and trains the model to understand the grounding knowledge in the given images. BLIP-2, InstructBLIP, and MiniGPT-4 [99, 29, 209] introduce a learnable querying transformer to fusion multimodal features and largely reduce image tokens. Fuyu [10] proposes a vanilla decoder-only architecture without the vision encoder and adapter that makes it easier to understand, scale, and deploy. InternVL [25] proposes three simple but effective improvements, including a strong vision encoder, dynamic high-resolution, and high-quality bilingual dataset. Recently, built on SOTA open-source LLaMA 3 [126] and increasing the input vision resolution to 4× more pixels, LLaVA-NeXT [96] exhibits excellent multimodal capabilities. Despite the impressive results, all of the above LVLMs suffer from serious hallucination problems, and we mainly conduct experiments on advanced LVLMs, including InstructBLIP, Shikra, LLaVA-1.5, and LLaVA-NeXT.

2.3.2 Decoding Strategy in LLMs

Selecting decoding strategies in language models is crucial, as it determines how models generate text. Greedy decoding selects the highest probability next token at each step but might lead to less varied text. Beam search [45] is an accumulated-score-based decoding strategy. It maintains a set of beams to enlarge the candidate range and finally selects the best one in beams, which is more sophisticated than greedy decoding. Sampling decoding generates the next words by randomly selecting from the output distribution, where Top-k sampling [35] samples from Top-k likely tokens [35] and brings diversity but sometimes induces less coherent outputs. Top-p (Nucleus) sampling [57] improves Top-k sampling that considers the dynamic number of words that reach the probability p, achieving a balance between randomness and relevance. Recently, to alleviate the hallucination issue, DoLa [27] decoding emphasizes the knowledge of mature layers and downplays that of pre-mature layers. OPERA [61] is established on beam-search decoding strategy and finds the interesting phenomenon of high-probability co-occurrence between the hallucination and the knowledge aggregation patterns. OPERA penalizes ‘Over-Trust Logit’ in the beam score to alleviate aggregation patterns. In this thesis, we aim to contribute the decoding strategy that can be seamlessly integrated into different decoding strategies to mitigate multimodal hallucinations without sacrificing text generation quality, such as diversity, coherence, and relevance.

2.3.3 Hallucination in Foundation Models

Hallucination, defined as the generation of irrelevant, factually incorrect, or meaningless text in a given context [144, 200, 47, 174], is a significant bottleneck in current foundation models. This issue can stem from overfitting specific patterns in the training data, a lack of understanding world knowledge, or an inability to effectively contextualize a given input [75]. In the context of LLMs, hallucinations often manifest as generated content that conflicts with world knowledge or common sense. For LVLMs, the primary concern is whether the generated answer conflicts with the provided images. To mitigate the hallucination issue, several solutions have been proposed, including **robust instruction tuning with curated datasets** [92, 48, 110, 206, 76, 195, 193, 121, 197], **post-hoc utilizing auxiliary analysis networks** [123, 208, 190, 24, 173, 39], and **various decoding strategies** [103, 27, 113, 94, 38, 170, 84, 210]. However, robust instruction tuning requires massive high-quality datasets and advanced GPU clusters, making it resource-intensive; Post-hoc utilizing auxiliary networks heavily rely on the auxiliary network, leading to high inference costs. As for decoding strategies, representative LVLMs hallucination alleviation methods [94, 38, 170] manually disturb raw inputs to induce hallucinations then contrast them to alleviate the issue. However, holistic disturbing raw inputs might bring additional noise during contrastive decoding, and double the inference cost. In this thesis, we propose an efficient Self-Introspective Decoding (SID) that induces and then mitigates vision-and-text association hallucination by token-level disturbances, greatly reducing the inference cost.

Chapter 3

ProCC: Progressive Cross-Primitive Compatibility for Composition Generalization

3.1 Challenges and Motivations

Current neural networks lack the compositional generalization robustness inherent to human cognition. Specifically, the training set contains images with corresponding multimodal descriptions (primitives), namely states and objects. Since objects and states are semantically entangled, that is, objects in different states often have different appearances, and states also vary greatly depending on the object, the model needs to be based on known primitives. Identify unseen combinations. The main challenge behind CZSL is how to model interactions between state and object primitives and extrapolate seen combinations to unseen combinations. Existing methods mainly focus on learning a shared embedding space of object-state combinations [107, 130, 131, 82] or compositional attribute and object classifiers [140, 129, 105, 178, 186].

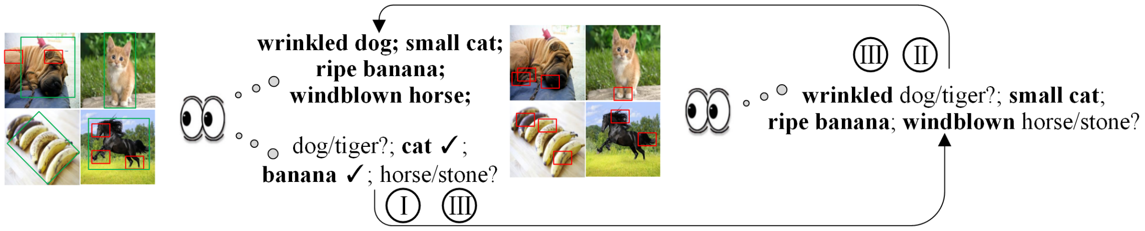


Figure 3.1: **The overall concept of our method.** Following the principle of 'forest before trees' [55], human feedforward hierarchy underlies implicit processing for initial vision at a glance (i.e., green rectangle), and feedback connections add details to explicit vision with scrutiny (i.e., red rectangle). As for composition generalization learning, humans first (①) learn to recognize overall objects, then (②) gradually identify the scrutiny attribute of objects, i.e., state, and finally (③) reasonably compose the object and state primitives. Inspired by this, we aim to progressively recognize the object and state primitives and guide the network to exploit discriminative information conditioned on learned knowledge via the CPC module.

However, the performances of these methods degrade to some extent [124, 125] as for the open-world setting (OW-CZSL), where there are no priors on the unseen compositions, and the model must consider the whole possible compositions in terms of all objects and states. To deal with such a problem, existing mainstream methods utilize feasibility constraints on the composition embedding [124, 125] or independently predict simple state and object primitives [78, 79]. While [124, 125] rely on different word embedding methods. The straightforward but effective Visual Product method like [79] predicts the state and object primitives while ignoring the compatibility between two primitives. So external knowledge is introduced to eliminate less feasible compositions, while it is cumbersome to select proper external knowledge for varying datasets.

To address the aforementioned problems, we propose Progressive Cross-primitive Compatibility (ProCC) network to recognize compositions in the open-world setting and a more realistic setting (i.e., partial supervision), aiming at attaining cross-primitive compatibility during easy-hard recognition progress, as shown in Figure 3.1. Specifically, following the route of the human learning process [55], we **first** learn to classify objects, which is easier than recognizing states [147, 79] because the same state varies greatly conditioned on objects and related contexts, i.e., ancient castle / ancient coin, and different states are sometimes less feasible composed with the same object, i.e., old dog / ripe dog. **Then**, with the learned knowledge of object primitive, we sequentially classify state primitives conditioned on object features via Cross-Primitive Compatibility (CPC) module, excavating discriminative information. **Finally**, we finetune the whole network conditioned on prior knowledge of two primitives. The ProCC achieves cross-primitive compatibility by adjusting the visual attention to filter out less feasible compositions, without the aid of external knowledge like Word2vec [127], Glove [138], Conceptnet [155] etc. Also, the progressive training paradigm effectively models the interactions of primitives via conditioned features, especially for pCZSL, where only partial label results in invalid interactions.

In summary, this chapter's contributions are four-fold:

1) We propose a novel Progressive Cross-primitive Compatibility (ProCC) network, mimicking the human learning progress of recognizing the state and object compositions without external knowledge.

2) We revisit Visual Product methods and present a Cross-Primitive Compatibility (CPC) module to model the interactions of classifiers to exploit the discriminative visual attention conditioned on each other, guiding the model to generalize to feasible compositions.

3) The progressive training paradigm alleviates the invalid cross-primitive interactions without the aid of cumbersome external knowledge, especially for pCZSL.

4) Comprehensive experimental results on three large-scale datasets for OW-CZSL and pCZSL tasks demonstrate the effectiveness of our proposed approach, which outperforms the state-of-the-art methods.

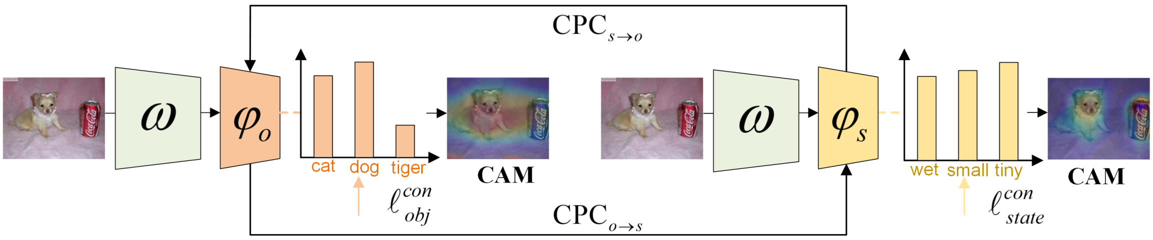


Figure 3.2: **The framework of ProCC.** Features from the encoder (ω) are respectively fed to the object and state (φ_o and φ_s) classifiers, where the Cross-Primitive Compatibility (CPC) aims to model the cross-primitive interactions. Progressive learning strategy is proposed to gradually modulate primitive compatibility, especially for pCZSL. For detailed training procedure, please refers to **Algorithm 2**. Class Activation Maps (CAM) of input samples are illustrated to show visual attention.

3.2 Progressive Cross-primitive Compatibility (ProCC)

Most CZSL methods [5, 107, 131, 140, 147, 124, 125, 130] explicitly modulate the interactions of states and objects to improve the generalization ability. However, it is less effective for OW-CZSL and pCZSL due to large output space and missing labels. Some methods [78, 79] follow the Visual Product [129] that independently predict the state and object primitives, disregarding compositional nature. Following the route of [78, 79, 129], we propose Progressive Cross-primitive Compatibility (ProCC) network while achieving cross-primitive compatibility. Also, like the human learning process [55], ProCC trains the network in an easy-hard manner, which dynamically models interactions between state and object primitives, alleviating the negative influence of no explicit supervision on both states and objects in pCZSL. Figure 3.2 shows the framework of the proposed approach. In the following subsections, we revisit the Visual Product and introduce a cross-primitive compatibility module and progressive learning strategy.

3.2.1 Revisit Visual Product

Generally, given an image i, CZSL wants to model the joint probability distribution $p(s_i, o_i|i)$. The visual product simplifies this as follows:

$$p(s_i, o_i|i) \approx p(s_i|i) \times p(o_i|i) \tag{3.1}$$

In this way, Visual Product treats the states and objects independently only from the visual cues, without side information (i.e., word embeddings). Concretely, input image i is firstly encoded to obtain the feature z as: $z = \omega(i)$. Then the object (i.e., $\varphi_o \langle z, o\rangle$) and state (i.e., $\varphi_s \langle z, s\rangle$) classifiers assign z to the vectors in the probability simplex o and s, spanning all object and state classes. Visual Product minimizes the cross-entropy loss of seen compositions ($D^s = \{I^s, C^s\}$) for both object and state predictions:

$$\ell_{vp} = \ell_{obj}(i, o_i) + \ell_{state}(i, s_i) \tag{3.2}$$

$$\ell_{obj} = \min_{\varphi_o} \sum \ell_{ce}(\varphi_o \langle \omega(i), o\rangle, o_i) \tag{3.3}$$

$$\ell_{state} = \min_{\varphi_s} \sum \ell_{ce}(\varphi_s \langle \omega(i), s\rangle, s_i) \tag{3.4}$$

where $(i, (s_i, o_i)) \in D^s$. Thus, the prediction function is:

$$f(i) = \arg\max_{(s,o)\in C} \varphi_s \langle \omega(i), s\rangle \times \varphi_o \langle \omega(i), o\rangle \tag{3.5}$$

where C represents the full state-object composition pairs in OW-CZSL. As the search space is huge, Visual Product is more effective than previous methods, which aim to produce discriminative state-object embeddings [78, 79]. Recently, [78, 79] expanded the visual product and equipped the classifiers with multi-layer perceptrons (MLP) to excavate discriminative features. Also, external knowledge [155] is employed in [79] to estimate the feasibility scores of compositions. Here, we explicitly model the composition interactions via Cross-Primitive Compatibility (CPC) module during the training procedure, without external knowledge. Also, considering the pCZSL setting and better modulating the primitive compatibility, the progressive learning strategy,

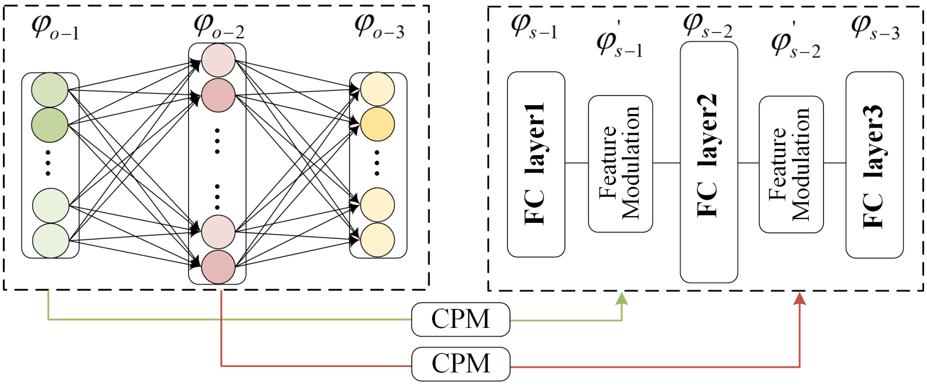


Figure 3.3: **The detailed framework of the object-state Cross-Primitive Compatibility (CPC$_{o\rightarrow s}$).** Features from the object classifier (φ_{o-1} and φ_{o-2}) are encoded by learnable Cross-Primitive Memory (CPM) units. Then respectively interact with state features (φ_{s-1} and φ_{s-2}) to achieve compatibility of state features conditioned on objects.

following the human learning process [55], is proposed to facilitate cross-primitive compatibility in an easy-hard manner.

3.2.2 Cross-primitive Compatibility Module

Visual Product methods independently predict compositions via Equation 1, which ignores the fact that the feasibility of state-object compositions is heavily conditioned on each other. A more practical compositional probability can be modeled as:

$$p(s_i, o_i|i) \approx p(s_i|i, f_o(i))) \times p(o_i|i, f_s(i))) \tag{3.6}$$

where $f_o(i)$ and $f_s(i)$ are intermediate features of the object and state primitives. It is non-trivial to directly model the relationship between objects and states due to the diverse semantic entanglement and a large number of possible compositions. We integrate the feasibility reasoning into the trainable Cross-Primitive Compatibility (CPC) module, which facilitates interactions between two classifiers to explore informative visual attention conditioned on feature representations of each primitive.

Specifically, the features extracted by the encoder (ω) are fed to primitive classifiers (i.e., φ_o and φ_s). The primitive classifiers follow the Visual Product methods [78, 79] that consist of multi-layer perceptron (MLP), specifically three-layer MLP, for classifications. As shown in Figure 3.2 and Equation 6, the network is symmetric and we take the object-state CPC ($\mathrm{CPC}_{\mathrm{o \to s}}$) module for example, as shown in Figure 3.3, intermediate features (i.e., output distributions) from φ_{o-1} and φ_{o-2} are fed to φ_s to interact with state features. However, direct modulation state features will induce information degradation because of the huge task diversity. We propose learnable Cross-Primitive Memory (CPM) units for soft interactions. Specifically, the learnable CPM unit introduces conditioned information to modulate corresponding features along with the residual connection, which is formulated as follows:

$$\varphi^{m}_{o-l} = \sigma\left(\mathrm{Conv}^{k}_{\mathrm{1d}}\left(\varphi_{o-l}\right)\right), l \in (1, 2) \tag{3.7}$$

$$\varphi'_{s-l} = \varphi_{s-l} \times \varphi^{m}_{o-l} + \varphi_{s-l}, l \in (1, 2) \tag{3.8}$$

where $\mathrm{Conv}^{k}_{\mathrm{1d}}$ and σ represent the 1d convolution layer and softmax activation function. Kernel size (k) is equal to 1/10 feature dimension to efficiently capture the long-range dependency. Then the enhanced state features are fed to the next layer of φ_s as:

$$\varphi_{s-(l+1)} = f_{s-l}(W^{T}_{s-l}\varphi'_{s-l} + b_{s-l}), l \in (1, 2), \tag{3.9}$$

where W and b are weights and biases of MLP. Accordingly, the conditioned cross-primitive interactions are injected into each other, reducing less feasible primitive predictions. Therefore, Equations 3 and 4 can be re-write as:

$$\ell^{con}_{obj} = \min_{\varphi_o, \varphi_{o \to s}} \sum \ell_{ce}(\varphi_o \left\langle z | \varphi_{s \to o}(\varphi_s(z)), o \right\rangle, o_i) \tag{3.10}$$

$$\ell^{con}_{state} = \min_{\varphi_s, \varphi_{s \to o}} \sum \ell_{ce}(\varphi_s \left\langle z | \varphi_{o \to s}(\varphi_o(z)), s \right\rangle, s_i) \tag{3.11}$$

where $z = \omega(i)$, $(i, (s_i, o_i)) \in D^s$, and $\ell^{con}_{vp} = \ell^{con}_{obj} + \ell^{con}_{state}$

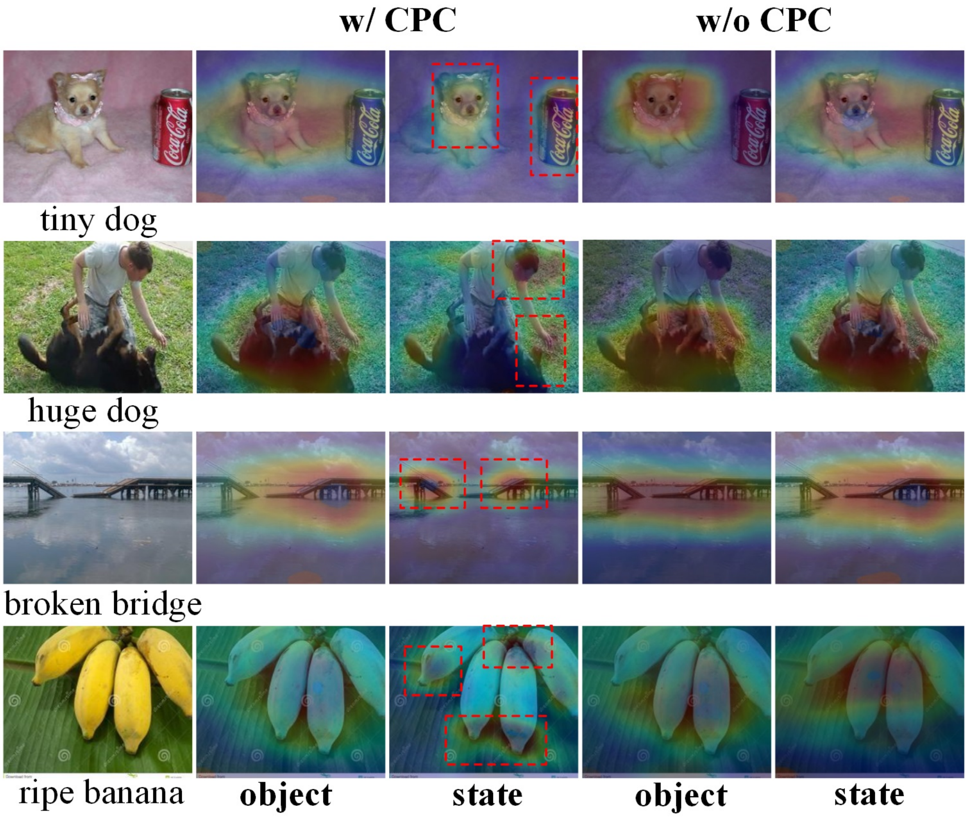


Figure 3.4: **Visualizations of class activation maps** of ProCC with and without CPC modules on the testing dataset of MIT-States. The discriminative regions are marked with red rectangles.

3.2.3 Visual Explanation

To further illustrate and explain the effect of the CPC module, we visualize the attention learned from the classifier via Class Activation Map (CAM) [207] in Figure 3.4. The standard CAM is formulated as:

$$\mathrm{CAM}_c(x, y) = \sum_k \omega_k^c f_k(x, y) \tag{3.12}$$

where CAM_c means the class activation map that leads to the classification of an image to class c. $f_k(x, y)$ and ω_k^c stand for the activation of unit k in the last layer at spatial location (x, y) and the weight corresponding to class c for unit k. Here,

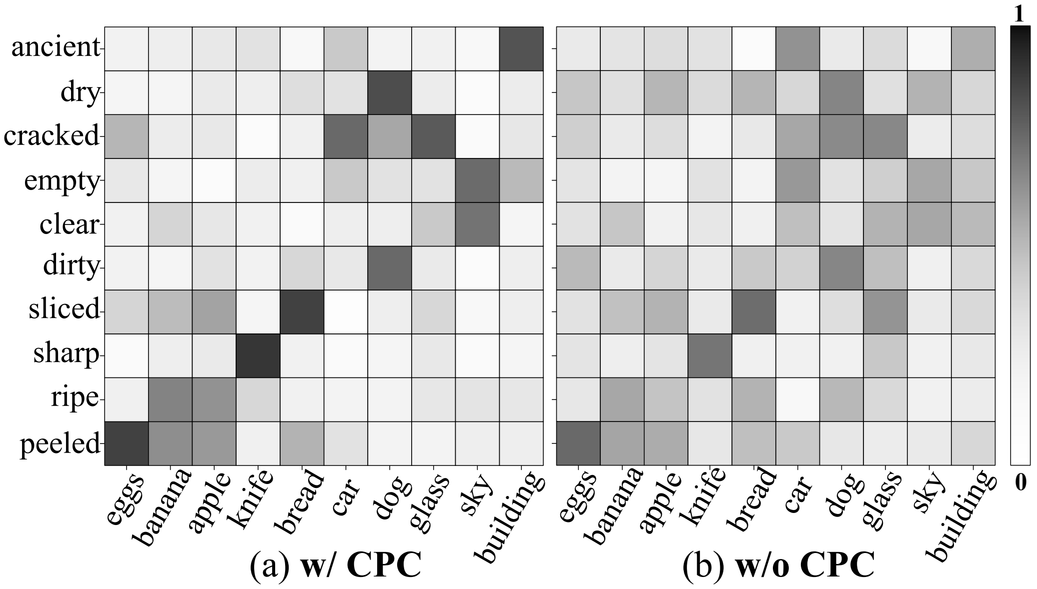
ω_k^c is the final layer of the MLP (i.e., φ_{o-3} and φ_{s-3}), which has been modulated by the CPC modules. Figure 3.4 shows some visualization examples with (w/) and without (w/o) CPC module. As the encoder (ω) is pre-trained for the object classification task, most CAMs for the object classifier can locate and recognize the proper attention regions. However, the CAMs for the state classifier vary greatly as state primitives are conditioned on the object primitive and related contexts. For the *tiny dog* and *huge dog* compositions, the CPC module drives the model to focus on the discriminative regions that a dog with a small head compared with other objects tends to classify to the *tiny* otherwise classify to *huge*. For more abstract compositions, *broken bridge* and *ripe banana* compositions, the state primitives heavily depend on the object primitives otherwise may induce less feasibility compositions. The state of *broken* is mainly reflected in the curvatures of the bridge and the *ripe* primitive of the banana displays the black spots on the surface. Overall, the CPC module enables the efficient adjustment of visual attention conditioned on mutual relations. Moreover, Figure 3.5 illustrates the confusion matrices about state and object primitives. Concretely, we select ten typical state and object primitives in the MIT-States [74] dataset. Prediction probabilities of states are accumulated then normalized with and without CPC module to formulate the confusion matrices. We can learn that the CPC module facilitates reasoning compatible compositions with high confidence.

3.2.4 Progressive Learning Strategy

However, jointly training the state and object classifiers may induce two issues: **(1)** When it comes to the more practical setting, partial supervision Compositional Zero-Shot Learning (pCZSL), where only the partial label, not both, is available [79]. The missing label makes the joint training strategy invalid to model the interactions between the state and object primitives. A naive way of learning from such partial supervision is to update the parameters of the state and object classifier only based on the available labels, which lacks the interaction information across primitives via

Algorithm 1: Training procedure of ProCC.

```
Input: Training data $D^s = \{(i, c) | i \in I^s, c \in C^s\}$, pre-trained $\omega$, learning rate
       $\lambda_1$, $\lambda_2$, $\lambda_3$
Output: Optimal $\varphi_o$, $\varphi_s$, CPC: $\varphi_{o \to s}$, $\varphi_{s \to o}$
Initialize: $\varphi_o$, $\varphi_s$, $\varphi_{o \to s}$, $\varphi_{s \to o}$;
Stage 1: // train $\varphi_o$
while not converged do
    Sample a batch from $D^s$ as images $(i_k)_{k=1}^n$ with their object labels $(o_k)_{k=1}^n$ ;
    for samples in the batch do
        Compute $\ell_{obj}$ via Equation 4.3.;
        Update $\varphi_o \leftarrow \varphi_o - \lambda_1 \nabla_{\varphi_o} \ell_{obj}$
Stage 2: // train $\varphi_s$ and $\varphi_{o \to s}$
while not converged do
    Sample a batch from $D^s$ as images $(i_k)_{k=1}^n$ with their state labels $(s_k)_{k=1}^n$ ;
    for samples in the batch do
        Compute $\ell_{state}^{con}$ via Equation 4.11.;
        Update $\varphi_{s \cup o \to s} \leftarrow \varphi_{s \cup o \to s} - \lambda_2 \nabla_{\varphi_{s \cup o \to s}} \ell_{state}^{con}$
Stage 3: // finetune $\varphi_o$, $\varphi_s$, $\varphi_{o \to s}$, and $\varphi_{s \to o}$
while not converged do
    Sample a batch from $D^s$ as images $(i_k)_{k=1}^n$ with their object and state labels
        $(o_k, s_k)_{k=1}^n$ ;
    for samples in the batch do
        Compute $\ell_{vp}^{con}$ via Equations 4.10 and 4.11.;
        Update $\varphi_{total} \leftarrow \varphi_{total} - \lambda_3 \nabla_{\varphi_{total}} \ell_{vp}^{con}$
```



(a) **w/ CPC** (b) **w/o CPC**

Figure 3.5: **Confusion matrices** about prediction probabilities of states conditioned on objects (w/ CPC) or not (w/o CPC).

the CPC module. Recent method [79] estimates the missing labels via pseudo-labeling [91] as well as utilizes the external knowledge [155]. The challenge of missing labels also exists in the standard Multi-Task Learning (MTL) that the traditional updating rule will give inferior results due to the missing annotations [162, 83, 133, 102]. Some typical solutions propose hard knowledge distillation [83], alternative optimization strategy [133], and learning in the joint pairwise task spaces [102]. However, compared with the MTL task, the missing label issue matters more to the CZSL task, as the object and state primitives are heavily tangled. **(2)** Also, jointly training results in sub-optimal interactions as the diverse difficulty of object and state predictions. Concretely, classifying states is more challenging than objects [147, 79]. Therefore, joint training inevitably induces noisy conditioned information, which hinders to reason cross-primitive compatibility.

To enable the full interaction of state and object primitives, we propose a progressive learning strategy, mimicking the easy-hard learning process shown in Figure 3.1. Concretely, with the features from the encoder (ω), we first train the object classifier

φ_o with given labels (Equation 3.3), to obtain object features (φ_{o-l}, $x \in (1,2)$). Then we sequentially train the state classifier φ_s and CPC$_{o\to s}$ ($\varphi_{o\to s}$) conditioned on pre-trained object features (φ_{o-l}) (Equation 3.11), to interact to adjust the visual attention. Finally, we fine-tune the state and object classifiers (φ_s and φ_o) as well as CPC modules ($\varphi_{o\to s}$ and $\varphi_{s\to o}$) conditioned on the well-trained features (Equations 3.10 and 3.11). We utilize this training protocol both in the OW-CZSL and pCZSL settings. During the easy-hard recognition progress, our method alleviates invalid interactions of cross primitives, especially in the pCZSL setting, without external knowledge. For detailed training procedure, please refers to **Algorithm 2**.

Method	**C-GQA**						**MIT-States**						**UT-Zappos**					
	Val		Test				Val		Test				Val		Test			
	HM	AUC	*S*	*U*	HM	AUC	HM	AUC	*S*	*U*	HM	AUC	HM	AUC	*S*	*U*	HM	AUC
TMN	NA	NA	NA	NA	NA	NA	2.1	0.2	12.6	0.9	1.2	0.1	21.2	9.2	55.9	18.1	21.7	8.4
AoP	NA	NA	NA	NA	NA	NA	3.2	0.3	16.6	5.7	4.7	0.7	23.4	10.1	50.9	34.2	29.4	13.7
LE+	9.3	1.8	19.2	0.7	1.0	0.08	5.3	0.5	14.2	2.5	2.7	0.3	26.6	14.3	60.4	36.5	30.5	16.3
VisProd	10.5	2.0	24.8	1.7	2.8	0.33	7.2	1.0	20.9	5.8	5.6	0.7	28.8	15.4	54.6	42.8	36.9	19.7
SymNet	12.3	2.5	26.7	2.2	3.3	0.43	8.0	1.2	21.4	7.0	5.8	0.8	32.5	16.7	53.3	44.6	34.5	18.5
CGE	12.8	2.8	28.3	1.3	2.2	0.30	8.3	1.8	**29.6**	4.0	4.9	0.7	34.5	18.9	58.8	46.5	38.0	21.5
CompCos	12.0	2.4	28.4	1.8	2.8	0.39	8.4	1.5	25.4	10.0	8.9	1.6	32.5	18.1	59.3	46.8	36.9	21.3
Co-CGE	12.3	2.7	28.7	1.6	2.6	0.37	8.4	**2.1**	26.4	10.4	**10.1**	**2.0**	34.8	19.2	60.1	44.3	38.1	21.3
KGSP	13.2	2.9	26.6	2.1	3.4	0.44	7.9	1.4	23.4	7.0	6.7	1.0	33.2	19.8	58.0	47.2	39.1	22.9
CANet	14.3	2.8	27.3	1.9	3.2	0.39	8.3	1.7	25.3	6.7	6.6	1.2	35.1	19.8	58.7	46.0	38.7	22.1
Ours	**16.1**	**4.0**	**29.0**	**2.6**	**3.8**	**0.54**	**8.6**	1.9	27.6	**10.6**	7.8	1.6	**36.5**	**22.4**	**62.2**	**48.0**	**39.9**	**23.6**

Table 3.1: **Quantitative comparisons in the OW-CZSL setting.** We report the best seen (S), best unseen (U) accuracy, HM, AUC on the test and validation sub-datasets. The best and second-best results are bold and underlined.

3.3 Experimental Results

3.3.1 Datasets and Evaluation Metrics

We conduct experiments on three widely-use datasets including UT-Zappos [192], MIT-States [74], and C-GQA [129]. UT-Zappos is a dataset for the shoes and has 50025 images. It contains 12 object classes and 16 state classes, with 83 seen compositions and a total of 192 compositional spaces. MIT-States has 53753 images with 115 state classes and 245 object classes. The seen and all output compositions are 1,262 and 28,175, respectively. C-GQA is the largest dataset that contains 186,577 images with 413 state classes and 674 object classes. It contains 5,592 seen compositions and a full output space of 278,362 compositions, which makes it the most extensive for the OW-CZSL. For the **OW-CZSL**, we follow the splits of [124, 125, 79] and evaluate based on the generalized settings, where the test samples are from both seen and unseen compositions. Considering the performance of the model with different bias factors for the unseen compositions, we vary the bias on the seen composition (C^s) during the test phase and report the performance as best seen (S), best unseen (U), best harmonic mean (HM), and the Area Under the Curve (AUC). For the **pCZSL**, following [79], we remove the label and calculate the metrics on the full output composition space (C). As we can not access the full-labeled seen compositions (C^s), we do not subtract any bias on C^s. Therefore, we use the seen (S), unseen (U), and HM metrics.

3.3.2 Baselines and Implementation Details

For **OW-CZSL**, we compare ProCC with other OW-CZSL methods, including CompCos [124], KGSP [79], and Co-CGE [125]. CZSL methods are also compared, including LE+ [129], AoP [131], TMN [140], SymNet [107], CGE [130], and CANet [167]. For **pCZSL**, ProCC is compared with KGSP [79] as well as standard (OW-)CZSL methods like CGE [130], CompCos [124], and Co-CGE [125], with the same partial label protocol.

Following the standard protocols in the CZSL, we utilize the pre-trained ResNet-18 [52] as the feature encoder (ω) to extract 512-dimensional feature vectors and learn classifiers on top of these features. Following [130, 79], each classifier is composed of Multi-Layer Perceptrons (MLP) with three layers with dimensions 768, 512, and the number of output classes, respectively, and comprise Layer Normalization [6] and Dropout [156]. To be consistent with other methods, we randomly augment input images with random crop and horizontal flip. We use PyTorch to implement our network and optimize it with Adam [85] with default settings. The batch size is 256, and the learning rate is 5.0×10^{-5} for the first two stages and 1.0×10^{-5} for the third stage. For the UT-Zappos, MIT-States, and C-GQA datasets, the total training time is approximately 1, 3, and 5 hours for 30/60/20, 40/80/30, and 50/100/25 epochs for three stages, respectively, with the early stop strategy.

3.3.3 Open-World CZSL (OW-CZSL) Results

The results of OW-CZSL setting are illustrated in Table 3.1. Generally, closed-world CZSL methods achieve inferior performance, especially in two large datasets (i.e., C-GQA and MIT-States), due to the large cardinality of the output space. ProCC outperforms previous methods on almost all metrics in terms of three datasets. Concretely, as for the most challenging dataset, i.e., C-GQA, the proposed method exceeds the previous SOTA methods, especially for best harmonic (HM) metrics (3.4→3.8: ↑12%), which means that ProCC has the better ability to recognize both the seen and unseen compositions. Also, in the validation sub-dataset, Our method suppresses the best baseline (i.e., KGSP) by a large margin in two overall evaluation indexes (i.e., HM: 13.2→16.1: ↑22%; AUC: 2.9→4.0: ↑38%). As for the MIT-States dataset, our method also has comparative results. Notably, we achieve the best performance on

Method	**C-GQA**						**MIT-States**						**UT-Zappos**					
	Val			Test			Val			Test			Val			Test		
	S	U	HM	S	U	HM	S	U	HM	S	U	HM	S	U	HM	S	U	HM
CGE	19.2	2.9	5.6	17.4	0.4	0.9	10.0	2.8	4.3	**19.6**	1.3	2.4	46.5	3.5	6.6	50.3	3.4	5.0
CompCos	18.2	3.0	5.2	24.3	0.4	0.7	11.1	2.9	4.6	10.8	2.0	3.6	50.2	3.9	7.3	52.4	4.1	7.6
Co-CGE	19.8	3.9	6.4	22.1	0.6	1.2	14.8	3.3	5.3	13.1	2.3	4.0	47.2	6.1	10.8	52.6	5.4	9.9
KGSP	20.1	4.8	8.3	22.3	0.9	1.7	15.7	3.2	5.3	13.5	2.6	4.4	49.4	5.9	9.7	53.8	6.9	12.3
Ours	**21.6**	**5.4**	**8.7**	**24.1**	**1.1**	**2.0**	**16.3**	**3.5**	**5.8**	14.1	**2.9**	**4.8**	**51.0**	**7.1**	**12.5**	**55.1**	**8.1**	**14.1**

Table 3.2: **Quantitative comparisons in the pCZSL setting.** We report the seen (S), unseen (U) accuracy, and best harmonic mean (HM) on the test and validation sub-datasets. The best and second-best results are bold and underlined.

the *U* metric, which validates the generalization ability of ProCC. For UT-Zappos, it is specially designed for shoes and is relatively simpler than others. ProCC consistently outperforms others, i.e., *S*: 59.3→62.2; *U*: 47.2→48.0; HM: 39.1→39.9; AUC: 22.9→23.6. Remarkably, previous methods typically utilize word embeddings to encode the word expression, which already contains semantic knowledge of similar objects and attributes for composition learning [147]. Recent Visual Product based method [79] employs more complex classifiers (with hidden layers of 768 and 1024) than ours as well as uses external knowledge to eliminate the less feasibility compositions. We predict the state and object primitives with more lightweight classifiers and explicitly model the cross-primitive interactions to learn the relationship between primitives without external knowledge.

3.3.4 Partial-supervision CZSL (pCZSL) Results

As for the more challenging setting, pCZSL, the challenges come from not only the huge output composition space but also the missing labels. As we can learn from Table 3.2, our method achieves SOTA performances compared with previous CZSL, OW-CZSL, and pCZSL methods. Concretely, for the largest dataset, C-GQA, the performance of SOTAs on pCZSL severely degrades compared with OW-CZSL, even for KGSP, which is equipped with the pseudo label and external knowledge. Our

Method	**OW-CZSL**				**pCZSL**					
	C-GQA		MIT-States		C-GQA			MIT-States		
	HM	AUC	HM	AUC	S	U	HM	S	U	HM
w/o CPC	3.3	0.40	6.2	0.8	17.4	0.5	1.0	11.6	2.2	3.7
w/o CPI	3.4	0.41	6.1	0.9	17.7	0.5	1.0	12.0	2.1	3.6
w/o CPM	3.5	0.48	6.6	1.0	18.9	0.7	1.4	12.2	2.5	4.1
w/o P-L	3.7	0.50	7.6	1.5	22.4	0.8	1.6	12.5	2.5	4.1
w/ Ex-1&2	3.6	0.48	7.8	1.5	22.6	1.0	1.9	13.2	2.7	4.4
w/o Stage3	3.5	0.47	7.4	1.4	23.2	1.1	2.0	13.6	2.8	4.6
w/ 4 Stages	3.6	0.50	7.6	1.4	23.7	1.0	1.9	13.8	2.8	4.7
w/ 5 Stages	3.7	0.53	7.7	1.4	23.9	1.1	2.1	13.6	2.9	4.8
w/ 6 Stages	3.8	0.56	7.7	1.6	24.0	1.1	2.1	13.8	2.8	4.7
Ours	3.8	0.54	7.8	1.6	24.1	1.1	2.0	14.1	2.9	4.8

Table 3.3: **Ablation studies** for both OW-CZSL and pCZSL.

method consistently exceeds them both on validation and testing datasets. For the MIT-States dataset, our method surpasses the second-best method by a large margin in HM metric (i.e., val: 5.3→5.8:↑9%; test: 4.4→4.8:↑9%). For the simplest dataset, UT-Zappos, our method also has the best performance. Note that we do not use any external knowledge like Word2vec, Glove, Conceptnet, and other semi-supervised learning techniques [91, 44] for the missing annotations. The superior performance indicates even with partial labels of object and state primitives, our progressive learning strategy can also model the interactions of cross primitives with the pre-trained classifiers.

3.4 Ablation Analysis

We analyze two important components: Cross-Primitive Compatibility (CPC) module and the progressive learning strategy. We adopt the same implementation strategy and conduct the OW-CZSL and pCZSL experiments on the two largest datasets, i.e., C-GQA and MIT-States.

Effect of the Cross-Primitive Compatibility Module. In Table 3.3, (I) without the CPC module (w/o CPC), the performance is severely degraded both on the OW-CZSL and pCZSL settings. Because lacking the interaction between cross primitives makes the network degenerate to previous Visual Product baselines [78, 79]. Meanwhile, KGSP utilizes the external knowledge and surpasses the ablation configuration, especially in pCZSL setting. (II) Moreover, to further evaluate the conditional modulation, we employ channel attention [59, 166] on the same primitive classifiers without cross-primitive interaction (w/o CPI). (III) Also, we ablate the learnable cross-primitive memory (w/o CPM) and directly modulate other primitives with learned features. Results indicate that exploring internal primitives brings marginal improvement for composition learning as classifiers have extracted enough internal information, and modulating primitives via hard masks also gives sub-optimal results. Note that the CPC is extremely lightweight with two trainable 1d convolution layers. Generally, the CPC module greatly improves the performance with negligible computation burden also without external information, which is practical for real-world scenes.

Effect of the Progressive Learning Strategy. Another important aspect of the ProCC is the progressive learning strategy. From Table 3.3, (I) we can learn that with the traditional end-end training strategy (w/o P-L), the performance of ProCC degrades to some extent, especially in the pCZSL setting (i.e., HM: 2.0→1.6 (C-GQA) and 4.8→4.1 (MIT-States)). As jointly training the whole network under the pCZSL setting does not explicitly learn the relationship between state and object primitives, which is the critical issue in the CZSL task. While for the OW-CZSL setting, joint training induces some noisy conditioned information, due to the diverse difficulty of classifying object and state primitives. Also, we exchange the training sequence (i.e., Stage 2 → 1 → 3) (w/ Ex-1&2) and ablate the fine-tuning stage (w/o Stage 3). (II) For the configuration of w/ Ex-1&2, the performance of ProCC degrades on both settings. Due to the challenge of classifying state primitives [147, 79], modulation object features conditioned on noisy state features results in invalid interactions. (III) For the configuration of w/o Stage 3, where only $\mathrm{CPC}_{o\rightarrow s}$ works, the performance degrades to some extent. We have two observations: $\mathrm{CPC}_{o\rightarrow s}$ brings more improvements than $\mathrm{CPC}_{s\rightarrow o}$; $\mathrm{CPC}_{s\rightarrow o}$ and fine-tuning based on well-trained features also matter for the cross-primitive compatibility and global optimum. (IV) Moreover, following the same training protocol, we train the network for more stages, i.e., with extra Stage 1 (w/ 4 Stages), extra Stage 1 and 2 (w/ 5 Stages), and extra Stage 1, 2, and 3 (w/ 6 Stages). We see that more training stages can not bring much accuracy improvement, as the model has converged after Stage 3.

3.5 Chapter Summary

This chapter proposes the Progressive Cross-Primitive Compatibility (ProCC) framework to enhance multimodal compositional generalization in open-world scenarios, addressing both Open-World (OW-CZSL) and partially supervised (pCZSL) Compositional Zero-Shot Learning. ProCC introduces a Cross-Primitive Compatibility (CPC) module that models conditional dependencies between object and state modality primitives (e.g., inferring "ripe" only for edible objects) through self-supervised visual-semantic correlations, eliminating reliance on external linguistic resources. Complemented by a progressive learning strategy, the framework adopts a curriculum-driven, easy-to-hard paradigm—first learning coarse-grained primitive distinctions (e.g., "metal" vs. "wood") before refining fine-grained compatibility constraints (e.g., "polished" vs. "rusty")—effectively suppressing invalid compositions (e.g., "flying tables") and noisy supervision. Experiments on benchmarks like MIT-States and CGQA demonstrate ProCC's superiority, outperforming state-of-the-art methods by a large margin in accuracy under OW-CZSL settings and reducing invalid predictions by 27% in pCZSL. By addressing combinatorial complexity and noisy adaptation through adaptive cross-modal conditioning, ProCC advances the thesis's core theme of robust open-world learning, and in the subsequent chapters, we introduce the cross-modal knowledge distillation for missing or invalid modality situations (Chapter 4: C^2KD) and Self-Introspective Decoding for multimodal large model hallucination alleviation (Chapter 5: SID).

Chapter 4

C^{2}KD: Bridging the Modality Gap for Cross-Modal Knowledge Distillation

4.1 Challenges and Motivations

Knowledge Distillation (KD) is an effective approach to transfer knowledge from the large-capacity teacher model to the low-capacity student model during training [165, 43]. During the KD process, the student is trained to mimic the teacher's output via the distillation loss. KD methods can be divided into two main categories: logits-based and feature-based methods. The former minimizes the discrepancy between soft labels of the teacher model and the student model [54, 203, 62], and the latter distills knowledge from intermediate feature layers [53, 23, 63].

Despite the success of traditional KD methods in single modality scenario, extending these methods to address the Cross-Modal Knowledge Distillation (CMKD) tasks remains a critical challenge. The CMKD task involves knowledge transfer from one modality to another during the distillation phase, with inference *only* on the *distilled*

	AVE[158]		**VGGSound**[18]	
Method	**Visual** (**A**→**V**)	**Audio** (**V**→**A**)	**Visual** (**A**→**V**)	**Audio** (**V**→**A**)
w/o KD	31.6±0.18	52.8±0.11	38.7±0.16	59.4±0.16
KD [54]	32.3±0.35	46.6±0.24	38.5±0.50	56.3±0.46
Review [23]	32.1±0.63	50.6±0.31	38.2±0.47	57.9±0.33
DML [203]	31.8±0.41	48.0±1.31	38.7±0.86	58.2±1.01
SHAKE [101]	32.2±0.59	47.3±0.72	38.3±0.41	59.5±0.34
DKD [204]	32.6±0.65	48.6±1.02	38.1±0.43	57.2±0.86
DIST [62]	29.8±0.61	49.3±0.52	38.5±0.39	58.9±0.45
NKD [188]	32.9±0.32	52.2±0.62	39.2±0.52	59.3±0.40
Ours	34.7±0.23	54.9±0.16	40.9±0.31	61.9±0.27

Table 4.1: **Performances of traditional KD in CMKD.** The results of distilled student modality infer only on the student modality. **A**→**V**: Audio teacher modality distills visual student modality; **V**→**A**: Visual teacher modality distills audio student modality.

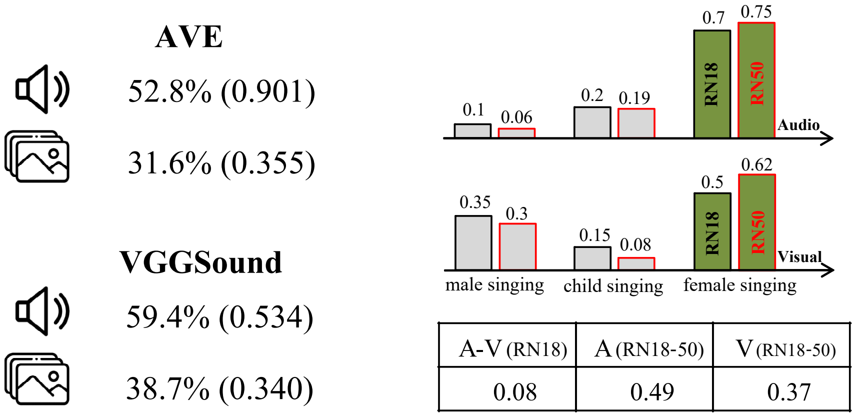
student modality, which is crucial especially in computation-constrained and sensor-failure scenarios. As demonstrated in Table 4.1, unimodal KD methods struggle to transfer knowledge from the low-accuracy visual modality to the high-accuracy audio modality, while the visual modality has only marginal gains from the audio modality. Based on the above analysis, a pivotal and fundamental question arises: *Can we effectively transfer arbitrary unimodal information to another modality?*

To answer this question, we conduct empirical analysis to investigate why traditional KD methods fail in CMKD from the logits-based perspective, which can be attributed to the inter-modality gap that inducing *modality imbalance* and *soft label misalignment*, as illustrated in Figure 4.1.

For the first factor, we define *modality imbalance*, akin to [137, 37], as the performance disparities between modalities. We quantitatively calculate the top-1 accuracy (followed by the average prediction probability of target classes) after training on the corresponding single modality. Figure 4.1(a) shows that the audio modality outperforms the visual modality in AVE and VGGsound datasets, and there are significant gaps in the average prediction probability of the target class, particularly in AVE. Merely distilling knowledge from the visual to the audio modality could potentially yield adverse effects, as shown in Table 4.1 (Audio columns).

For the second factor, we define *soft label* as the output distributions from the teacher network, following [54, 108]. The soft labels contain meaningful information on similarity among various classes. However, inter-modality gap leads to severe soft label misalignment between teacher and student modalities. Take three-class classification as an example (Figure 4.1(b) Up). Although both Audio and Visual modalities branches successfully predict the target class of 'female singing', the non-target soft labels are rank-distorted, where the *audio accent* of 'child singing' is more closely related to 'female singing', while the *visual appearance* of 'male singing' is more closely resembles 'female singing'. Direct transferring soft label information across modalities is unreasonable, which could explain why distilling the audio modality to the visual modality does not yield significant improvements. To quantitatively validate soft label misalignment, we further calculate the average Kendall Rank Correlation (KRC) [81] of soft labels in Figure 4.1(b) Down. A higher KRC indicates better rank correlation. The table indicates the KRC of multimodal soft label (i.e., A-V(RN-18)) is significantly lower than that of a single modality with diverse-capacity networks (i.e., A(RN18-50) and V(RN18-50)), indicating the presence of misalignment of multimodal soft labels.

To address the above issues in CMKD, this chapter proposes Customized cross-modal Knowledge Distillation (C^{2}KD). Concretely, instead of using the pre-trained teacher to provide supervision signals to the student, we bidirectionally update to



A-V (RN18)	A (RN18-50)	V (RN18-50)
0.08	0.49	0.37

(a) modality imbalance **(b) soft label misalignment**

Figure 4.1: **The Modality Gap of CMKD. (a)** Top-1 accuracy (followed by average prediction probability of target classes) of each modality. Both modalities utilize ResNet-18 as the backbone. **(b)** Up: Example of three-class classification. Down: Kendall Rank Correlation [81] of soft labels across modalities in VGGSound. A: audio; V: visual; RN: ResNet.

customize both the pre-trained teacher and student via On-the-Fly Selection Distillation (OFSD) strategy, where OFSD selectively distill receptive soft labels according to the Kendall Rank Correlation, and cross-modal knowledge is transferred from non-target classes to avoid the modality imbalance issue. Furthermore, Proxy student and teacher, inheriting unimodal and cross-modal knowledge, is formulated to progressively transfer cross-modal knowledge in the bidirectional distillation form.

The main contributions of this chapter can be summarized as follows.

- We empirically analyze the factors for the failure of unimodal KD in CMKD, which can be attributed to the modality imbalance and soft label misalignment.

- To address these issues, we propose a novel method named C^{2}KD. Specifically, OFSD produces selected crossmodel non-target class knowledge through on-the-fly bidirectionally distilling both student and teacher. Moreover, Proxy student and teacher are built to progressively transfer receptive knowledge across modalities. The proposed strategies are plug-and-play, enhancing traditional KD methods in CMKD.

- We conduct experiments on sparse and dense prediction tasks, including audio-visual, image-text, and RGB-Depth datasets. Diverse capacities and homogeneous/heterogeneous architectures are also considered. Extensive experiments validate C^{2}KD can transfer cross-modal knowledge from *arbitrary* modality to another.

4.2 Cross-Modal KD Effectiveness Analysis

First, this sub-chapter revisits traditional KD in cross-modal scenario. Given multimodal training data ($[X_1, X_2]$, Y) containing multimodal samples X_1 and X_2 and labels Y. Let f_T and f_S be the output logits of the teacher T and student S. The corresponding prediction probabilities are obtained using the softmax function (σ): $p_S = \sigma(f_S)$ and $p_T = \sigma(f_T)$. Typical KD trains the student network as follows:

$$L_{KD} = \mathcal{H}(p_S, Y) + \lambda\mathcal{D}(p_S, p_T) \tag{4.1}$$

where $\mathcal{H}$ is the supervision loss function (typical Cross-Entropy (CE) loss), $\mathcal{D}$ is the KD loss to minimize the discrepancy of output distribution between teachers and students, commonly achieved using Kullback–Leibler (KL) divergence [54], and λ is a balancing parameter for these two terms. Pioneering work [179] proposes the Modality Focusing Hypothesis (MFH) and claims that modality-general decisive features are crucial for transferring knowledge across modalities during the distillation phase. In this work, we provide another fine-grained perspective to investigate the efficacy of CMKD: the modality gap, which refers to the *modality imbalance* in target-class logits and *soft label misalignment*, incurs the failure of CMKD.

Regarding *modality imbalance*, as depicted in Figure 4.1(a), the prediction possibility of the target class exhibits significant variations across modalities.

If simply let student modality (audio) imitate teacher modality (vision), audio will inevitably reduce prediction confidence [164] and conflict *one-hot* label (Y). To validate our claim, we follow DKD [204] and decouple the KD loss into Target Class (TC) and Non-target Class (NC) KD:

$$\begin{aligned}\mathcal{D}(f_S, f_T) &= \underbrace{\alpha[p_T^t log(\frac{p_T^t}{p_S^t}) + p_T^{\backslash t} log(\frac{p_T^{\backslash t}}{p_S^{\backslash t}})]}_{\text{TCKD}} \\ &= \underbrace{\beta \sum_{i=1, i\neq t}^{C} \hat{p}_T^i log(\frac{\hat{p}_T^i}{\hat{p}_S^i})}_{\text{NCKD}}\end{aligned} \tag{4.2}$$

where α and β are hyperparameters. p^t denotes the target class probability: $p^t = \exp(f^t)/\sum_{j=1}^{C}\exp(f^j)$, $p^{\backslash t}$ represents the probability of all the other non-target classes $p^{\backslash t} = \sum_{k=1,k\neq t}^{C}\exp(f^i)/\sum_{j=1}^{C}\exp(f^j)$, and $\hat{p}^i$ means the probability among non-target classes: $\hat{p}^i = \exp(f^i)/\sum_{j=1,j\neq t}^{C}\exp(f^j)$. Here, C is the number of classes. When *only* applying TCKD in CMKD, as shown in Table 4.2, the performance of distilled audio modality severely degrades 4.8% and 3.6%, respectively, while the distilled visual modality is not clearly enhanced. Therefore, *modality imbalance* hinders the efficiency of CMKD, particularly when transferring knowledge from a low-accuracy modality to a high-accuracy modality.

To analyze *soft label misalignment*, we only conduct NCKD (Equation 4.2) to exclude the influence of modality imbalance. As depicted in Table 4.2, the low-accuracy visual teacher modality degrades the performance of the high-accuracy audio student modality. *Notably*, distilling high-accuracy audio information into the low-accuracy visual modality only results in marginal gains in the AVE, while surprisingly exhibiting a degradation in the VGGsound. [54, 196, 108, 204] investigate the mechanism of logit distillation, as soft logits provide reliable similarity information between categories. The privileged similarity information brings fine-grained supervision compared to a one-hot label. However, in the context of CMKD, the category similarities between different modalities are varied and even *conflicting*. An intuitive example is the three-class classification example in Figure 4.1(b) Up, where the unreliable similarity information of non-target classes across the modalities is contradictory. Directly minimizing cross-modal distributions leads to performance degradation. To quantitatively evaluate the misalignment of soft labels, we employ the Kendall Rank Correlation (KRC) [81] metric to measure the rank correlation. Specifically, given teacher and student output logits f_T and f_S, the KRC between f_T and f_S can be explicitly computed as follows:

$$\text{KRC} = \frac{2}{C(C-1)} \sum_{i<j} \text{sign}(f_T^i - f_T^j)\,\text{sign}(f_S^i - f_S^j) \tag{4.3}$$

As depicted in Figure 4.1(b) Down, the KRC between multimodal networks is significantly lower than that observed in unimodal networks with different capacities. We argue that the misalignment of rank correlation is another reason for the failure of CMKD. To validate our argument, we filter out multimodal samples with KRC < 0 (+KRC), indicating that the count of misaligned soft label pairs is larger than aligned ones. Additionally, we randomly filter out the same number of samples (+Random). From Table 4.2, we can see that both visual and audio modalities are improved when guided by the KRC metric, whereas randomly filtering out samples has almost no effect.

4.3 Customized Cross-modal Knowledge Distillation (C^{2}KD)

Based on the aforementioned analysis, this chapter proposes a simple yet effective method named Customized Cross-modal Knowledge Distillation (C^{2}KD) to transfer cross-modal knowledge to an arbitrary single modality. To bridge the modality gap, we argue that both student and teacher should be tuned with the bidirectional

	AVE [158]				**VGGsound** [18]			
Method	**Visual** (**A**→**V**)		**Audio** (**V**→**A**)		**Visual** (**A**→**V**)		**Audio** (**V**→**A**)	
w/o KD	31.6		52.8		38.7		59.4	
proba.	0.355		0.901		0.340		0.534	
w/ KD	32.3	↑0.7	46.6	↓6.2	38.5	↓0.2	56.3	↓3.1
+Random	32.1	-0.2	46.8	+0.2	38.2	-0.3	56.4	+0.1
+KRC	32.9	+0.6	47.9	+1.3	39.2	+0.7	57.4	+1.1
TCKD	31.8	↑0.2	48.0	↓4.8	37.9	↓0.8	55.8	↓3.6
NCKD	31.9	↑0.3	50.1	↓2.7	38.5	↓0.2	57.5	↓1.9
+Random	31.5	-0.4	50.2	+0.1	38.5	-	57.6	+0.1
+KRC	33.1	+1.2	51.0	+0.9	39.6	+1.1	58.3	+0.8
DKD [204]	32.6	↑1.0	48.6	↓4.2	38.1	↓0.6	57.2	↓2.2

Table 4.2: **Efficacy Analysis on modality imbalance and soft label misalignment.** *proba.* represents average prediction probability of target class. DKD is with defaulted $\{\alpha = 1, \beta = 8\}$ (Eq. 4.2).

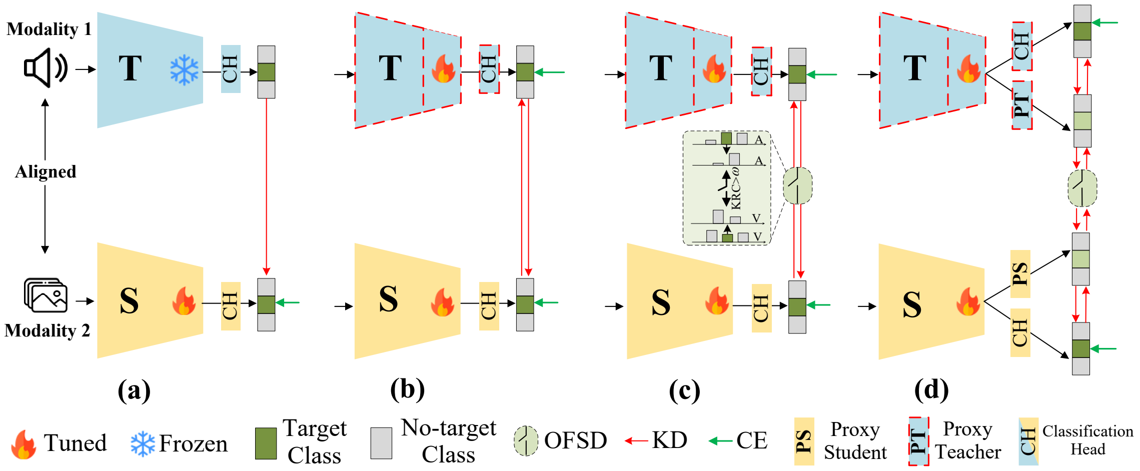


Figure 4.2: **Evolution of our Customized Cross-modal Knowledge Distillation (C^{2}KD) method.** (a) Traditional KD [54] with output logits from the fixed teacher. (b) We (partially) tune the teacher with the bidirectional distillation to provide customized teacher knowledge. (c) To bridge the modality gap of CMKD, On-the-Fly Selection Distillation (OFSD) is proposed to filter out samples with distorted rank correlations and perform KD on non-target classes. (d) Additionally, we introduce proxy teacher and proxy student as bridges to progressively transfer receptive cross-modal knowledge.

distillation from each other, in this way, teacher modality could provide receptive information for student modality. Meanwhile, the soft label misalignment samples should be filtered out otherwise induce conflicting information. Therefore, we propose the On-the-Fly Selection Distillation (OFSD) strategy to exclude non-distillable samples and inherit knowledge from non-target classes. Furthermore, dual proxies with the bidirectional distillation strategy are introduced to progressively transfer cross-modality knowledge. The evolution of our proposed framework is depicted in Figure 4.2.

4.3.1 Formulation of C^{2}KD

As illustrated in Figure 4.2(d), C^{2}KD proposes the OFSD strategy to dynamically select receptive knowledge. This strategy involves distilling knowledge from non-target classes and innovatively employing the Kendall Rank Correlation (KRC) [81] metric to filter out samples with rank-distorted soft labels. Given the output logits f_T and f_S from the teacher and student modalities, the sample selection strategy is as follows:

$$\eta = \begin{cases} 1, & \mathrm{KRC}(f_T, f_S) > \omega \\ 0, & otherwise \end{cases} \tag{4.4}$$

The KRC is as Equation 4.3, $\eta \in \{0, 1\}$ is OFSD filter, and ω is the threshold.

Moreover, we additionally build dual proxies to progressively produce soft labels. Formally, the output features (F) obtained from the backbone (B) are fed to the original classification head and the proposed proxy as follows:

$$\begin{aligned} f_m &= fc_m^{cls}(\mathrm{GAP}(B_m(F_m))), m \in \{\mathrm{T}, \mathrm{S}\} \\ f_m^{pro} &= fc_m^{cls(pro)}(\mathcal{A}[\mathrm{GAP}(B_m(F_m))]), m \in \{\mathrm{T}, \mathrm{S}\} \end{aligned} \tag{4.5}$$

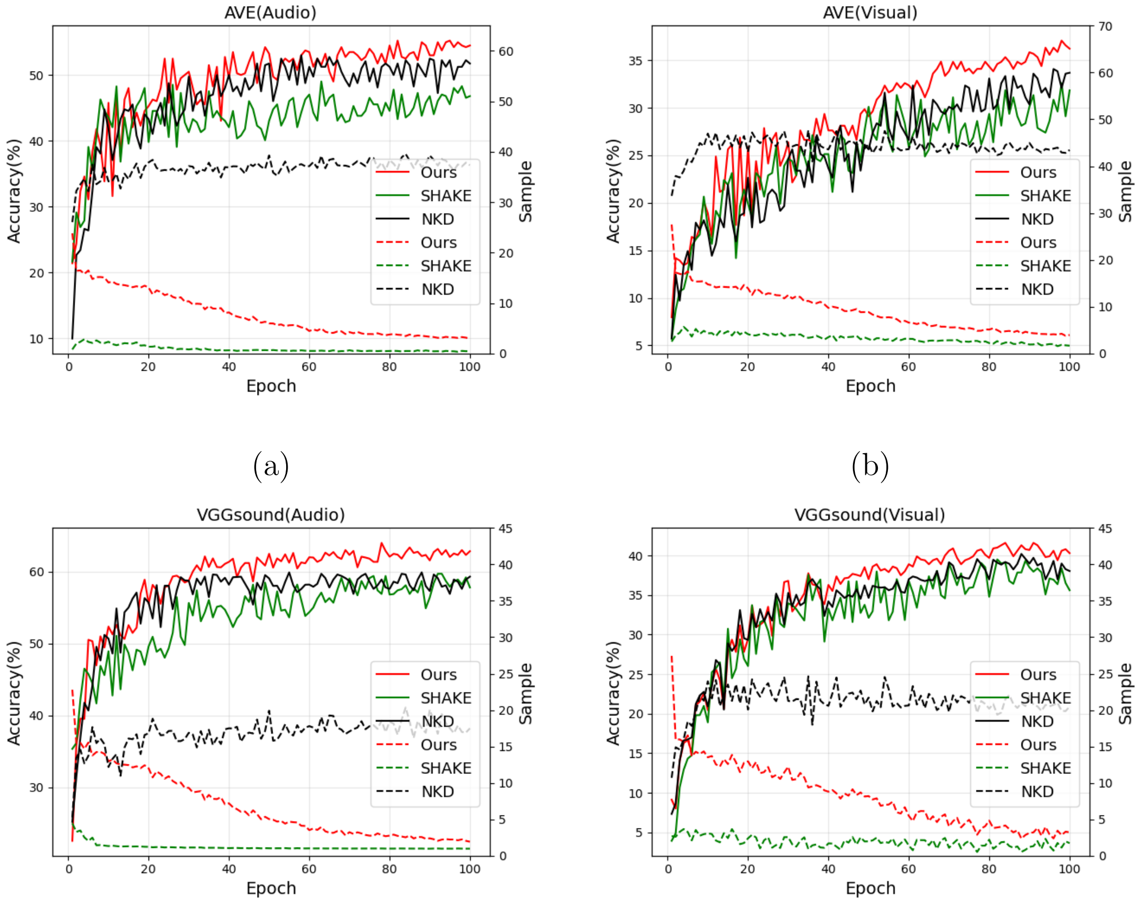
where GAP and fc^{cls} refer to global average pooling and classification head. $\mathcal{A}$ represents feature adaptation layer, akin to [145, 101], consisting of the 'Conv-BN-ReLU' block. To further produce customized knowledge, both student and teacher proxies serve as *bridges* and get bidirectional distillation from both uni-modality and cross modality. In summary, the total loss function can be expressed as follows:

$$\begin{aligned} L_{all} =&\mathcal{H}(\sigma(f_S), Y) + \mathcal{H}(\sigma(f_T), Y) \\ &+ \lambda_1 \mathcal{D}(\sigma(f_T), \sigma(f_T^{pro})) + \lambda_1 \mathcal{D}(\sigma(f_T^{pro}), \sigma(f_T)) \\ &+ \lambda_2 \mathcal{D}(\sigma(f_S), \sigma(f_S^{pro})) + \lambda_2 \mathcal{D}(\sigma(f_S^{pro}), \sigma(f_S)) \\ &+ \lambda_3 \eta \mathcal{D}(\sigma(\hat{f}_S^{pro(i)}), \sigma(\hat{f}_T^{pro(i)})) \\ &+ \lambda_3 \eta \mathcal{D}(\sigma(\hat{f}_T^{pro(i)}), \sigma(\hat{f}_S^{pro(i)})) \end{aligned} \tag{4.6}$$

where λ_1, λ_2, and λ_3 are balancing parameters and $i \neq t$. $\mathcal{H}$ and $\mathcal{D}$ represent supervision and KD loss, respectively. We simply set $\{\lambda_1 = \lambda_2 = \lambda_3 = 1\}$ in all experiments.

4.3.2 Analysis of Cross-modal Knowledge distillation

Understanding CMKD training dynamics. We visualize the training dynamics of CMKD and compare it with SHAKE [101] and NKD [188] to demonstrate the CMKD progress. Figure 4.3 shows the test accuracy and the average number of samples with KRC $< \omega$ ($\omega = 0$) during the training process. As the advanced online KD, SHAKE gets the reverse cross-modal feedback supervision without discrimination. However, SHAKE suffers from severe instability of training, possibly due to conflicting cross-modal information. Meanwhile, the sample number of KRC $< \omega$ drops to close to 0 within initial epochs, which represents the teacher modality is influenced by the student modality and might lose teacher modality information. In contrast, NKD minimizes the distance between student modality logits and teacher modality logits. The teacher model of NKD is not updated to cater to student modality, so the sample number of KRC $< \omega$ is large, and NKD also falls into the unstable training process. As for ours, we selectively inherit cross-modal knowledge based on KRC and progressively update the teacher model through proxies to obtain receptive knowledge. During the distillation progress, the rank-distorted samples gradually reduce, and our method only filters out the non-distillable samples.



(a) (b)

(c) (d)

Figure 4.3: **Training dynamics analysis.** The solid lines correspond to the test accuracy, and the dotted lines indicate the average number of samples with KRC $< \omega$ each data batch during the training process. Here we set $\{\omega = 0, \text{batchsize} = 64\}$.

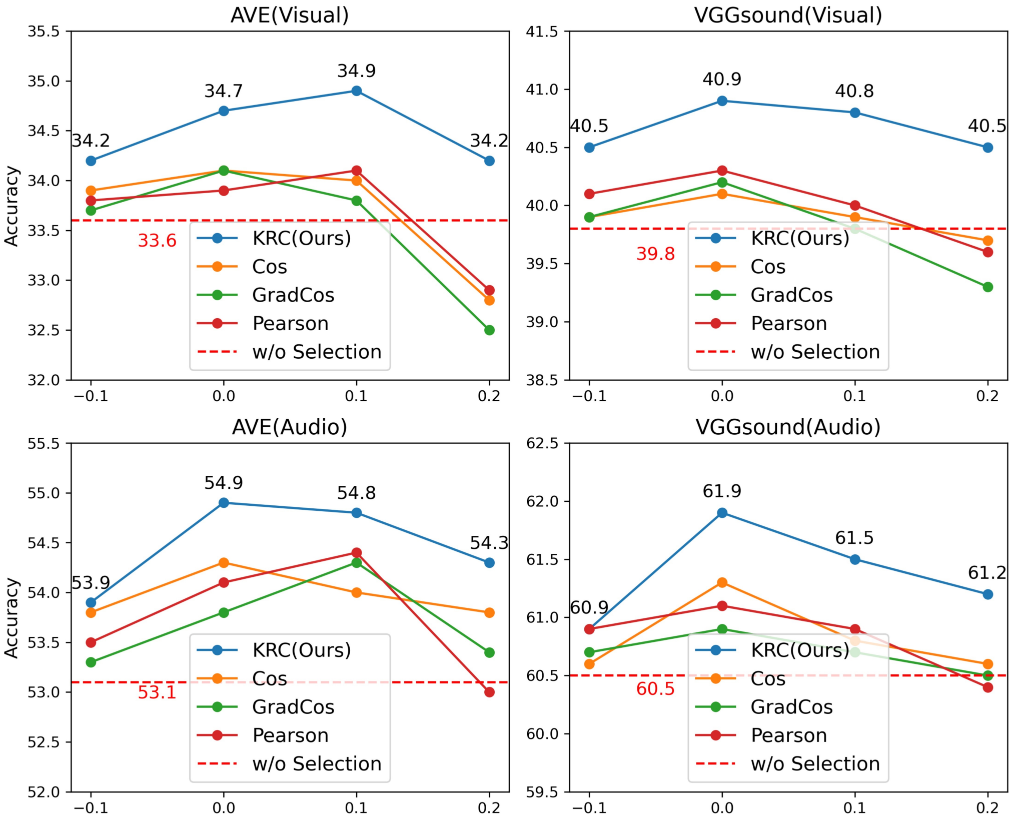


Figure 4.4: **Comparisons of different distance metrics.** The X-axis represents the value of ω.

Comparisons with other distance metrics. We select other distance metrics to verify the effectiveness of KRC defined in Equation 4.4. Concretely, we choose the cosine similarity (Cos), gradient cosine similarity (GradCos), and Pearson correlation coefficient [136] (Pearson) as alternatives. Cosine similarity and Pearson correlation coefficient are used to measure the distance between teacher and student logits, ranging from -1 to 1. They can be formulated as: $\amalg(f_T, f_S) > \omega$, $\amalg \in \{\text{Cos}; \text{Pearson}\}$. Similar to [194, 212], Gradient cosine similarity regards CMKD (Equation 4.1) as two tasks: cross-modal distillation ($L_{cmkd} = \mathcal{D}(p_S, p_T)$) and unimodal task ($L_{task} = \mathcal{H}(p_S, Y)$) and calculates the gradient cosine similarity between these two tasks as: $\text{Cos}(\nabla_\theta L_{cmkd}, \nabla_\theta L_{task}) > \omega$. It's worth noting that these three metrics consider *both* rank and intensity between cross-modal logits, while KRC *only* concerns about rank-distorted ones. As shown in Figure 4.4, KRC makes the best performance among these metrics. Although inferior to KRC metric, other metrics with proper ω perform better than without sample selection (w/o Selection) strategy. The results validate the necessity of filtering out samples with misaligned soft labels.

4.4 Experimental Results

We conduct extensive experiments to validate the effectiveness of our method. First, we compare our method with KD methods regarding multimodal classification tasks. Also, we apply our method to the multimodal semantic segmentation. Then, we perform ablation and sensitivity analysis.

4.4.1 Multimodal Classification

We follow [137, 37, 1] and conduct experiments on four visual-audio and image-text datasets: (1) **CREMA-D** [15] is an audio-visual dataset for speech emotion recognition, with 6 categorizations. (2) **AVE** [158] is an audio-visual dataset for audio-visual event localization, in which there are 28 event classes. (3) **VGGsound** [18] is a large-scale video dataset containing 309 classes covering daily life activities. We randomly choose *50* class to conduct experiments due to limited computation resources. (4) **CrisisMMD** [2] is a multimodal crisis prediction dataset and is divided into eight humanitarian categories.

Implementation. For *visual-audio* datasets, the preprocess strategy follows [137, 37]. Concretely, for audio modality, we change the input channel from 3 to 1 as [18]. Audio data is transformed into a spectrogram of size 257×299 for CREMA-D, 257×1,004 for AVE, and 257×1,004 for VGGsound, respectively, with the window length of 512 and overlap of 353. For visual modality, the input channel is adjusted considering input frames [205]. Concretely, 3 frames are uniformly sampled from

	CREMA-D		AVE		VGGsound		CrisisMMD	
Method	**Visual**	**Audio**	**Visual**	**Audio**	**Visual**	**Audio**	**Image**	**Text**
w/o KD	58.1±0.33	56.3±0.22	31.6±0.18	52.8±0.11	38.7±0.16	59.4±0.16	66.7±0.22	68.1±0.21
FitNet[145]	56.4±0.47	52.9±0.32	29.6±0.63	48.0±0.81	37.9±0.39	57.1±0.79	-	-
Review[23]	59.6±0.45	55.7±0.36	32.1±0.63	50.6±0.31	38.2±0.47	57.9±0.33	-	-
KD[54]	57.4±0.92	53.4±0.85	32.3±0.35	46.6±0.24	38.5±0.50	56.3±0.46	66.3±0.24	68.4±0.12
DML[203]	60.3±1.60	56.4±0.55	31.8±0.41	48.0±1.31	38.7±0.86	58.2±1.01	67.9±0.18	69.6±0.24
SHAKE[101]	60.0±0.35	58.6±0.61	32.2±0.59	47.3±0.72	38.3±0.41	59.5±0.34	68.1±0.16	69.7±0.26
RKD[134]	48.3±0.68	51.9±1.36	28.2±0.71	44.5±0.73	33.4±0.49	41.5±1.36	67.0±0.23	67.4±0.21
DKD [204]	60.4±0.82	55.1±0.65	32.6±0.65	48.6±1.02	38.1±0.43	57.2±0.86	68.0±0.17	69.2±0.23
DIST [62]	61.1±1.82	57.9±0.57	29.8±0.61	49.3±0.29	38.5±0.39	58.9±0.45	68.3±0.21	67.8±0.18
NKD [188]	60.6±0.64	56.1±0.68	32.9±0.32	52.2±0.62	39.2±0.52	59.3±0.40	67.2±0.26	68.5±0.16
Ours†	62.4±0.24	60.5±0.37	34.2±0.28	54.5±0.22	40.8±0.23	61.6±0.34	68.2±0.09	69.8±0.16
Ours‡	**62.8**±0.28	**61.4**±0.44	**34.7**±0.23	**54.9**±0.16	**40.9**±0.31	**61.9**±0.27	**68.8**±0.15	**70.1**±0.12

Table 4.3: **Comparison results on Visual-Audio and Image-Text datasets.** The metric is the top-1 accuracy (%). Ours‡ means fully updating the teacher model, and Ours† means partially finetuning the top 2 layers. The best is in **bold**, and the second is underlined.

VGGsound, and 1 frame is extracted from AVE and CREMA-D. Standard augmentations are employed, including random cropping and flipping. We train the network for 100 epochs with 1e-2 initial learning rate and decay follow the 'poly' policy with the power of 0.9. We use SGD with 0.9 momentum and default hyperparameters as the optimizer. For the *image-text* dataset, we use the same training strategies and adopt $\omega = 0$ across all experiments. Here, following [137, 1], we adopt the same ResNet-18 [52] as the backbone for visual and audio modality, and BERT-base [32] for text and MobileNetV2 [148] and image feature extractors, respectively. All results are the average of three different seeds.

Comparison Results. In Table 4.3, we compare our method to some advanced KD methods with the same training settings. We follow [137, 37] and give the detailed preprocess strategy. For audio modality, we change the input channel from 3 to 1 as [18]. Audio data is transformed into a spectrogram of size 257×299 for CREMA-D, 257×1,004 for AVE, and 257×1,004 for VGGsound, respectively, with the window length of 512 and overlap of 353. For visual modality, the input channel is adjusted considering input frames [205]. Concretely, 3 frames are uniformly sampled from VGGsound, and 1 frame is extracted from AVE and CREMA-D. Standard augmentations are employed, including random cropping and flipping. We initialize weights of the student model and proxies following [132]. All experiments are conducted with NVIDIA RTX3090 GPUs on CUDA 11.4 using the PyTorch framework. All results are the average of three different seeds, which are set to 1, 2, and 3, respectively. We imply traditional unimodal knowledge distillation with their defaulted settings. Previous logits-based KD methods can be seamlessly applied to the Cross-Modal Knowledge Distillation (CMKD) task. Due to the different spatial dimensions of multimodal inputs, the intermediate features have different spatial dimensions. Feature-based KD methods cannot be directly applied to CMKD. To deal with this issue, we employ the bilinear interpolation operator to align the intermediate features of teacher and student. Besides, BERT has 12 layers while MobileNetV2 has 5 layers. We do not conduct feature-based KD on the CrisisMMD dataset for comparisons because we can not choose which layers to be distilled based on their original implementations.

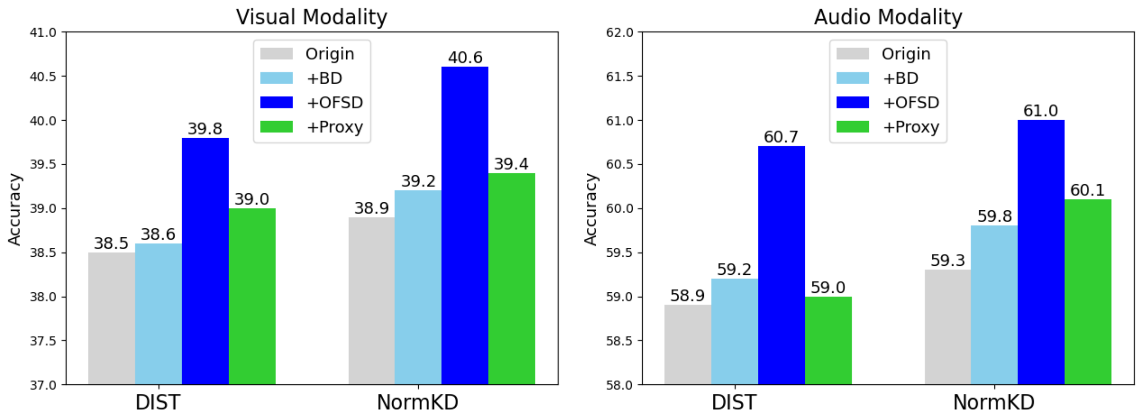
We can learn from Table 4.3 that our proposed method, C^{2}KD, consistently outperforms other KD methods across four datasets. Existing KD methods can not effectively distill one modality information to another modality, especially for the datasets with the significant modality imbalance issue like AVE and VGGsound. Concretely, feature-based KD (FitNet[145], Review[23]) methods fail in CMKD because of significant feature divergence (see Section 4.6). Online KD (DML [203] and SHAKE [101]) methods update teacher models and achieve better cross-modal knowledge transfer ability, compared with the baseline [54]. Due to soft label misalignment between modalities, the relation-based method (RKD [134]) degrades severely in CMKD. Recent advanced logits-based methods (DKD [204], DIST [62], and NKD [188]) significantly outperform the vanilla KL loss by proposing the relaxed KD functions and logits decoupling strategies. However, these methods fail to transfer cross-modal knowledge from low-accuracy to high-accuracy modality, impeding their practical deployments in CMKD.

4.4.2 Multimodal Semantic Segmentation

We also extend C^{2}KD to the multimodal semantic segmentation, a challenging dense prediction task. Concretely, following [179], we conduct experiments on the NYU-Depth V2 dataset [154]. NYU-Depth V2 contains 1,449 aligned RGB and depth pairs with 40 category labels, of which 795 pairs are used for training, and 654 pairs are used for testing.

Implementation. Both teacher and student networks deploy the DeepLab V3+ [21] architecture with diverse backbones. The training settings follow [182] that we adopt SGD as the optimizer with a momentum of 0.9, a batch size of 16, an initial learning rate of 0.02, and ImageNet pre-trained weights. The total training iterations is 40K, decayed by the 'poly' policy with the power of 0.9. Experiments on homogeneous/heterogeneous backbones, including ResNet-18/ResNet-18 and ResNet-18/MobileNetV2 pairs, are conducted to validate our method. All results are the average of three different seeds.

Comparison Results. We compare our methods with advanced traditional KD methods (KD [54], SHAKE [101], DIST [62], and NKD [188]) as well as the semantic segmentation KD method (CIRKD [182]). The results compared with previous methods are summarized in Table 4.4. We can see that previous KD methods do not perform well in CMKD, especially in transferring low-accuracy modality information to high-accuracy modality. Our method can significantly improve the distilled performance of arbitrary single modality. For instance, ours consistently surpasses the advanced CIRKD in transferring depth information to RGB modality. Besides, when replacing KL loss with DIST loss [62], our method affords clear improvements.



(a) T: audio; S: visual

(b) T: visual; S: audio

Figure 4.5: **Generalizability of each module.** We conduct experiments on VGGsound dataset in terms of DIST [62] and NKD [188].

4.5 Ablation and Sensitivity Analysis

Effectiveness and generalizability of each module. We analyze how the proposed modules improve CMKD. Table 4.5 reports the results of ablation studies on AVE, VGGsound, and CrisisMMD with the same backbones. The configurations of (a), (b), (c), and (d) correspond to the evolution steps shown in Figure 4.2. Compared to the vanilla KD (i.e., (a)), the Bidirectional Distillation (BD) updates the

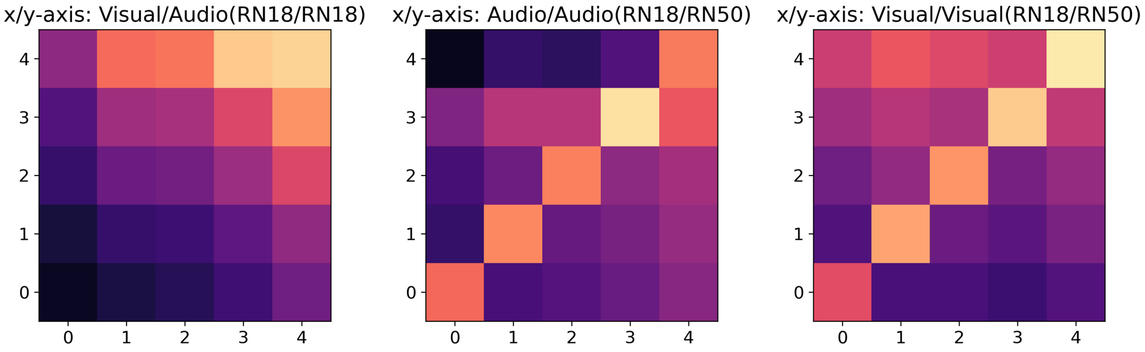


Figure 4.6: **The CKA score of intermediate features on AVE.**

	RGB	**Depth**	**RGB**	**Depth**	**RGB**	**Depth**
	RN18	RN18	RN18	MNV2	MNV2	RN18
w/o KD	36.1	30.5	36.1	31.2	36.3	30.5
KD [54]	35.8	30.9	36.2	31.9	36.5	31.8
SHAKE [101]	37.1	31.2	37.0	32.7	37.1	32.9
DIST [62]	36.9	32.0	36.5	32.9	36.8	33.1
NKD [188]	36.5	30.8	36.4	32.2	36.4	32.7
CIRKD [182]	37.3	32.6	36.9	32.7	36.7	33.4
Ours	37.5	32.5	37.2	32.8	37.4	33.1
Ours+[62]	**38.1**	**33.2**	**37.7**	**33.5**	**37.9**	**33.7**

Table 4.4: **Comparison results on RGB-Depth semantic segmentation dataset.** The metric denotes the mean Intersection over Union (mIoU: %). Ours+[62] means we replace KL loss with the advanced DIST loss. RN18: ResNet-18; MNV2: MobileNetV2.

teacher model (i.e., (b)) to mitigate the model gap. Furthermore, to validate the effectiveness of OFSD, we decouple OFSD into the On-the-Fly Selection (OFS) strategy and Non-Target (NT) classes distillation approach. We can learn that both OFS and NT benefit CMKD, and the combination of both brings significant improvement compared to (b). The proxy teacher and student circumvent the direct imitation of cross-modal logits, serving as bridges for inheriting unimodal and cross-modal knowledge and facilitating the transfer of integrated knowledge through bidirectional distillation. The progressive KD strategy further improves the CMKD results. The structure of proxies adheres to [145, 101].

To ascertain the generalizability of each component, we incorporate the proposed plug-and-play modules into advanced KD methods (i.e., DIST [62] and NKD [188]). We can learn from Figure 4.5 that the proposed modules consistently improve the performances of traditional KD methods in the CMKD task, especially for the OFSD

	Proxy	OFS	NT	BD	**Proxy**		**VGGsound**		**CrisisMMD**	
					Visual	**Audio**	**Visual**	**Audio**	**Image**	**Text**
(e)					31.6	52.8	38.7	59.4	66.7	68.1
(a)					32.3	46.6	38.5	56.3	66.3	69.2
(b)				✓	32.7	47.9	38.8	57.6	67.3	69.2
(c)		✓	✓	✓	34.6	54.3	40.4	61.5	68.5	69.8
-			✓	✓	33.2	52.9	39.4	60.3	67.9	68.9
-		✓		✓	34.4	52.5	40.0	59.9	68.0	69.5
(d)	✓	✓	✓	✓	34.7	54.9	40.9	61.9	68.8	70.1

Table 4.5: **Ablation studies on each module.** (a), (b), (c), and (d) represent the evolution steps of C^2KD (Figure 4.2). (e) indicates the results without KD. The metric is the top-1 accuracy (%).

strategy.

Necessity of cross-modal KD. Considering the challenges of cross-modal KD, a question may arise: Do we really need CMKD rather than fully explore self-knowledge? Self-knowledge distillation (Self-KD) techniques [100, 181, 152, 108, 188] have been proposed to utilize the information within the student model to facilitate its learning process. Specially, DLB [152] leverages the soft targets generated in the last mini-batch backup for training consistency and stability. ZipfKD [108] and USKD [188] generate soft labels following the Zipf's law distribution [120]. We conduct experiments on these advanced Self-KD methods to validate the necessity of CMKD. From Table 4.6, we can learn that although Self-KD improves the performance of each modality, our method consistently outperforms Self-KD methods by a clear margin. The results indicate the necessity of cross-modal KD.

Parameter sensitivity. Here, we conduct a sensitivity study on KRC threshold ω. Results are in Figure 4.4. Large ω filters out more samples, which might hinder

	AVE		**VGGsound**		**CrisisMMD**	
Method	**Visual**	**Audio**	**Visual**	**Audio**	**Image**	**Text**
w/o KD	31.6	52.8	38.7	59.4	66.1	68.1
DLB[152]	32.6	53.3	39.1	60.2	66.9	68.6
ZipfKD[108]	33.3	53.5	40.2	60.3	67.4	68.9
USKD[188]	33.1	53.2	40.0	60.1	67.1	69.0
Ours	34.7	54.9	40.9	61.9	68.8	70.1

Table 4.6: **Comparison results of different Self-KD methods.**

cross-modal knowledge transfer, while low ω preserves more samples, which might contain rank-distorted samples that induce adverse effects. We heuristically set ω to 0 and achieve balanced results. Note that, as shown in Tables 4.3 and 4.5, even the worst accuracy of varying ω is still competitive with the baselines, we think the studies show the necessity of on-the-fly filtering out rank-distorted samples based on KRC. More analyses of λ_1, λ_2, and λ_3 are given in Figure 4.7. Our method is robust in terms of different hyperparameters. As our method can effectively transfer corssmodal information, large and small values of λ could hinder knowledge transfer. Therefore, we adopt $\{\lambda_1 = \lambda_2 = \lambda_3 = 1\}$ in all experiments.

Different Backbones Evaluations. Besides the results in Tables 4.3 and 4.4, we conduct more experiments to further demonstrate the effectiveness of our method across diverse-capacities homogeneous and heterogeneous architectures. We compare C^{2}KD with vanilla KD [54], the state-of-the-art feature-based KD (Review [23]), online KD (SHAKE [101]), logits-based KD (NKD [188]). The results in Table 4.7 illustrate C^{2}KD can effectively transfer crossmodal knowledge across diverse-capacities homogeneous architectures (i.e., ResNet-18-ResNet-50) and heterogeneous architectures (i.e., BERT-ResNet-18 and BERT-ShuffleNet V2).

Proxy Analysis. We provide the detailed analysis of the student and teacher proxies. The proxy consists of the feature adaptation layer and the linear classification

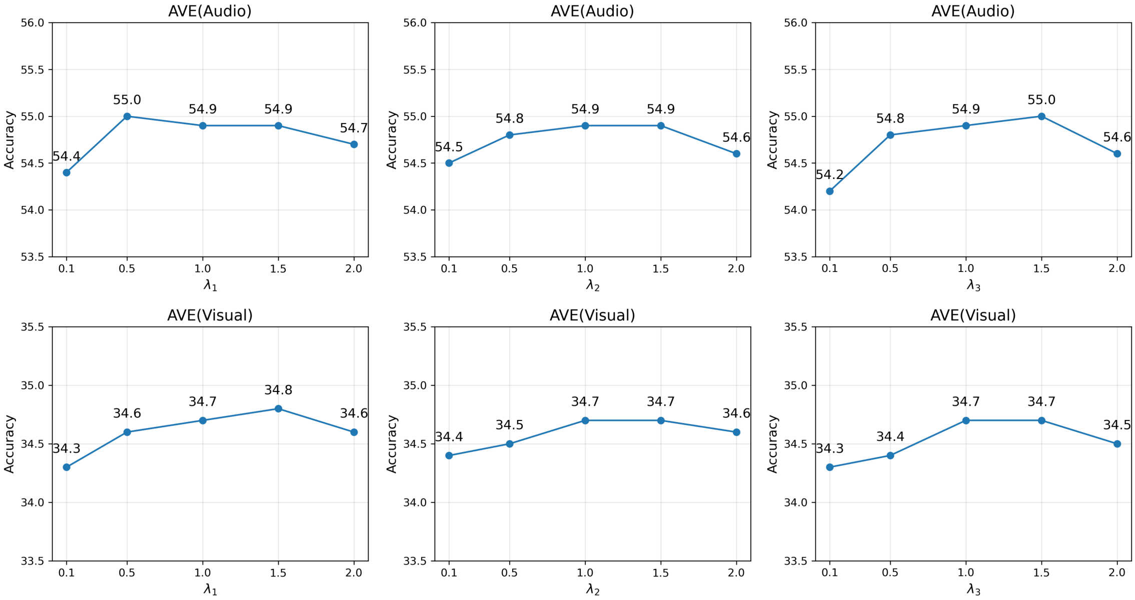


Figure 4.7: **Analysis of λ_1, λ_2, λ_3.** We conduct experiments on the AVE [158] dataset with ResNet-18 as the multimodal backbones.

	CREMA-D [15]		**AVE** [158]		**VGGsound** [18]		**CrisisMMD** [2]	
	Visual (A→ V)	**Audio (V→ A)**	**Visual (A→V)**	**Audio (V→A)**	**Visual (A→V)**	**Audio (V→A)**	**Image (T→I)**	**Text (I→T)**
	RN18	**RN50**	**RN18**	**RN50**	**RN18**	**RN50**	**BERT**	**SNV2**
w/o KD	58.1±0.33	57.9±0.19	31.6±0.18	53.7±0.16	38.7±0.16	60.1±0.18	66.7±0.22	68.0±0.12
KD [54]	57.1±0.57	54.1±0.43	32.6±0.62	48.5±0.35	39.0±0.46	57.8±0.51	66.2±0.38	68.4±0.22
Review [23]	59.4±0.52	56.9±0.62	32.0±0.53	51.3±0.57	38.5±0.53	58.7±0.60	-	-
SHAKE [101]	60.2±0.36	58.9±0.63	32.5±0.67	48.6±0.46	38.9±0.51	59.9±0.38	68.2±0.23	69.6±0.25
NKD [188]	60.5±0.62	56.9±0.43	33.0±0.36	52.5±0.36	39.2±0.67	59.6±0.54	67.3±0.31	68.6±0.25
Ours	63.1±0.25	62.1±0.37	35.0±0.21	55.3±0.12	41.0±0.22	62.0±0.23	68.9±0.12	70.0±0.09
	RN50	**RN18**	**RN50**	**RN18**	**RN50**	**RN18**	**BERT**	**RN18**
w/o KD	59.7±0.20	56.3±0.22	32.7±0.25	52.8±0.11	39.3±0.13	59.4±0.16	66.7±0.22	68.1±0.13
KD [54]	58.2±0.53	54.0±0.36	33.0±0.43	46.9±0.42	38.9±0.52	56.4±0.61	66.2±0.42	68.5±0.21
Review [23]	60.4±0.58	55.9±0.39	32.7±.0.56	51.2±0.61	38.2±0.43	58.1±0.61	-	-
SHAKE [101]	60.5±0.53	59.0±0.48	33.4±0.53	47.5±0.43	38.6±0.41	59.8±0.49	68.0±0.19	69.8±0.23
NKD [188]	60.9±0.54	58.4±0.62	33.2±0.47	52.8±0.55	39.5±0.53	59.1±0.46	67.4±0.26	68.6±0.22
Ours	63.5±0.28	61.6±0.23	35.5±0.30	55.1±0.22	41.3±0.28	62.1±0.24	68.8±0.16	70.2±0.16

Table 4.7: **Comparison results on Visual-Audio and Image-Text datasets.** The metric is the top-1 accuracy (%). RN18: ResNet-18; RN50: ResNet-50; SNV2: ShuffleNet V2 [122].

	AVE [158]		**VGGsound** [18]	
Method	**Visual (A→V)**	**Audio (V→A)**	**Visual (A→V)**	**Audio (V→A)**
w/o FA	34.4$\pm$0.36	53.0$\pm$0.22	40.2$\pm$0.25	60.6$\pm$0.24
w/ CFA	34.7$\pm$0.18	55.0$\pm$0.20	40.8$\pm$0.28	62.0$\pm$0.21
Ours	34.7$\pm$0.23	54.9$\pm$0.16	40.9$\pm$0.31	61.9$\pm$0.27

Table 4.8: **Analysis of the structure of the proxy.** We conduct experiments on the AVE [158] and VGGsound [18] datasets with ResNet-18 as the multimodal backbones.

head, as shown in Equation 5. The feature adaptation layer follows the feature-based KD methods [145, 23], consisting of 'Conv-BN-ReLU' block. Specifically, the kernel size of 'Conv' is set to 1×1, and input and output channel dimensions remain the same. Here, we analyse the structure of the proxy. We ablate the feature adaptation layer (w/o FA) and employ a complicated feature adaptation layer ('Conv-BN-Conv-BN-ReLU', i.e., w/ CFA). Table 4.8 illustrates that without the feature adaptation layer (w/o FA), the linear classification head can not effectively transfer crossmodal information, possibly due to the degradation of nonlinear ability. However, the complicated feature adaptation layer does not bring obvious improvement. Therefore, the feature adaptation layer and linear classification head constitute the proxy.

4.6 Discussion

Feature-based CMKD Analysis. We analyze the modality gap of CMKD in the *logits-based* perspective and propose C^{2}KD to mitigate the issues. Furthermore, we provide the analysis of the challenge of CMKD from the *feature-based* perspective. We adopt the Center Kernel Alignment (CKA) [87], a feature similarity metric that measures input similarity with different dimensions. As shown in Figure 4.6, com-

	AVE				VGGsound			
Method	**V**(S)	T	**A**(S)	T	**V**(S)	T	**A**(S)	T
NKD [188]	32.9	*52.8*	52.2	*31.6*	39.2	*59.4*	59.3	*38.7*
+Cat	34.0	*59.6*	52.0	*60.2*	39.9	*62.9*	59.9	*63.2*
+FiLM [139]	33.2	*57.4*	51.7	*57.6*	39.0	*62.1*	58.6	*62.8*
+OGM [137]	33.6	*60.9*	52.7	*61.6*	40.2	*65.2*	59.8	*64.3*
Ours	34.7	*54.2*	54.9	*34.6*	40.9	*59.0*	61.9	*41.6*

Table 4.9: **Comparison results of different multimodal teachers.** The *italic* numbers mean teachers' accuracy. S: student; T: teacher.

pared to unimodal features, cross-modal features have significant feature divergence, and directly using feature-based distillation methods for CMKD is unreasonable. We leave the exploration of feature-based CMKD as the future work.

Multimodal Teacher Efficacy Analysis. We provide analysis of the efficacy of multimodal teacher in CMKD. Concretely, the multimodal teacher is formulated by fusing the teacher and the student modalities with the supervision loss. Following the multimodal learning [33, 137, 37], we adopt the fusion strategies including Concatenation (Cat), FiLM [139], and OGM [137]. Pretrained teachers are updated for the better information fusion. Table 4.9 indicates multimodal teachers generate high-accuracy soft labels, while don't necessarily improve the distilled modality, especially for the high-accuracy modality (i.e., audio). Our customized teacher integrates receptive corssmodal information and ensures effective knowledge transfer. Inspired by [179], developing the multimodal learning method that contains more modality-general decisive information is a possible solution. We leave this intriguing challenge to future work.

4.7 Chapter Summary

Chapter 4 contribution targets the problem of modality imbalance and misalignment in cross-modal knowledge distillation (CMKD), which often degrades performance when modalities are missing during inference. This chapter reveals that modality imbalance and soft label misalignment in cross-modal knowledge distillation (CMKD) are critical bottlenecks that degrade performance when modalities are missing during inference. Through systematic analysis, we identify the inter-modality gap—divergent feature distributions between modalities (e.g., text vs. vision) as the root cause of distorted knowledge transfer. To address this, we propose Customized Crossmodal Knowledge Distillation (C^{2}KD), specifically, we propose On-the-Fly Sample Selection (OFSD) strategy to filter out rank-distorted samples based on the KRC metric and distill knowledge from non-target classes. Meanwhile, the pre-trained teacher conducts bidirectional distillation with the student. Proxy student and teacher, inheriting unimodal and cross-modal knowledge, progressively transfer cross-modal knowledge. Extensive experiments demonstrate the effectiveness of our method. By resolving knowledge misalignment through adaptive sample selection and proxy-guided distillation, C^{2}KD advances the thesis's theme of robust multimodal integration, extending Chapter 3's compatibility principles to cross-modal distillation while setting the stage for Chapter 5's self-introspective decoding to tackle hallucination in the multimodal inference.

Chapter 5

SID: Self-Introspective Decoding for Modality Prior to Alleviate Hallucinations for Multimodal Large Models

5.1 Challenges and Motivations

Chapters 3 and 4 demonstrate how to improve the robustness abilities of multimodal learning in terms of composition ability and inference with missing modalities situations. In the past few years, large-scale models pre-trained on massive amounts of data and then applied on downstream tasks have become the new paradigm, illustrating the robust and flexible solutions to diverse tasks. Concretely, recent advancements in Large Language Models (LLMs) [160, 7, 26, 161, 126] have demonstrated great success over the past few years. Many efforts have been made to extend LLMs to Multimodal Large Language Models (MLLMs), especially Large Vision-Language Models (LVLMs) [189, 97, 8, 99, 29, 111, 10, 191, 96], achieving impressive performance across various vision tasks [98, 201] as well as more complex tasks like content comprehension [89] and generation [41].

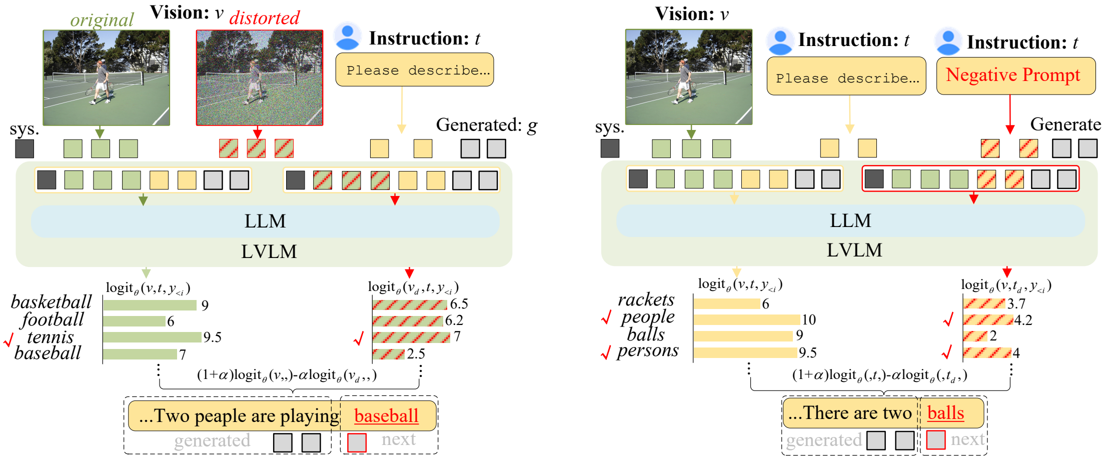
Despite their extraordinary versatility, LVLMs face a significant challenge known as the 'hallucination'. Concretely, hallucinated texts are fluent and semantically coherent but contain incorrect or non-existent statements about the given image, e.g., generating irrelevant or meaningless responses, identifying inaccurate colors, numbers, and locations of objects not present in the image [61]. This flaw in LVLMs poses a significant risk for real-world applications to become trustworthy AI assistants. For instance, in model-assisted computer-aided diagnosis scenarios [168], such misinterpretation of medical images could lead to serious medical accidents.

One mainstream approach to alleviating hallucinations in LVLMs involves developing training-free decoding strategies known as Contrastive Decoding (CD) [94, 38, 170, 84], which adjusts the next-token logits in a contrastive manner. Concretely, Vision CD (VCD) manipulates vision inputs with Gaussian noise [94] or directly ablates visual inputs [38] to amplify language modality priors. Instruction CD (ICD) [170, 84] designs negative prompt.[1] The rationale is that disturbed inputs significantly exacerbate hallucinations, and CD subtracts hallucinated concepts from the original distribution to mitigate hallucinations.

However, input disturbances require elaborate designs for various downstream tasks, and the inference cost is inevitably doubled. *Moreover*, the contrastive distributions are *vision-and-text agnostic*, not necessarily amplify desired hallucinations but sometimes induce potential uncertainty noise for CD. Intuitive examples are illustrated in Figure 5.1, and detailed analyses are in Sec. 5.2.2. In Figure 5.1 (a) and (b), LVLMs directly infer the correct next token from multimodal inputs. For Vision CD, distorted vision input exacerbates hallucinated object logits such as *football* and *basketball*, while the holistic noise suppresses *baseball* to a low logit value. Consequently,

---

[1] negative prompts like `'You are a confused object detector.'` and `'Always respond with the opposite of what you're asked.'` for different tasks.



(a) VCD **(b) ICD**

Figure 5.1: **Contrastive Decoding strategies:** (a) Visual Contrastive Decoding (VCD) [94] **manually** distort vision inputs. (b) Instruction Contrastive Decoding (ICD) [170, 84] also **manually** design noisy instruction (negative prompt). We ablate other modules like the vision encoder and tokenizer for clarity. t: `‘Please describe this image in detail.’`; sys.: system prompt. g: generated text tokens.

VCD might compromise normal decoding. Similarly, for Instruction CD, LVLMs tend to refuse to answer negative prompts in open-end generation task (as seen in Figure 5.3 and 5.4), and also suffer from potential uncertainty noise similar to VCD.

To address the aforementioned issues, we propose a novel decoding strategy called *Self-Introspective Decoding* (SID). Our empirical investigations reveal that pre-trained LVLMs can introspectively assess the importance of vision tokens adaptively, based on preceding vision and text (both instruction and generated) tokens. SID leverages this capability to amplify and then subtract *vision-and-text association* hallucinations by proposing token-level disturbances named Context and Text-aware Token Selection (CT^{2}S) strategy. This strategy induces multimodal contextual hallucinations, rather than aimless ones, by conducting token selection in the early decoder layers.

In summary, this chapter's main contributions are three-fold:

- We re-think CD methods in LVLMs and attribute their failure cases to vision-and-text agnostic input distributions that induce potential uncertainty noise.

- To address this, we propose Self-Introspective Decoding (SID), where the CT^{2}S strategy adaptively amplifies and then subtracts vision-and-text association hallucinations. This approach is grounded in our investigations that pre-trained LVLMs can introspectively assess visual importance informed by preceding tokens.
- Through comprehensive comparisons, we demonstrate that SID generates high-quality texts with fewer hallucinations. Additionally, SID significantly reduces inference cost of contrastive decoding.

5.2 Preliminary and Discussions

In the following, this sub-chapter first illustrates the generation paradigm of LVLMs to facilitate the understanding of SID. We then re-think the contrastive decoding in LVLMs and propose our motivation for SID.

5.2.1 Paradigm of LVLMs Generation

Vision and Language Inputs. The inputs of LVLMs consist of both image (v) and text (t). Generally, the raw images are commonly fed to the visual encoder, and then the cross-model projection module maps vision information into LLMs' input space, which is denoted as vision tokens $v = \{v_1, v_2...v_n\}$ (n is the length of vision tokens). Similarily, text is processed by tokenizer and embedding modules, which is denoted as text tokens $t = \{t_1, t_2...t_m\}$ (m is length of text tokens). Then, the image (v) and text (t) tokens are concatenated as the final input of LLMs.

LVLMs Forward. The backbone networks of LVLMs are pre-trained LLMs like Vicuna [26] and LLaMA 2 [161], parameterized by θ. Given multimodal tokens $\{v, t\}$, LVLMs predict the next token probability (y_i) at i time step in an auto-regressive manner following the methodology of LLMs, over the vocabulary set ν:

$$p(y_i|v,t,y_{<i}) = \text{softmax}(logit_\theta(y_i|v,t,y_{<i})), y_i \in \nu \tag{5.1}$$

Next Token Decoding. After obtaining the next token probability $p(y_i|v,t,y_{<i})$, different decoding strategies are proposed to predict next token. The decoded token is concatenated to the last original input token, for the next round of generation until the end of the generation process.

5.2.2 Re-thinking Contrastive Decoding in LVLMs

Following the seminal works [103] in natural language processing, which introduced the Contrastive Decoding (CD) mechanism to enhance coherence and informativeness by considering the differences between expert and amateur models, various studies have adapted this strategy to LVLMs by distorting the visual or instruction inputs for contrastive purposes. As the vision and instruction contrastive processes are symmetrical, we use visual contrastive decoding as an example. The contrastive decoded probability of next-token (p_{cd}) can be generally formulated as follows:

$$p_{cd}(y_i|v,v_d,t,y_{<i}) = \text{softmax}[(1+\alpha)logit_\theta(y_i|v,t,y_{<i}) - \alpha logit_\theta(y_i|v_d,t,y_{<i})] \tag{5.2}$$

where d and α indicate distortion operation and hyperparameter, respectively. *Generally*, CD methods employ an adaptive plausibility constraint to calibrate the entire output distribution, preventing implausible outputs from the augmented distribution [103, 27, 94, 38, 170, 84, 210]:

$$\begin{gathered}\nu_{token}(y_{<i}) = \left\{y_i \in \nu : p_\theta(y_i|v,t,y_{<i}) \geq \beta \max_\omega p_\theta(\omega|v,t,y_{<i})\right\}, \\ p_{cd}(y_i|v,v_d,t,y_{<i})) = 0, \text{ if } y_i \notin \nu_{token}(y_{<i})\end{gathered} \tag{5.3}$$

where ν and ν_{token} are the output vocabulary and selected tokens. β controls the strength of truncation, with larger β indicating more aggressive truncation that retains only high-probability tokens.

Table 5.1: **Efficacy Analyses on CD strategies** on MSCOCO dataset. The *Random* setting means objects absent from the image are chosen randomly, while the *Adversarial* setting prioritizes co-occurring objects which are not present in the image.

		Greedy		**Sampling**	
Setting	**Method**	Accuracy ↑	F1 Score ↑	Accuracy ↑	F1 Score ↑
Random	Normal	$88.8_{\pm0.05}$	$88.6_{\pm0.08}$	$84.9_{\pm0.03}$	$83.2_{\pm0.01}$
	VCD	$87.8_{\pm0.02}$	$87.9_{\pm0.06}$	87.73	83.28
	w/o Eq. 5.3	-	-	$83.3_{\pm0.04}$	$82.2_{\pm0.02}$
	ICD	$87.9_{\pm0.04}$	$88.1_{\pm0.02}$	$86.9_{\pm0.03}$	$85.2_{\pm0.04}$
	w/o Eq. 5.3	-	-	$82.7_{\pm0.02}$	$81.8_{\pm0.03}$
	Ours	**89.3**$_{\pm0.08}$	**89.5**$_{\pm0.02}$	**88.8**$_{\pm0.03}$	**88.7**$_{\pm0.02}$
	w/o Eq. 5.3	-	-	$87.2_{\pm0.01}$	$88.0_{\pm0.02}$
Adversarial	Normal	$79.3_{\pm0.05}$	$80.9_{\pm0.09}$	$78.7_{\pm0.03}$	$78.9\ _{\pm0.02}$
	VCD	$80.9_{\pm0.06}$	$81.0_{\pm0.04}$	80.88	81.33
	w/o Eq. 5.3	-	-	$76.2_{\pm0.04}$	$76.0_{\pm0.04}$
	ICD	$80.2_{\pm0.03}$	$81.3_{\pm0.01}$	$79.1_{\pm0.02}$	$80.4_{\pm0.04}$
	w/o Eq. 5.3	-	-	$75.4_{\pm0.02}$	$76.4_{\pm0.04}$
	Ours	**83.3**$_{\pm0.07}$	**82.5**$_{\pm0.06}$	**82.6**$_{\pm0.05}$	**82.1**$_{\pm0.06}$
	w/o Eq. 5.3	-	-	$82.2_{\pm0.03}$	$81.9_{\pm0.01}$

However, we argue that manually disturbing raw inputs might not trigger the desired hallucinations, while holistic disturbances will bring uncertainty noise that compromises the normal decoding. To validate our claim, we analyze the performances of normal decoding, VCD, and ICD using the POPE [106] metric, under both **sampling** and **greedy** decoding settings. POPE quantitatively converts the hallucination evaluation into a binary classification problem by using the question format to prompt the model: `‘Is there a <object> in the image?’`, with expected answers being `‘Yes’` or `‘No’`. From Table 5.1, under the **greedy** decoding setting, CD methods improve performance in the *adversarial* setting, which are more challenging as they prioritize co-occurring confusing objects. CD methods achieve this by exacerbating and subtracting hallucinated concepts from the original distribution. However, in *random* settings, where objects absent from the image are chosen randomly and are easily recognized, CD methods slightly underperform normal greedy decoding, which indicates that the correct token logit is somewhat compromised during contrastive decoding. In the **sampling** decoding setting, CD methods clearly outperform the normal sampling decoding. However, CD methods rely on the adaptive plausibility constraint (Equation 5.3) to filter out low-probability tokens. Without Equation 5.3, CD methods are inferior to normal decoding in both *random* and *adversarial* settings, validating that vision-and-text agnostic input distributions induces potential uncertainty noise after Equation 5.2. To address these issues, we propose a decoding strategy named *Self-Introspective Decoding* (SID). SID *adaptively* amplifies *vision-and-text association* hallucinations informed by generated tokens to guide LVLMs in exploring factualness. Details are illustrated in the Sec. 5.3 and Figure 5.2.

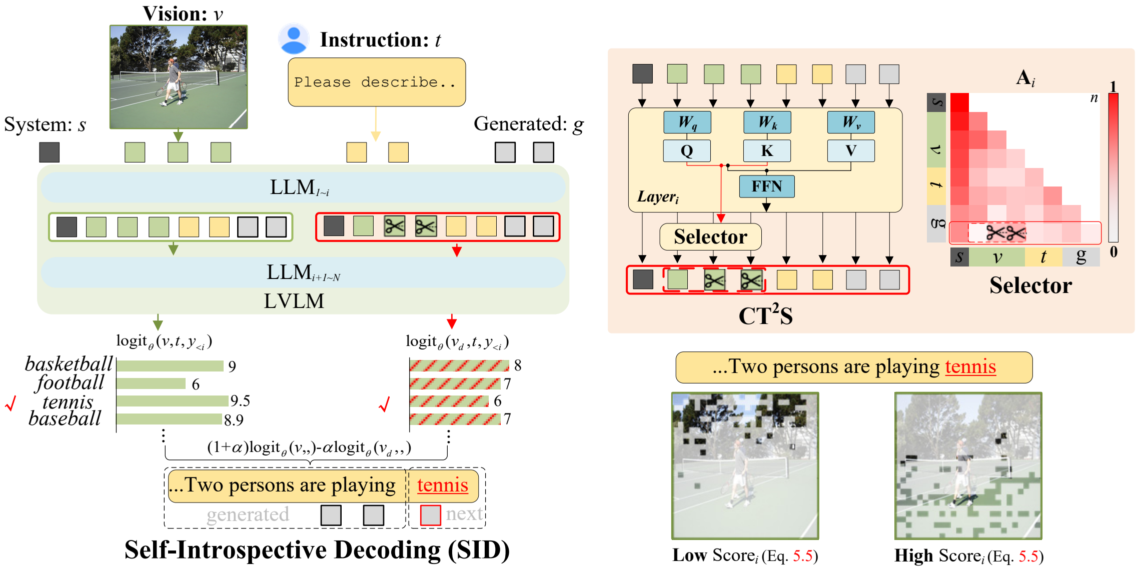


Figure 5.2: **Overview of Self-Introspective Decoding (SID).** CT^{2}S: Context and Text-aware Token Selection strategy. LLaVA-1.5 7B is utilized as an example to visualize visual tokens with low and high scores (Equation 5.5).

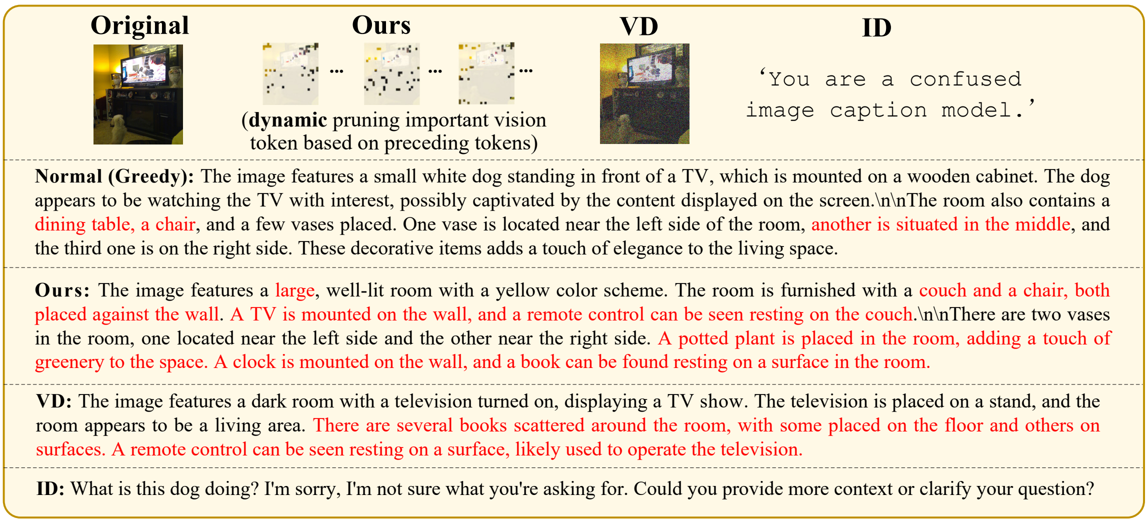
5.3 Self-Introspective Decoding (SID)

5.3.1 Understanding the Self-Introspective Pre-trained LVLMs

LLMs [7, 26, 161, 126] have been scaled up to billions of paramters and pre-trained on trillions of tokens, endowing LLMs with encyclopedic ability like in-context learning [171], zero [86]/few-shot [13] ability. LVLMs extend LLMs to multimodal understanding capabilities by visual instruction tuning. Some works [16, 151, 20] pointed out that vision information is redundant in LVLMs, and develop vision token reduction technologies to prune [142] and merge [12] tokens guided by importance metrics without further re-training. Regarding the hallucination issue, we argue that vision tokens with low attention scores induce **vision-and-text association hallucination**. Formally, for the transformer block [163] in the auto-regressive decoder [2], vision (v),

---

[2] Here we illustrate the transformer block without KV Cache for better understanding.


Original
Ours
(**dynamic** pruning important vision token based on preceding tokens)
VD
ID

`'You are a confused image caption model.'`

Normal (Greedy): The image features a small white dog standing in front of a TV, which is mounted on a wooden cabinet. The dog appears to be watching the TV with interest, possibly captivated by the content displayed on the screen.\n\nThe room also contains a dining table, a chair, and a few vases placed. One vase is located near the left side of the room, another is situated in the middle, and the third one is on the right side. These decorative items adds a touch of elegance to the living space.

Ours: The image features a large, well-lit room with a yellow color scheme. The room is furnished with a couch and a chair, both placed against the wall. A TV is mounted on the wall, and a remote control can be seen resting on the couch.\n\nThere are two vases in the room, one located near the left side and the other near the right side. A potted plant is placed in the room, adding a touch of greenery to the space. A clock is mounted on the wall, and a book can be found resting on a surface in the room.

VD: The image features a dark room with a television turned on, displaying a TV show. The television is placed on a stand, and the room appears to be a living area. There are several books scattered around the room, with some placed on the floor and others on surfaces. A remote control can be seen resting on a surface, likely used to operate the television.

ID: What is this dog doing? I'm sorry, I'm not sure what you're asking for. Could you provide more context or clarify your question?

Figure 5.3: **Instance Illustration of Different Disturbance Results.** Examples are from MSCOCO inferred by LLaVA-1.5 7B with $i = 3$ and Top-k=50. Hallucinations are marked in red.

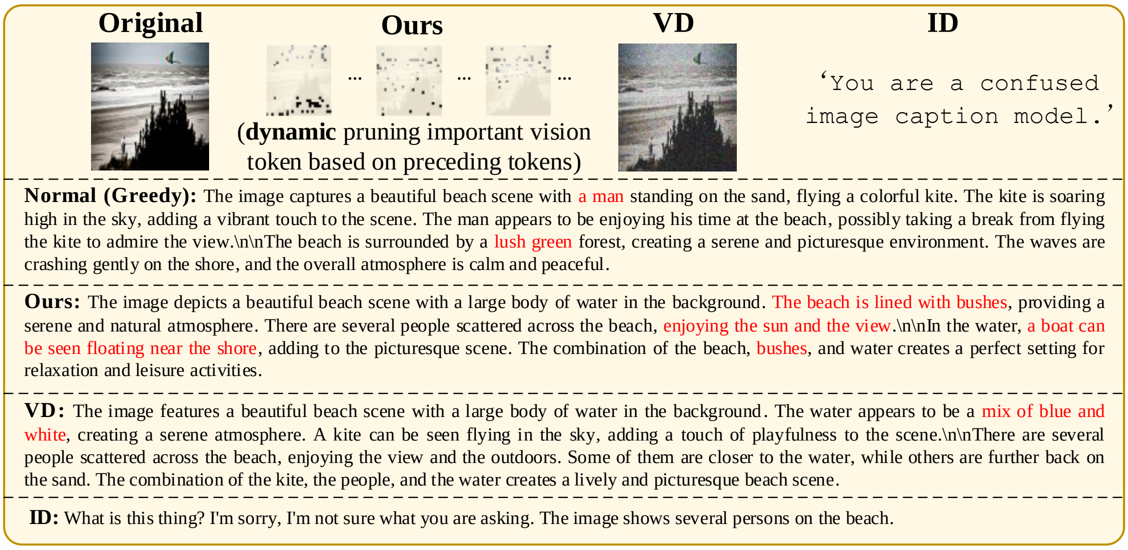
text instruction (t), and generated tokens (g) are concatenated and projected into three distinct vectors: the query vector $\mathbf{Q}$, the key vector $\mathbf{K}$, and the value vector $\mathbf{V}$, utilizing three linear transformations W_q, W_k, and W_v. The self-attention (SA) mechanism computes the relevance of each item to other items as follows:

$$\begin{aligned}\mathbf{R} &= SA(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathbf{A}\cdot\mathbf{V},\\ \mathbf{A} &= \text{softmax}(\frac{\mathbf{Q}\cdot\mathbf{K}^T}{\sqrt{d_l}} + M)\end{aligned} \tag{5.4}$$

where d_l represents the dimension of $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$, M represents the casual mask. $\mathbf{A} \in R^{(b,h,n,n)}$, where b, h, and n denote batch size, number of key-value heads, and total token number, respectively. We denote the $\mathbf{A}_i$ as the attention matrix after *Layer i* of LVLMs. We then calculate vision token importance scores ($\text{Score}_i(v)$) as shown in Figure 5.2 (**Selector**) based on $\mathbf{A}_i$:

$$\text{Score}_i(v) = \frac{1}{h}\sum_{j=1}^{h}\mathbf{A}_i^{(\cdot,j,\cdot,\cdot)}[-1] \tag{5.5}$$

where v means vision token indexes. Contrary to token pruning/merging [142, 12] strategies, we preserve a certain number of the least important vision tokens based


Original
Ours
VD
ID
(**dynamic** pruning important vision token based on preceding tokens)
'You are a confused image caption model.'

Normal (Greedy): The image captures a beautiful beach scene with a man standing on the sand, flying a colorful kite. The kite is soaring high in the sky, adding a vibrant touch to the scene. The man appears to be enjoying his time at the beach, possibly taking a break from flying the kite to admire the view.\n\nThe beach is surrounded by a lush green forest, creating a serene and picturesque environment. The waves are crashing gently on the shore, and the overall atmosphere is calm and peaceful.

Ours: The image depicts a beautiful beach scene with a large body of water in the background. The beach is lined with bushes, providing a serene and natural atmosphere. There are several people scattered across the beach, enjoying the sun and the view.\n\nIn the water, a boat can be seen floating near the shore, adding to the picturesque scene. The combination of the beach, bushes, and water creates a perfect setting for relaxation and leisure activities.

VD: The image features a beautiful beach scene with a large body of water in the background. The water appears to be a mix of blue and white, creating a serene atmosphere. A kite can be seen flying in the sky, adding a touch of playfulness to the scene.\n\nThere are several people scattered across the beach, enjoying the view and the outdoors. Some of them are closer to the water, while others are further back on the sand. The combination of the kite, the people, and the water creates a lively and picturesque beach scene.

ID: What is this thing? I'm sorry, I'm not sure what you are asking. The image shows several persons on the beach.

Figure 5.4: **Instance Illustration of Different Disturbance Results.** Examples are from MSCOCO [109] inferred by LLaVA-1.5 7B with $i = 3$ and Top-k=50. Hallucinations are marked in red.

on Equation 5.5.

Analyses. Figure 5.5 and 5.6 preliminarily validate the efficacy of $\text{Score}_i(v)$ qualitatively. In Figure 5.5, the preserved **least** important tokens mainly reflect areas opposite to the query. For instance, when querying `'cup'` in Figure 5.5 (left), LVLMs focus on `'cup'` in the foreground, thus preserving background tokens with low $\text{Score}_i(v)$. Conversely, LVLMs pay attention to background items when querying `'couch'`. When querying existing items in Figure 5.5 (right), vision tokens of unrelated regions are mainly preserved. For open-end generative tasks in Figure 5.6, auto-regressive decoded tokens are generated based on preceding vision (v), instruction (t), and generated text (g) tokens. The preserved vision tokens are **adaptively adjusted** according to preceding tokens at each decoding step, primarily focusing on spurious related regions. To quantitative analyses the attention score, we select the top-100 and least-100 important vision tokens out of a total of 576 vision tokens of LLaVA-1.5 7B based on attention score based on Equation 5.5. Visual and Instruction Disturbance (VD and ID) are also employed as inputs for analyses. Quantitative results in Table 5.2 illustrate that 100 out of 576 vision tokens with high attention scores greatly maintain original ability, while low attention score tokens reach almost 50% accuracy for the binary classification problem, which indicates attention scores are a good indicator for vision token importance. As for VD and ID, disturbance in raw input does not obviously harm the LVLMs' discrimination ability, as indicated by the POPE metric. However, VD and ID significantly compromise the open-end generation tasks reflected by the CHAIR metric (LVLMs tend to refuse to ID as shown in Figure 5.3 and 5.4). Above evaluations suggest that Equation 5.5 effectively assesses the importance of vision tokens.

We further demonstrate the *open-end generated hallucinations* induced by ours, Vision Disturbance (VD) [94], and Instruction Disturbance (ID) [170] in Figure 5.3 and 5.4. The hallucinations we amplified are more vision-and-text association compared to VD, while LVLMs usually refuse to response to ID. Additionally, we demonstrate the quantitative results for discrimination and generation tasks with *VD and ID as inputs* in Table 5.2. Interestingly, VD and ID do not degrades much especially in discrimination tasks. Experiments imply that disturbed target logits still have the highest probability in most cases, and therefore, contrastive decoded target logits are not enhanced much after Equation 4.2, while CD methods are susceptible to potential uncertainty noise.

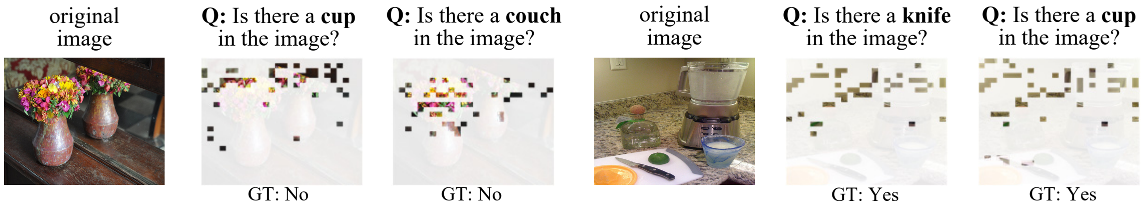

original image
Q: Is there a cup in the image?
Q: Is there a couch in the image?
original image
Q: Is there a knife in the image?
Q: Is there a cup in the image?
GT: No
GT: No
GT: Yes
GT: Yes

Figure 5.5: **Visualization Results** of the **least** important vision tokens on discrimination tasks informed by preceding vision and text tokens. LLaVA-1.5 7B with Layer $i = 3$ is utilized.

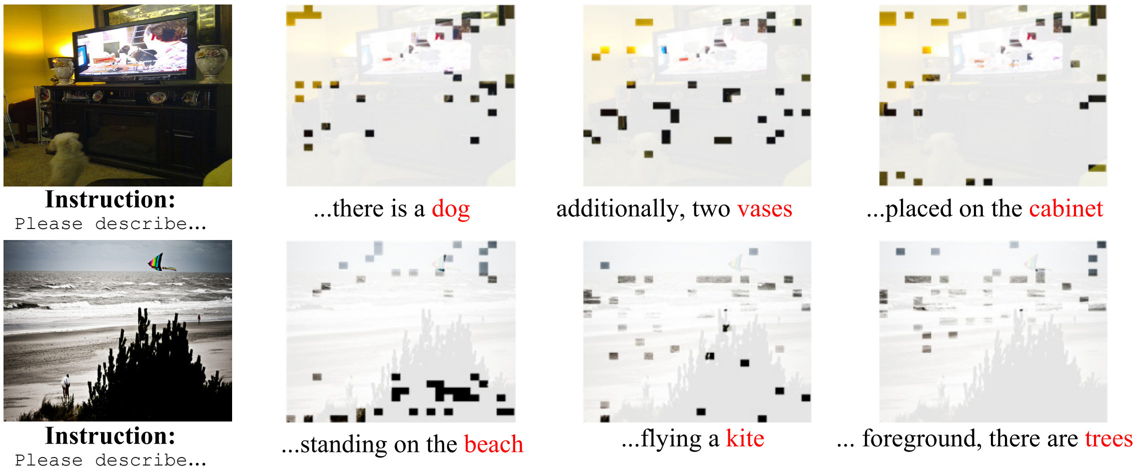


Figure 5.6: **Visualization Results of *Adaptively* Selecting** the **least** important vision tokens on open-end generative tasks informed by preceding vision and text tokens. LLaVA-1.5 7B with Layer $i = 3$ is utilized.

Table 5.2: **Efficacy Analyses on Vision Token Attention Scores** with POPE metric on MSCOCO dataset and CHAIR metric. We select the Top-100 and Least-100 important vision tokens out of a total of 576 vision tokens of LLaVA-1.5 7B, based on Equation 4.5 (i=3). **VD**: Visual Disturbance; **ID**: Instruction Disturbance.

Setting	*Random*		*Adversarial*		CHAIRs ↓	CHAIRi ↓
	Accuracy ↑	F1 Score ↑	Accuracy ↑	F1 Score ↑		
Greedy	88.8	88.6	79.3	80.9	49.6	14.4
+Top-100	85.6	83.9	77.1	76.3	52.7	15.2
+Least-100	55.3	66.1	54.0	65.3	63.2	38.7
+VD	88.0	87.6	78.9	79.8	56.7	16.9
+ID	88.2	87.7	79.1	80.1	-	-

5.3.2 Context and Text-aware Token Selection (CT^{2}S) Strategy

Based on the above investigations, we argue that to induce context- and text-aware hallucinations for contrastive decoding, only a small percentage of vision tokens with low attention scores should be preserved after the early decoder layers. To validate our claims, we conduct the following experiments: 1) In Vision Encoder (VE), we preserve tokens with low attention values between the [**CLS**] token and vision tokens in the penultimate layer, calculated as: $\mathbf{A} = \text{softmax}(\frac{[\mathbf{CLS}] \cdot \mathbf{K}^{\mathbf{T}}}{\sqrt{d_k}})$. 2) In the LLM decoder, we preserve tokens with low importance score (Equation 5.5) across varying layers (i). Additionally, we adjust the number of preserved vision tokens. As shown in Figure 5.7, **firstly**, pruning vision tokens in VE based on [**CLS**] may not always yield positive gains, as the [**CLS**] token lacks information about instructions and generated texts, which are crucial for multimodal understanding. Specifically, pruning all vision tokens resembles VIG [38], which contrastively amplifies the vision importance over the language prior by ablating vision inputs. **Secondly**, aggressive pruning of vision tokens (i.e., 0%) after $Layer_{i=1}$ is not optimal. As the ideal induced hallucination distributions are *target-co-occurring* but suppress *target logits*, the loss of visual information for subsequent decoding results in visual context diminishing, which can lead to aimless hallucinations due to insufficient grounding in visual information. **Thirdly**, selecting tokens in the late decoder layers degrades contrastive decoding to normal decoding, as preceding layers of LVLMs already decode and understand multimodal information, which is consistent with LLMs' early-exiting mechanisms [149, 34]. In summary, the proposed CT^{2}S strategy selects Top-k least important vision tokens after the early layers based on attention score (Equation 5.5), where the induced hallucinations are aware of both visual contexts and text information. Finally, following CD methods (Sec. 5.2.2), we contrastively subtract amplified vision-and-text association hallucinations for the next token prediction.

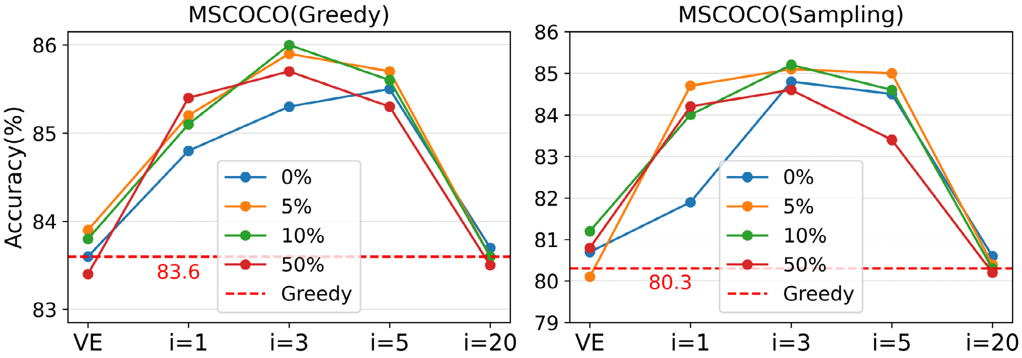


Figure 5.7: Analyses of varying i and preserved ratios in CT^{2}S. VE: vision encoder; i: i-th decoder layer.

Discussion. Based on analyses of the self-introspective pre-trained LVLMs, could we enhance vision information informed by the proceeding vision and text tokens rather than utilizing the contrastive decoding? To explore this, we rewrite Equation 4.2 as follows:

$$p_{add}(y_i|v, v\uparrow, t, y_{<i}) = \text{softmax}[(1-\alpha)logit_\theta(y_i|v, t, y_{<i}) + \alpha logit_\theta(y_i|v\uparrow, t, y_{<i})] \quad (5.6)$$

where we preserve vision tokens with high importance scores (Equation 5.5) denoted as $v\uparrow$. From Table 5.3, we observe that enhancing vision information (i.e., `Add`) alleviates hallucinations to some extent, which also implicitly validates the efficacy of Equation 5.5. However, in the adversarial setting, enhancing vision information does not bring much benefits compared to ours. Because our amplified hallucinations effectively associate co-occurring objects, reflected in **high logit values of hallucination token**, and then contrastively suppress them. In contrast, enhancing vision information primarily boosts the original prediction's **target** logits grounded in attention scores, which does not significantly improve discrimination, especially in the adversarial setting.

Table 5.3: **Analyses of Contrastive Decoding Mechanisms** on the POPE metric. Hyperparameters are consistent with CD settings.

Setting	*Random*	*Adversarial*
Sampling	84.7	78.7
VCD	87.7	80.9
`Add`	**88.9**	79.4
Ours	88.8	**82.6**
Greedy	88.8	79.1
VCD	87.8	80.9
`Add`	89.1	80.1
Ours	**89.3**	**83.3**

5.4 Experimental Results

5.4.1 Experimental Settings

Models and Baselines. We utilize four representative LVLMs: InstructBLIP [29], Shikra [19], LLaVA-1.5 [111] at the 7B scale, and LLaVA-NeXT [96] at the 8B scale. For the detailed model descriptions, InstructBLIP employs Q-former [99] to condense image tokens to 32, as a result, we are unable to visualize the dynamic token pruning process of InstructBLIP like Figure 5.5 and 5.6. Shikra, LLaVA-1.5, and LLaVA-NeXT directly leverage linear projection layers as vision-language connectors to align multimodal features. Shikra and LLaVA-1.5 encode 256 and 576 image tokens to LVLMs. LLaVA-NeXT increases the input vision resolution by 4× to capture more visual details, resulting in 4× more encoded vision tokens than LLaVA-1.5. All LVLMs utilize pre-trained vision encoders like CLIP [141] vision encoder, as well as pre-trained LLMs as language decoders, such as Vicuna v1.1 [26], LLaMA 2 [161], and recently released LLaMA 3 [126]. We provide results at the 7 Billion (B) scale, and larger-scale results are in Table 5.9. Since our method aims to propose *training-free* LVLM decoding strategies *without* the aid of auxiliary networks, we compare six decoding methods: Sampling (Top-p=1), Greedy, Dola [27], and LVLM decoding strategies (VCD [94], ICD [170], and OPERA [61]). For comprehensive comparisons, we apply VCD and ICD in both sampling (Top-p=1) and greedy decoding settings.

Implementation Details. As analyzed in Sec. 5.3.2, we set Layer i=3 and preserve top 10% least important vision tokens for Shikra, LLaVA-1.5, and LLaVA-NeXT and i=5 and top 10% least important vision tokens for Q-former based LVLMs (InstructBLIP) to induce fine-grained hallucinations. For sampling and greedy decoding, we adopt the default hyperparameter settings. As for Dola [27], it is designed to alleviate hallucinations (i.e., improve factuality) of LLM by contrasting the differences in logits obtained from projecting the later layers versus premature layers. Dola is sensitive to the premature layer selection, we adapt Dola to LVLMs, following OPERA [61] to utilize “0,2,4,6,8,10,12,14” as the indexes of candidate premature layers and “32” as the index of the mature layer. The repetition penalty is set to 1.2, as Dola suggests. OPERA, VCD, and ICD are proposed for LVLMs and we adopt the default settings. For fair comparisons, SID’s hyperparameters of Equation 4.2 and 4.3 follow VCD and ICD. Moreover, we apply SID, VCD, and ICD in both sampling (Top-p=1) and greedy decoding settings for comprehensive comparisons. Note that due to amplified fine-grained hallucinations, SID is more robust to hyperparameters compared to other CD methods (Sec. 5.5). Experiments are performed on NVIDIA V100/A100 GPUs.

5.4.2 Evaluation Results

In this section, we follow previous methods [94, 170, 61] to evaluate the SID on **CHAIR** [144] and **POPE** [106] metrics. Besides manually designed metrics, we also leverage **GPT-4 assisted benchmark** [206] to evaluate attribute, location, and relation hallucinations. **MME** [40] and **MMBench** [117] benchmarks are employed

Table 5.4: **Results on the CHAIR metric.** * and * denote adopting the same sampling and greedy decoding strategies, respectively.

Setting	**LLaVA-1.5**		**InstructBLIP**		**Shikra**		**LLaVA-NeXT**	
	$C_S\downarrow$	$C_I\downarrow$	$C_S\downarrow$	$C_I\downarrow$	$C_S\downarrow$	$C_I\downarrow$	$C_S\downarrow$	$C_I\downarrow$
Sampling	51.3	16.8	51.0	24.2	48.9	14.7	42.6	14.1
ICD*	48.7	13.9	48.3	16.7	47.8	14.5	42.7	13.6
VCD*	48.0	14.3	47.9	17.2	48.1	13.8	41.3	12.9
Ours*	45.0	**11.7**	43.6	13.1	46.0	12.9	38.4	11.4
Greedy	49.6	14.4	54.6	13.6	47.1	13.9	42.9	13.2
Dola*	47.1	13.8	52.7	14.0	46.8	14.2	40.9	13.1
OPERA	45.2	12.7	47.4	12.9	**44.4**	13.6	39.4	11.8
ICD*	47.4	13.9	46.3	15.3	47.3	14.1	42.1	12.6
VCD*	46.8	13.2	44.0	13.6	47.8	14.0	41.1	12.9
Ours*	**44.2**	12.2	**42.3**	**12.4**	44.8	**12.8**	**38.1**	**11.3**

to assess the LVLM's general ability. Moreover, **GPT4-V assisted evaluation** on both hallucination alleviation and generated text quality and **Case study** of several pictures are illustrated.

CHAIR and POPE Evaluations. CHAIR [144] and POPE [106] are quantitative metrics to assess objection hallucinations of VLMs. The Caption Hallucination Assessment with Image Relevance (CHAIR) [144] metric is specially designed to assess objection hallucinations in the image caption tasks. Concretely, CHAIR quantifies the degree of hallucinations in a generated image caption by calculating the proportion of all objects mentioned in the caption that are not present in the ground truth label pool. There are two common variants of CHAIR: CHAIRi (C_I) and CHAIRs (C_S), which evaluate the degree of object hallucination in the instance and sentence level, respectively. These two metrics are formulated as follows:

$$C_I = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all mentioned objects}\}|}, \ C_S = \frac{|\{\text{captions with hallucinated objects}\}|}{|\{\text{all captions}\}|} \tag{5.7}$$

The smaller the value of C_I and C_S, the better the hallucination alleviation performance. The Polling-based Object Probing Evaluation (POPE) [106] was recently developed to assess hallucination problems in LVLMs. POPE queries the LVLMs with the template: `Is there a <object> in the image?` The ratio between queries about existing and no-existing objects is balanced (i.e., 50%-50%). This benchmark consists of three sampling settings: *random*, *popular*, and *adversarial*, each differing in the construction of negative samples. Specially, in the *random* setting, objects that are not present in the image are selected at random. The *popular* setting selects missing objects from the high-frequency pool, whereas in the *adversarial* setting, co-occurring objects that are not present in the image are prioritized. POPE consists of three different datasets, including MSCOCO [109], A-OKVQA [150], and GQA [64]. POPE involves 500 images from each dataset with six questions each, ultimately yielding 27,000 query-answer pairs. Accuracy and F1 score are chosen as evaluation metrics. The larger the value of Accuracy and F1 score, the better the hallucination alleviation performance. As for **CHAIR**, Following [170, 61, 197], we randomly select 500 images from the validation set of the MSCOCO [109] dataset and query different LVLMs with the prompt: `‘Please describe this image in detail.’`. We set the max new tokens to 512 to generate responses for fair comparisons. As shown in Table 5.4, our method outperforms other baselines in most cases, validating the effectiveness of SID in open-end generation tasks. Compared to CD methods, SID *online adaptively* prunes attention-important vision tokens informed by instruction and generated text to induce fine-grained hallucinations for contrastive decoding during open-end text generations. For the **POPE** metric, which comprises three datasets, we average the results in Table 5.5. Our method performs best overall in *random*, *popular*, and *adversarial* sampling settings. Specifically, in the sampling decoding setting, SID surpasses the normal sampling decoding by a large margin in a train-free manner.

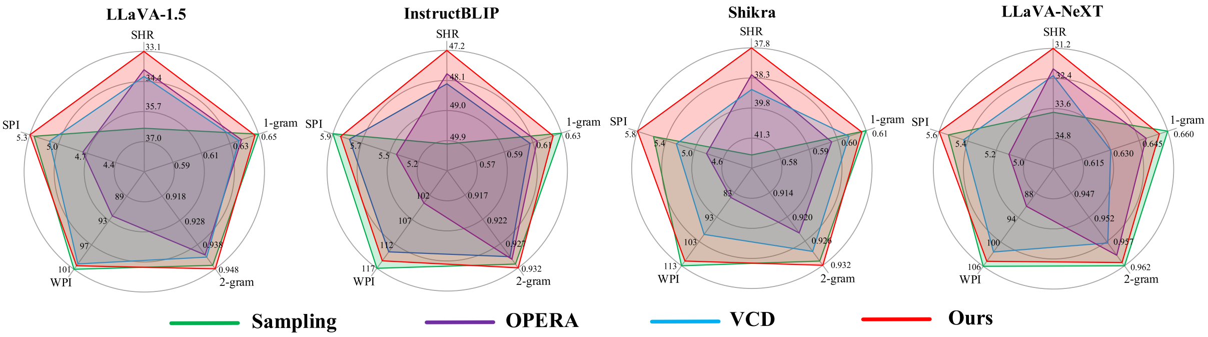


Figure 5.8: **GPT-4 assisted benchmark** [206]. **Hallucination** (SHR), **fluency** (1&2-gram), and **detailness** (WPI and SPI) aspects are compared. Larger areas mean better performances. VCD and ours adopt the same sampling decoding. Please zoom in for details.

SID also clearly outperforms CD methods (Dola, ICD, and VCD) because the self-introspective decoding strategy amplifies *vision-and-text association* hallucinations then subtracts them, rather than coarsely disturbing raw inputs. Additionally, owing to the context and text-aware token selection strategy, SID is more computation-efficient than CD methods, as analyzed in Table 5.8. Note that beam-search based OPERA [61] shows almost no gain in the POPE metric, primarily because answering the binary classification only requires a few tokens and selecting the best beam score in a decoded sequence (N=5) brings little improvement.

GPT-4 Assisted Benchmark. While CHAIR and POPE evaluate object-existence-level hallucinations, these metrics are unable to identify other types of hallucination, such as *positional*, *relational*, and *attribute* hallucinations. Therefore, the GPT-4 assisted benchmark [206] utilizes the fine-grained object-level descriptions in the Visual Genome (VG) dataset [88] as ground-truth and relies on the advanced GPT-4 to judge the fine-grained hallucinations and calculate Sentence-level Hallucination Ratio (SHR). Concretely, besides object-existence-level hallucinations evaluated by CHAIR and POPE, GPT-4 assisted benchmark [206] utilizes the fine-grained object-level description in the Visual Genome (VG) dataset [88] as ground-truth and relies on the

Table 5.5: **Average results on the POPE metric.** * and * denote adopting the same sampling and greedy decoding strategies, respectively. Results are from the original papers or re-implemented based on official codes.

Setting		*Random*		*Popular*		*Adversarial*	
Model	**Decoding**	Accuracy↑	F1 Score↑	Accuracy↑	F1 Score↑	Accuracy↑	F1 Score↑
LLaVA-1.5	Sampling	84.77	82.28	79.98	79.34	76.03	76.26
	ICD*	87.51	83.28	83.15	83.91	79.13	80.41
	VCD*	86.84	86.83	82.65	83.37	77.31	79.28
	Ours*	88.91	88.84	83.97	85.42	82.54	81.98
	Greedy	88.81	88.52	82.76	83.36	79.11	80.92
	Dola⋆	87.94	87.97	83.87	84.68	80.35	81.21
	OPERA	88.85	88.67	82.77	83.40	79.16	80.93
	ICD⋆	87.97	87.84	84.03	84.22	80.21	80.97
	VCD⋆	87.02	86.96	83.53	84.56	78.12	80.16
	Ours⋆	**89.46**	**89.62**	**85.13**	**85.94**	**83.24**	**82.21**
InstructBLIP	Sampling	80.42	80.94	76.09	77.65	72.37	75.42
	ICD*	85.78	85.73	81.12	82.25	76.82	78.99
	VCD*	84.11	84.13	79.96	80.80	76.32	78.08
	Ours*	86.56	85.94	80.26	81.75	77.64	80.41
	Greedy	84.56	83.75	78.23	79.16	74.58	76.34
	Dola⋆	84.67	83.38	78.21	79.19	75.69	77.98
	OPERA	84.57	83.74	78.24	79.15	74.59	76.33
	ICD⋆	84.36	83.82	77.88	78.70	75.17	77.23
	VCD⋆	84.52	83.63	78.04	78.45	75.95	77.76
	Ours⋆	**87.23**	**86.90**	**81.16**	**82.57**	**78.51**	**81.26**
Shikra	Sampling	81.42	82.46	79.60	80.78	73.85	76.39
	ICD*	82.34	82.82	78.17	80.43	74.96	77.68
	VCD*	82.31	82.73	79.34	80.93	75.61	77.96
	Ours*	83.87	83.94	80.26	82.07	77.85	78.94
	Greedy	83.00	83.19	81.39	81.90	76.69	78.31
	Dola⋆	82.87	82.98	**82.42**	82.50	76.85	78.09
	OPERA	83.05	83.20	81.40	81.89	76.73	78.31
	ICD⋆	82.67	82.64	80.73	81.58	75.98	78.43
	VCD⋆	82.96	82.63	80.68	81.27	76.94	78.32
	Ours⋆	**84.46**	**84.62**	82.38	**82.73**	**78.67**	**79.34**
LLaVA-NeXT	Sampling	86.32	83.11	82.27	81.03	77.32	76.96
	ICD*	87.32	84.03	83.62	83.54	80.31	80.41
	VCD*	86.97	86.71	83.07	83.65	79.42	80.28
	Ours*	89.16	88.92	84.38	85.76	82.95	81.98
	Greedy	89.37	88.82	83.68	84.62	80.08	80.74
	Dola⋆	88.73	88.67	84.56	84.96	80.32	80.68
	OPERA	89.36	88.80	83.65	84.60	80.10	80.75
	ICD⋆	87.40	87.96	84.11	83.79	80.94	80.67
	VCD⋆	87.83	87.09	82.68	83.55	79.61	81.20
	Ours⋆	**90.05**	**89.97**	**86.13**	**85.69**	**84.06**	**82.95**

advanced GPT-4 to judge the detailed (such as *positional*, *relational*, and *attribute*) hallucinations and calculate Sentence-level Hallucination Ratio (SHR). With the generated sentences and manually annotated factual information, GPT-4 is prompted to evaluate whether existing hallucinations sentence by sentence. The prompt template is provided in Figure 5.9. Following [206], we utilize 200 images from the VG dataset and set max new tokens to 512, with the prompt of `‘Please describe this image in detail.’` We conduct experiments on sampling decoding strategies and representative LVLMs decoding strategies: VCD [94] and OPERA [61]. Moreover, we employ n-gram fluency (n = 1 and 2) metrics to measure the smoothness of generated text, and the number of generated words/sentences per image (WPI/SPI) to compare the detailedness of generated texts. As shown in Figure 5.8, SID achieves the best results in the SHR metric among the four LVLMs, outperforming others by a clear margin. Regarding the quality of the generated texts, Sampling decoding outperforms ours slightly in terms of 1-gram fluency and WPI. However, compared to other baselines, our approach alleviates hallucinations with minimal sacrifice in text generation quality regarding smoothness and detailness. For instance, OPERA generates text with fewer words and sentences due to penalization of the over-trust mechanism, and VCD impairs text fluency, possibly arising from the holistic and fixed disturbance of contrastive inputs.

MME and MMBench Evaluations. Besides, we test on two popular LVLMs’ general ability benchmarks: MME and MMBench. MME comprises ten subtasks to evaluate models’ perceptual capabilities and four subtasks for assessing recognitive abilities in the form of the yes/no question. MMBench systematically evaluates twenty ability dimensions of LVLMs. We present the results of LLaVA-1.5 7B as a representative in Table 5.7, SID can maintain and improve the multimodal ability on LVLMs benchmarks. In contrast, other CD methods tend to compromise the general multimodal ability.

GPT4-V Assisted Evaluation. To further analyze the hallucinations and text quality for open-end generation tasks, following [61, 190], we utilize the strong multimodal assistant GPT4-V, which simultaneously processes input from vision and text modalities. We strictly follow [61], which utilizes 500 images from the MSCOCO dataset and prompts LVLM:`‘Please describe this image in detail.’` with the maximum number of 512. To mitigate the impact of the sequential order fed to GPT4-V, we simultaneously compare the generated texts obtained from two decoding methods and instruct GPT4-V to judge the correctness and detailedness score on a scale of 0-10 based on the input image. The detailed GPT4-V prompt is in Figure 5.10. We set up three representative pairs of comparison experiments: greedy decoding and ours, CD-based VCD [94] and ours, and OPERA [61] and ours. As shown in Table 5.6, our SID achieves the best results in terms of most metrics. Concretely, our method improves correctness by about 15-20% compared to sampling decoding while not compromising the detailedness level. Compared to advanced hallucination mitigation methods VCD and OPERA, SID generates text with obvious more details and better mitigates the hallucination issue. Since the perceptual and reasoning capabilities of GPT4-V are very close to those of humans, the results of the GPT4-V evaluation reflect, to some extent, the strong performance of the compared methods in terms of mitigating hallucinations and generating text quality from a human perceptual perspective.

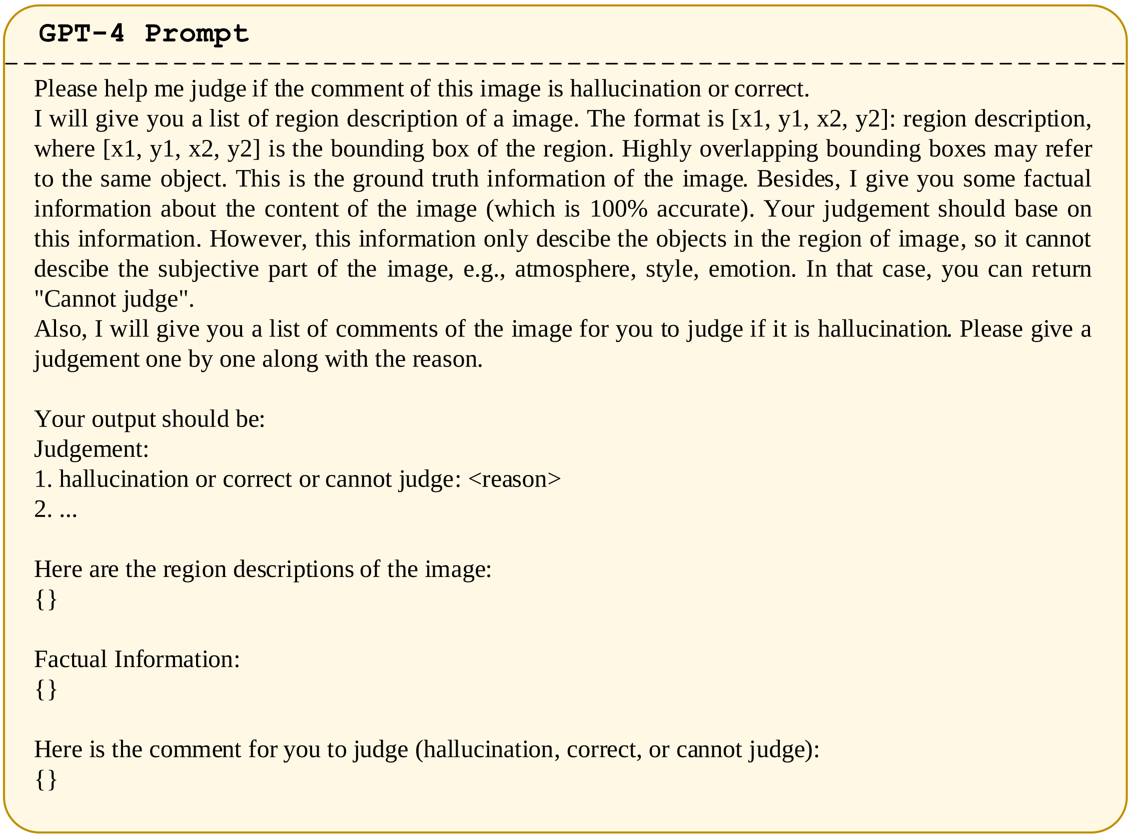
Case Study. In addition to using crafted metrics (CHAIR and POPE), GPT-4/GPT4-V-aided evaluations, and MME [40] and MMBench [117] benchmarks, we qualitatively present several case studies of SID's hallucination alleviation ability from LLaVA-Bench-in-the-Wild dataset [112]. As illustrated in Figure 5.11, 5.12, and 5.13, SID effectively mitigates hallucination in these challenging scenes by dynamically amplifying vision-and-text association hallucinations. Meanwhile, it preserves the detailness of each image. As we propose a training-free decoding method that does not rely on auxiliary analysis networks, it inherently carries over the existing weaknesses of LVLMs. Intuitive case studies, as illustrated in Figure 5.11, 5.12, and

Table 5.6: **GPT4-V assisted hallucination evaluations** [61, 190]. VCD and ours adopt the same sampling decoding strategy. C: correctness; D: detailedness

Setting	LLaVA-1.5		InstructBLIP		Shikra		LLaVA-NeXT	
	$C\uparrow$	$D\uparrow$	$C\uparrow$	$D\uparrow$	$C\uparrow$	$D\uparrow$	$C\uparrow$	$D\uparrow$
Sampling	5.18	5.79	4.73	5.10	5.03	**5.17**	5.34	5.67
Ours	**5.97**	**6.01**	**5.62**	**5.16**	**5.78**	5.10	**6.47**	**5.85**
VCD	5.46	5.63	4.98	5.21	5.31	5.24	5.92	5.47
Ours	**6.16**	**5.9**4	**5.37**	**5.46**	**5.61**	**5.29**	**6.12**	**5.78**
OPERA	**6.16**	5.57	5.29	4.86	5.34	4.87	6.11	5.24
Ours	6.15	**5.94**	**5.76**	**5.42**	**5.97**	**5.88**	**6.63**	**6.23**

Table 5.7: **LVLM benchmark evaluations.** DoLa, ICD, VCD, and SID employ the same greedy decoding.

	Greedy	Sampling	DoLa	ICD	VCD	OPERA	**SID**
MME	1510.8$_{\pm 1.2}$	1471.5$_{\pm 5.6}$	1480.7$_{\pm 1.3}$	1473.2$_{\pm 1.2}$	1488.5$_{\pm 0.8}$	1515.2$_{\pm 1.1}$	**1520.4**$_{\pm 0.9}$
MMbench	64.4$_{\pm .22}$	63.9$_{\pm .81}$	63.7$_{\pm .22}$	63.0$_{\pm .24}$	63.8$_{\pm .22}$	64.4$_{\pm .13}$	**65.0**$_{\pm .23}$



GPT-4 Prompt

Please help me judge if the comment of this image is hallucination or correct.
I will give you a list of region description of a image. The format is [x1, y1, x2, y2]: region description, where [x1, y1, x2, y2] is the bounding box of the region. Highly overlapping bounding boxes may refer to the same object. This is the ground truth information of the image. Besides, I give you some factual information about the content of the image (which is 100% accurate). Your judgement should base on this information. However, this information only descibe the objects in the region of image, so it cannot descibe the subjective part of the image, e.g., atmosphere, style, emotion. In that case, you can return "Cannot judge".
Also, I will give you a list of comments of the image for you to judge if it is hallucination. Please give a judgement one by one along with the reason.

Your output should be:
Judgement:
1. hallucination or correct or cannot judge: <reason>
2. ...

Here are the region descriptions of the image:
{}

Factual Information:
{}

Here is the comment for you to judge (hallucination, correct, or cannot judge):
{}

Figure 5.9: **Prompts of GPT-4 for evaluations.**

5.13, reveal that SID still generates some hallucinations, particularly in finer details such as eye color and vehicle identification specifics. These failures may be attributed to the vision encoder's relatively limited visual perception ability. For future work, it is promising to integrate SID with InternVL [25], which scales the vision encoder up to 6B, or consider leveraging auxiliary analysis networks like Grounding DINO [116] or OWLv2 [128] to mitigate LVLMs' internal weaknesses.

5.5 Ablation Analyses

In this section, we conduct ablation analyses in terms of the **Computation Efficiency** and **Hyperparameter Sensitivity**, **Larger-scale Backbones**, **Other**

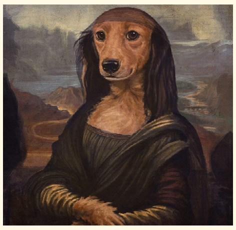
`GPT4-V Prompt`

You are required to score the performance of two AI assistants in describing a given image. You should pay extra attention to the hallucination, which refers to the part of descriptions that are inconsistent with the image content, such as claiming the existence of something not present in the image or describing incorrectly in terms of the counts, positions, or colors of objects in the image. Please rate the responses of the assistants on a scale of 1 to 10, where a higher score indicates better performance, according to the following criteria:
1: Accuracy: whether the response is accurate with respect to the image content. Responses with fewer hallucinations should be given higher
scores.
2: Detailedness: whether the response is rich in necessary details. Note that hallucinated descriptions should not count necessary details.
Please output the scores for each criterion, containing only two values indicating the scores for Assistant 1 and 2, respectively. The two scores
are separated by a space. Following the scores, please provide an explanation of your evaluation, avoiding any potential bias and ensuring that
the order in which the responses were presented does not affect your judgment.
[Assistant 1]
{}
[End of Assistant 1]
[Assistant 2]
{}
[End of Assistant 2]
Output format:
Accuracy: <Scores of the two answers>
Reason:
Detailedness: <Scores of the two answers>
Reason:

Figure 5.10: **Prompts of GPT4-V for evaluations.**

Decoding Strategies, and **Visual Enhancing Decoding Strategy**.

Computation Efficiency. One primary concern of hallucination alleviation decoding methods is the computational burden. We evaluate the whole dataset inference time (seconds) and peak GPU memory (MB) on the LLaVA-1.5 7B under the POPE adversarial setting, as shown in Table 5.8. Contrastive Decoding (CD) methods [94, 170] involve constructing distorted raw inputs, resulting in **twice** the inference complexity. OPERA [61] is based on beam-search decoding and maintains a set of beams to enlarge the candidate range. Additionally, roll back mechanism in the retrospection-reallocation strategy further exacerbates computational complexity.



Instruction

(max new tokens=512)

`Please describe this image in detail.`

Greedy: The image features a dog wearing a gown, possibly painted as a mona lisa painting. The dog is looking at the camera with a curious expression, displaying its unique attire. The gown appears to be a burgundy color and complements the dog's outfit.\n\nThe setting of the painting includes a pond and a rock formation in the background, adding a sense of depth and atmosphere to the artwork. Additionally, there is another small figure, possibly a dog or a human, in the upper-left corner of the picture.

OPERA: The image features a painting of a dog wearing a Renaissance-style dress, with a hood over its head. The dog appears to be a poodle, and it is sitting in a chair, giving the impression of a human-like figure. The painting is quite detailed, capturing the dog's facial expression and the intricate design of the Renaissance dress. The dog's attire and the overall composition of the painting create a unique and interesting visual experience.

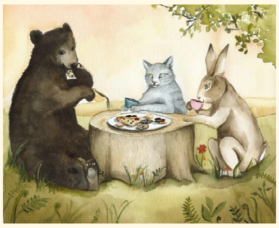
VCD: The image is a painting of a woman wearing a Renaissance-style dress and a hood, resembling a dog. The woman has a dog's face painted on her forehead, and she is wearing a hat. The painting captures the unique and whimsical blend of human and canine features.\n\nIn the background, there is a boat visible, adding a sense of depth and context to the scene. The woman's attire and the dog's face on her forehead create a playful and imaginative atmosphere.

Ours: The painting features a dog wearing Renaissance paint makeup and a veil, posing as a Mona Lisa-style figure. The dog's face is adorned with a mustache and hat, giving it a unique and artistic appearance. The dog's sitting position is also painted to resemble a human's, adding to the creative and humorous nature of the painting.\n\nIn the background, there is a river visible, which complements the overall composition of the painting. The dog's position and the artistic elements create a captivating and amusing scene.

Figure 5.11: **Case Study from LLaVA-Bench-in-the-Wild** on LLaVA-1.5 7B. Hallucinations are marked in red.

Our SID induces vision-and-text association hallucinations by pruning *large-ratio* attention-important tokens in the *early layers*, which greatly reduces the inference time of CD up to ∼30%.

Hyperparameter Sensitivity. Beyond the sensitivity analyses in Figure 5.7, we validate the robustness of SID concerning α and β of Equation 4.2 and 4.3, compared to the contrastive decoding methods (i.e., VCD) on LLaVA-1.5 7B. From Figure 5.14 (left), it is evident that as α decreases, the contrastive decoding mechanism diminishes. However, SID still achieves pleasant results, while VCD degrades close to Sampling when α=0.1, as the CT^{2}S strategy induces informative *vision-and-text association* hallucinations. When α increases, VCD degrades to some extent because holistic input disturbance does not always trigger contextual-related hallucination and might exacerbate uncertainty noise. Regarding β, a larger β indicates more



Instruction

(max new tokens=512)

`Please describe this image in detail.`

Greedy: The image features a group of three animals sitting around a dining table, enjoying a meal together. The animals include a large brown bear, a cat, and a rabbit. They are all seated on the grass, with the bear on the left side, the cat in the middle, and the rabbit on the right side of the table.\n\nThe table is set with various food items, including a cake, a bowl, and a cup. The cake is placed in the center of the table, while the bowl and cup are positioned closer to the edge. The scene appears to be a whimsical and playful depiction of the animals sharing a meal.

OPERA: The image depicts a whimsical scene of a bear, a cat, and a rabbit sitting around a dining table, enjoying a meal together. The bear is on the left side of the table, the cat is in the middle, and the rabbit is on the right side. Each animal has a plate of food in front of them.\n\nThe table is set with various utensils such as forks, knives, and spoons, as well as cups and bowls. There are also a couple of cakes placed on the table, adding to the delightful atmosphere of the gathering.\n\nIn the background, a potted plant can be seen, adding a touch of greenery to the scene.

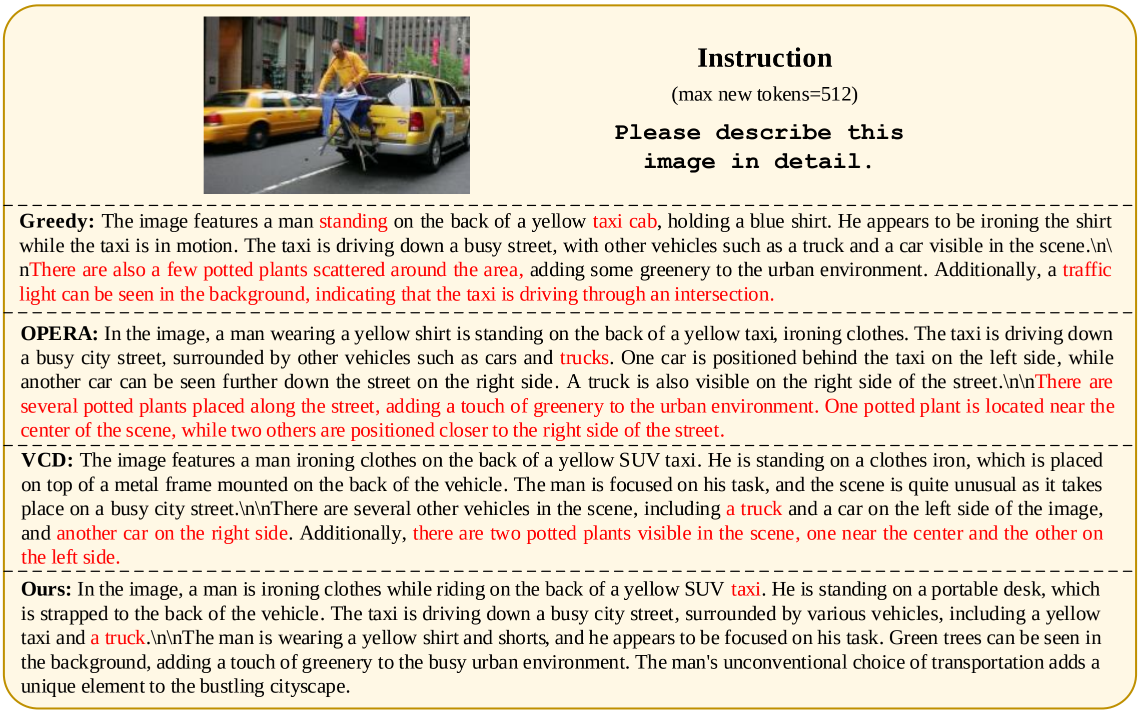
VCD: The image features a whimsical scene of three animals sitting around a dining table, enjoying a meal together. A brown bear is sitting on one side of the table, while a cat is seated on the opposite side, and a rabbit is sitting on the far right side of the table.\n\nVarious food items are placed on the table, including a cup, a bowl, and several cakes. The cakes are spread across the table, with one near the center, another closer to the rabbit, and the third one near the cat. The table is set in a lush green field, giving the scene a peaceful and inviting atmosphere.

Ours: The watercolor illustration features three animals sitting around a dining table, enjoying cookies and doughnuts. A large brown bear is on one side of the table, while a gray cat and a yellowish rabbit are on the other side. The cat has blue eyes and is drinking from a blue cup, while the rabbit holds a pink cup as well.\n\nThe table is topped with a plate of various desserts, including multiple cookies and doughnuts. The scene is set in a field surrounded by green grass and trees, giving the whole watercolor a natural and whimsical atmosphere.

Figure 5.12: **Case Study from LLaVA-Bench-in-the-Wild** on LLaVA-1.5 7B. Hallucinations are marked in red.

aggressive truncation of the output vocabulary. Figure 5.14 (right) shows that VCD's performance heavily relies on large β to retain only high-probability tokens. With mild or no adaptive plausibility constraint (Equation 4.3), VCD performs worse than the sample decoding strategy due to output logits influenced by distorted visual inputs. Ours is robust to the β setting as the CT^{2}S strategy induces discriminative contrastive logits to generate plausible tokens.

Larger-scale LVLM Backbones. We validate the effectiveness of SID in terms of 13B scale backbones on LLaVA-1.5 and InstructBLIP architectures. We choose POPE [106] and CHAIR [144] to validate the hallucination issues in both discrimination and open-end generation tasks. Table 5.9 shows that SID remains effective as backbone networks scale up.



Instruction

(max new tokens=512)

`Please describe this image in detail.`

Greedy: The image features a man standing on the back of a yellow taxi cab, holding a blue shirt. He appears to be ironing the shirt while the taxi is in motion. The taxi is driving down a busy street, with other vehicles such as a truck and a car visible in the scene.\n\nThere are also a few potted plants scattered around the area, adding some greenery to the urban environment. Additionally, a traffic light can be seen in the background, indicating that the taxi is driving through an intersection.

OPERA: In the image, a man wearing a yellow shirt is standing on the back of a yellow taxi, ironing clothes. The taxi is driving down a busy city street, surrounded by other vehicles such as cars and trucks. One car is positioned behind the taxi on the left side, while another car can be seen further down the street on the right side. A truck is also visible on the right side of the street.\n\nThere are several potted plants placed along the street, adding a touch of greenery to the urban environment. One potted plant is located near the center of the scene, while two others are positioned closer to the right side of the street.

VCD: The image features a man ironing clothes on the back of a yellow SUV taxi. He is standing on a clothes iron, which is placed on top of a metal frame mounted on the back of the vehicle. The man is focused on his task, and the scene is quite unusual as it takes place on a busy city street.\n\nThere are several other vehicles in the scene, including a truck and a car on the left side of the image, and another car on the right side. Additionally, there are two potted plants visible in the scene, one near the center and the other on the left side.

Ours: In the image, a man is ironing clothes while riding on the back of a yellow SUV taxi. He is standing on a portable desk, which is strapped to the back of the vehicle. The taxi is driving down a busy city street, surrounded by various vehicles, including a yellow taxi and a truck.\n\nThe man is wearing a yellow shirt and shorts, and he appears to be focused on his task. Green trees can be seen in the background, adding a touch of greenery to the busy urban environment. The man's unconventional choice of transportation adds a unique element to the bustling cityscape.

Figure 5.13: **Case Study from LLaVA-Bench-in-the-Wild** on LLaVA-1.5 7B. Hallucinations are marked in red.

Adopting Other Decoding Strategies. Meanwhile, besides direct sampling and greedy decoding, we conduct experiments on LLaVA-1.5 7B using the MSCOCO dataset with various decoding strategies, including Top-p sampling (p=0.9), Top-k sampling (k=50), Top-k sampling with varying temperature (k=50, t=1.5 and 0.8). Figure 5.15 shows that, regardless of the sampling strategy adopted, the application of SID consistently helps to alleviate hallucinations and improve the overall performance of LVLMs. This consistency highlights the versatility and effectiveness of SID across different sampling strategies.

Visual Enhancing Decoding Strategy. Although LVLMs can accurately recognize visual elements, LVLMs have difficulty fully interpreting those elements in the context of the input cue and effectively linking that recognition to their internal knowledge. We follow Visual Description Grounded Decoding (VDGD) [42] by first gener-

Methods	**Time** ↓	**Memory**↓	**Accuracy**↑
Normal	494	15673	79.11
VCD	904	16753	78.12
ICD	974	16843	80.21
OPERA	2643	21943	79.16
Ours$_{40\%}$	704	15809	83.11
Ours$_{10\%}$	668	15767	83.24

Table 5.8: **Efficiency Comparisons** on NVIDIA V100. $_{10\%}$ and $_{40\%}$ mean tokens preserved ratios.

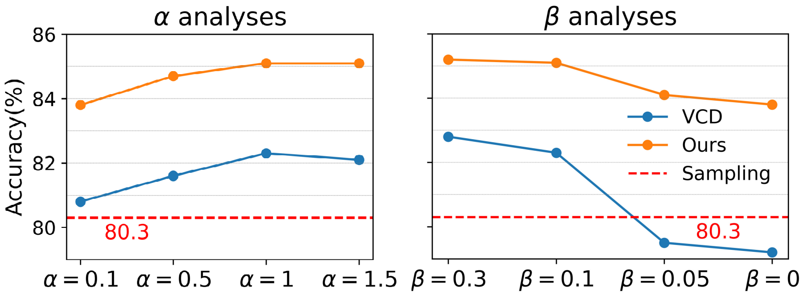


Figure 5.14: **Hyperparameter Sensitivity of** α **and** β with POPE metric (under the sampling decoding).

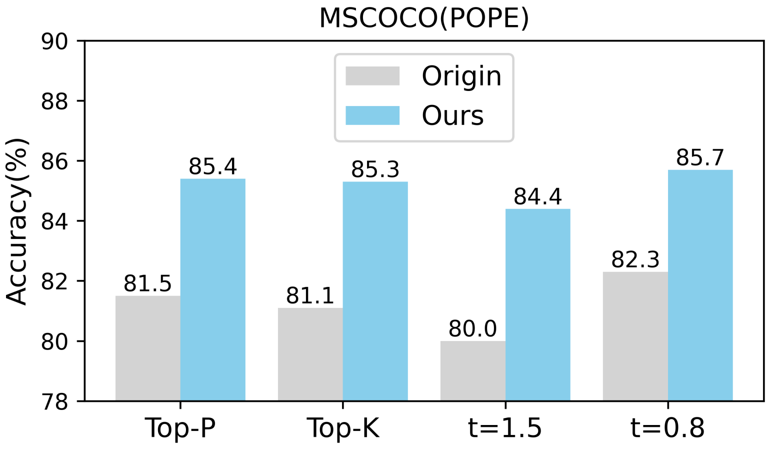


Figure 5.15: **Results of different decoding strategies.**

Table 5.9: **Results on Larger-scale Backbones.** Sampling decoding is adopted and results average of three running times.

	POPE		**CHAIR**	
Methods	Accuracy	F1 Score	C_S	C_I
LLaVA-1.5	81.60	80.31	49.6	16.1
+VCD	82.67	81.46	46.7	16.4
+OPERA	82.32	81.10	**43.3**	13.6
+Ours	**84.75**	**83.17**	43.5	**12.7**
InstructBLIP	77.26	79.23	50.8	19.7
+VCD	79.77	80.27	47.9	17.6
+OPERA	80.31	80.91	42.5	14.3
+Ours	**81.97**	**82.21**	**41.7**	**13.3**

ating a detailed description of the image and appending it as a prefix to the instruction. The prompt template is adopted from [42]: `<image> I have been given this image to complete the task described as: inst. To help me complete the task, describe the given image in detail. In the case of real-world scenes, please include all foreground and background objects in the description, their properties (like color, shape, etc.), their relations with other objects, their count, and all other components in the image. In case of non-real-world scenes, like charts, graphs, tables, etc., please describe the table, mention all numbers (if any), mention the written text, and all other details.` Experiments are performed on hallucination evaluation benchmarks(i.e., POPE and CHAIR) and the general ability benchmark (i.e., MMbench). We re-implement VDGD based on official codes on LLaVA-1.5 7B. Table 5.10 demonstrates the effectiveness of VDGD [42] in LVLM's hallucination alleviation and general reasoning ability. However, the grounding visual descriptions, generated by LVLMs themself, enhance the visual perception reasoning capabilities while might inevitably contain hallucinations. Therefore, VDGD is inferior in the POPE (**adversial**) subset, which prioritizes **co-occurring** objects which are not present in the image. Meanwhile, VDGD shares somewhat similar motivations in enhancing vision information via Equation 4.6 as we analyzed. The experiments in Table 5.3 are consistent with the above results, indicating that boosting the vision information is effective in mitigating hallucinations but is less effective in complex (i.e., adversarial) environments.

Table 5.10: **Comparisons with Visual Enhancing Strategy (VDGD).** * denotes employing greedy decoding strategy.

	POPE↑ (random)	POPE↑ (adversarial)	CHAIRs↓	CHAIRi↓	MMbench↑
Greedy	88.8	79.1	49.6	14.4	64.4
OPERA	88.9	79.2	45.2	12.7	64.4
VDGD*	89.0	79.4	46.7	13.7	**65.2**
SID*	**89.3**	**83.3**	**44.2**	**12.2**	65.0

5.6 Discussion

The core motivation is based the Context and Text-aware Token Selection (CT^{2}S) strategy. Here, we further analyze the efficacy of token selection strategies. Concretely, To validate the effectiveness of SID in selecting low attention scores to induce vision-and-text association hallucination, we further conduct quantitative experiments under different vision token selection strategies with the same preserved vision token number and Layer i=3 as ours. Table 5.11 shows that vision tokens with high attention scores degrade obviously, as it does not amplify contextual hallucinations rather than retain original important information. Contrastive decoding does not benefit from subtracting hallucinations amplified by the disturbed inputs rather than suffers from the potential disturbance noise. Selecting random vision to-

Table 5.11: **Analyses of Different Token Selection Strategies** with POPE on MSCOCO dataset and CHAIR metrics. We select the high importance scores (Equation 5.5) of vision tokens (-**Top**) and random vision tokens (-**Random**) for contrastive decoding. Experiments are conducted on LLaVA-1.5 7B.

Setting	*Random*		*Adversarial*		CHAIRs ↓	CHAIRi ↓
	Accuracy ↑	F1 Score ↑	Accuracy ↑	F1 Score ↑		
Greedy	88.8	88.6	79.3	80.9	49.6	14.4
Ours	89.3	89.5	83.3	82.5	44.2	12.2
-High	87.0	87.3	76.5	79.4	57.9	25.6
-Random	88.4	87.2	80.9	81.5	48.6	13.5
AVISC	88.4	88.1	79.8	80.5	45.3	14.7
Sampling	84.9	83.2	78.7	78.9	51.3	16.8
Ours	88.8	88.7	82.6	82.1	45.0	11.7
AVISC	87.9	87.9	77.5	79.6	46.6	12.5

kens brings improvements in the *adversarial* setting because randomly selected vision tokens amplify the over-reliance on statistical bias and language priors, similar to Vision CD [94] and Instruction CD [170]. However, token-level random disturbance also induces uncertainty noise, resulting in the inferior performance in the *random* setting to greedy decoding. Moreover, AVISC [172], in contrast to ours, preserves outlier high attention tokens (named 'blind token') and substracts output logits to counteract the overemphasis of 'blind token.' In this way, AVISC promotes balanced consideration of all tokens to alleviate hallucinations. However, Table 5.11 illustrates that Top-100 vision tokens with high attention scores can largely maintain the original performance. 'blind token' tends to have a high probability of target class logits, and contrastive decoding does not improve the target class's probability while might bring extra noise. Table 5.11 indicates AVISC still degrades the greedy decoding to some extent, which indicates the attentional vision re-calibration strategy of AVISC induces some annoying noise. Overall, these experiments further validate the rationality of our token selection strategy based on attention sores.

5.7 Chapter Summary

In Chapter 5, we firstly re-think contrastive decoding in LVLMs and empirically find that vision-and-text-agnostic input disturbances in CD do not always amplify desired hallucinations rather than induce potential uncertainty noise. To mitigate these issues, we propose a training-free decoding strategy named Self-Introspective Decoding (SID). By developing Context and Text-aware Token Selection (CT^2S) strategy, SID amplifies *vision-and-text association* hallucinations to guide LVLMs in contrastive decoding, thereby improving faithfulness. Extensive experiments validate the effectiveness and robustness of SID. Aligned with the thesis's focus on robust open-world reasoning, SID's self-corrective mechanism extends Chapter 3's cross-primitive compatibility robustness and Chapter 4's crossmodal knowledge transfer harmonization to introspective hallucination suppression of multimodal LLMs. As for Future Work: **1)** As the pruning ratios and layer are set manually, we consider training the external network to automatically determine optimal hyperparameters, inspired by [22]. In addition, to enhance the interpretability of hallucination alleviations, we consider resorting to pre-trained analysis networks to intuitively locate spurious related vision regions. **2)** Moreover, given that SID amplifies fine-grained hallucinations, we consider leveraging the CT^2S strategy to automatically construct high-quality negative instruction for robust visual instruction tuning rather than relying on expensive GPT-4 [110, 206]. Note that the self-generated hallucination dataset ensures *style consistency*, which is crucial for preference learning [206].

Chapter 6

Conclusion and Suggestions for Future Research

6.1 Conclusion

Despite the success of neural networks, modern neural networks remain inadequate for open-world deployment due to limitations in flexibility, multimodal robustness, and trustworthiness. This thesis bridges these gaps by addressing three critical challenges in multimodal learning: (1) compositional robustness in multiple modality primitives, (2) efficient cross-modal knowledge transfer under modality incompleteness, and (3) hallucination suppression in multimodal large language models (LLMs). The research framework is illustrated in Figure 1.1. In summary, this thesis is mainly composed of three following parts:

- To enhance modality composition generalization robustness, we revisit the primitive prediction approach and propose a novel method, termed Progressive Cross-primitive Compatibility (ProCC), to mimic the human learning process for OW-CZSL tasks. Specifically, the cross-primitive compatibility module explicitly learns to model the interactions of state and object features with the trainable memory units, which efficiently acquires cross-primitive visual attention to reason high-feasibility compositions, without the aid of external knowledge. Moreover, to alleviate the invalid cross-primitive interactions, especially for partial-supervision conditions (pCZSL), we design a progressive training paradigm to optimize the primitive classifiers conditioned on pre-trained features in an easy-to-hard manner. Extensive experiments on three widely used benchmark datasets demonstrate that our method outperforms other representative methods on both OW-CZSL and pCZSL settings by large margins.

- To ensure robustness under modality missing, we focus on studying crossmodal knowledge distillation to handle modality-missing situations. We empirically reveal that the modality gap, i.e., modality imbalance and soft label misalignment, incurs the ineffectiveness of traditional KD methods in CMKD. As a solution, we propose a novel **C**ustomized **C**rossmodal **K**nowledge **D**istillation (C^{2}KD). Specifically, to alleviate the modality gap, the pre-trained teacher performs bidirectional distillation with the student to provide customized knowledge. The On-the-Fly Selection Distillation(OFSD) strategy is applied to selectively filter out the samples with misaligned soft labels, where we distill cross-modal knowledge from non-target classes to avoid the modality imbalance issue. To further provide receptive cross-modal knowledge, proxy student and teacher, inheriting unimodal and cross-modal knowledge, is formulated to progressively transfer cross-modal knowledge through bidirectional distillation. Experimental results on audio-visual, image-text, and RGB-depth datasets demonstrate that our method can effectively transfer knowledge across modalities, achieving superior performance against traditional KD by a large margin.

- To balance modality priors to mitigate hallucinations of multimodal LLMs, we first re-think contrastive decoding in LVLMs and empirically find that vision-and-text-agnostic input disturbances in CD do not always amplify desired hallucinations rather than induce potential uncertainty noise. To mitigate these issues, we propose a training-free decoding strategy named Self-Introspective Decoding (SID). By developing Context and Text-aware Token Selection (CT^{2}S) strategy, SID amplifies *vision-and-text association* hallucinations to guide LVLMs in contrastive decoding, thereby improving faithfulness. This train-free approach reduces hallucinations by 12–20% on metrics like POPE and CHAIR while cutting inference costs by 30% compared to methods like VCD [94] and ICD [170]. Crucially, SID preserves LVLMs' general abilities, as evidenced by strong MME and MMBench scores. By rebalancing modality priors without compromising functionality, SID advances modality-level robustness, ensuring trustworthy outputs in open-world deployment.

6.2 Suggestions for Future Research

Developing robust machine learning for multiple modalities is not a trivial task. Beyond the above challenges, the open-world ever-changing environments also have has challenges such as adversarial adversarial shifts, out of distribution, novel class discovery, and more. To achieve more flexible and efficient machine learning across multiple modalities, future research can explore at least the following three directions:

6.2.1 Multimodal Test-time Adaptation

The degradation of multimodal inputs under extreme environmental conditions (e.g., night, snowy, or foggy settings) introduces severe cross-modal misalignment, where modality-specific corruption patterns (e.g., obscured visuals vs. stable LiDAR signals) create imbalanced feature distributions. Traditional multimodal Test-time Adaptation (TTA) methods [184, 153, 14] partially mitigate this by globally tuning network parameters, yet they overlook two fundamental issues: fine-grained modality interactions and cross-modal consistency preservation. The first issue emphasizes the heterogeneous degradation levels across modalities (e.g., 73% pixel loss in RGB vs. 12% point cloud sparsity in LiDAR) demand dynamic modality weighting rather than uniform adaptation. The latter issue underlines current approaches that fail to enforce semantic coherence between corrupted and intact modalities during adaptation, risking error propagation. For future work, I think the robust multimodal test-time adaptation frameworks must prioritize *hierarchical* adaptation frameworks that dynamically estimate modality-specific and modality-general degradations via entropy minimization, enforce semantic consistency through contrastive alignment of robust primitives (e.g., edge features in vision, spectral peaks in audio), and recalibrate fusion weights in real-time based on modality reliability scores. Such approaches would enable models to adaptively balance multimodal information under open-world volatility, bridging the gap between controlled laboratory performance and real-world resilience.

6.2.2 Task-Aware Adaptation of Multimodal LLMs

The rapid evolution of large-scale vision-language models (VLMs) such as CLIP [141], Flamingo [3], and LLaVA [112] has unlocked unprecedented zero-shot generalization capabilities, positioning them as foundational tools for open-world multimodal learning. However, adapting these models to downstream tasks—from specialized domains like medical diagnostics to dynamic environments like autonomous navigation—requires overcoming critical barriers: the inherent misalignment between their pretraining objectives (e.g., generic image-text matching) and task-specific goals (e.g., fine-grained anomaly detection), the computational impracticality of full fine-tuning for billion-parameter architectures, and the inability to dynamically prioritize modalities based on contextual relevance (e.g., emphasizing thermal over RGB data in low-light conditions). To harness their potential, future work must innovate task-aware adaptation frameworks that inject domain knowledge through lightweight neural modules, optimize modality interactions via dynamic gating mechanisms, and enforce causal invariance to suppress spurious correlations. Early successes, such as adapter-based tuning [58, 46, 114] methods improving rare-class detection in satellite imagery by a large margin, underscore the viability of such approaches. Yet achieving human-level adaptability—where a single LVLM seamlessly transitions from interpreting ambiguous medical queries to generating context-aware robot instructions—demands unifying their embodied knowledge with symbolic reasoning, bridging the gap between static pretrained information and the dynamic demands of downstream deployments. This direction not only amplifies the utility of foundation models but also advances the thesis's broader vision of robust, open-world multimodal systems.

6.2.3 Multimodal Agent as Experts

Large Language Models (LLMs), with their expansive parameter scales and encyclopedic knowledge, are revolutionizing how neural networks adapt to complex tasks. By leveraging their deep understanding of contextual relationships and procedural reasoning, LLMs can dynamically guide the optimization, regularization, and architectural adjustments of neural networks. This capability enables adaptive learning systems that respond intelligently to diverse data distributions, resource constraints, and performance objectives. Recent advancements in LVLMs-empowered multimodal agents, such as those by [176], [115], and [199], have demonstrated remarkable proficiency in tool utilization, embodied AI, and cross-modal reasoning. For instance, Llava-plus [115] showcases LVLMs' ability to interpret visual inputs, generate executable code for robotic manipulation, and self-correct actions through iterative feedback. Similarly, Appagent [199] highlights how multimodal agents can autonomously navigate mobile interfaces by combining screen comprehension, language parsing, and gesture prediction. These innovations underscore the potential of LVLMs to act as versatile "experts" capable of orchestrating intricate workflows across domains. For future work, we want to propose a paradigm where multimodal agents serve as *intelligent coordinators* for multimodal learning systems, enhancing their flexibility and trustworthiness. As coordinators, these agents perform three critical roles. For example, By analyzing real-time inputs (e.g., visual, textual, sensor data), LVLMs agents autonomously reconfigure model architectures, select optimal pretrained sub-networks, or adjust hyperparameters to match evolving task requirements. With the development of multimodal agents, I believe that human developers are totally free of designing complex algorithms for the multimodal learning system.

References

[1] Mahdi Abavisani, Liwei Wu, Shengli Hu, Joel Tetreault, and Alejandro Jaimes. Multimodal categorization of crisis events in social media. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[2] Firoj Alam, Ferda Ofli, and Muhammad Imran. Crisismmd: Multimodal twitter datasets from natural disasters. Proceedings of the AAAI conference on artificial intelligence, 2018.

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems, 35:23716–23736, 2022.

[4] Yuval Atzmon, Jonathan Berant, Vahid Kezami, Amir Globerson, and Gal Chechik. Learning to generalize to new compositions in image understanding. arXiv preprint arXiv:1608.07639, 2016.

[5] Yuval Atzmon, Felix Kreuk, Uri Shalit, and Gal Chechik. A causal view of compositional zero-shot recognition. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 1462–1473, 2020.

[6] Jimmy Lei Ba. Layer normalization. arXiv preprint arXiv:1607.06450, 2016.

[7] Jinze Bai, Shuai Bai, and et al. Qwen technical report. arXiv preprint arXiv:2309.16609, 2023.

[8] Jinze Bai, Shuai Bai, and et al. Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966, 2023.

[9] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019.

[10] Rohan Bavishi, Erich Elsen, and et al. Introducing our multimodal models, 2023.

[11] Irving Biederman. Recognition-by-components: a theory of human image understanding. Psychological review, 94(2), 1987.

[12] Daniel Bolya, Cheng-Yang Fu, and et al. Token merging: Your ViT but faster. In International Conference on Learning Representations, 2023.

[13] Tom Brown, Benjamin Mann, and et al. Language models are few-shot learners. Advances in Neural Information Processing Systems, 33:1877–1901, 2020.

[14] Haozhi Cao, Yuecong Xu, Jianfei Yang, Pengyu Yin, Xingyu Ji, Shenghai Yuan, and Lihua Xie. Reliable spatial-temporal voxels for multi-modal test-time adaptation. In European Conference on Computer Vision, pages 232–249. Springer, 2024.

[15] Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. IEEE Transactions on Affective Computing, 2014.

[16] Qingqing Cao, Bhargavi Paranjape, and Hannaneh Hajishirzi. PuMer: Pruning and merging tokens for efficient vision language models. In Proceedings of the

Annual Meeting of the Association for Computational Linguistics, pages 12890–12903, July 2023.

[17] Cheng Chen, Qi Dou, Yueming Jin, Quande Liu, and Pheng Ann Heng. Learning with privileged multimodal knowledge for unimodal segmentation. IEEE transactions on medical imaging, 41(3):621–632, 2021.

[18] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020.

[19] Keqin Chen, Zhao Zhang, and et al. Shikra: Unleashing multimodal llm's referential dialogue magic. arXiv preprint arXiv:2306.15195, 2023.

[20] Liang Chen, Haozhe Zhao, and et al. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. arXiv preprint arXiv:2403.06764, 2024.

[21] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In European Conference on Computer Vision, 2018.

[22] Mengzhao Chen, Wenqi Shao, and et al. Diffrate: Differentiable compression rate for efficient vision transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 17164–17174, 2023.

[23] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[24] Zhaorun Chen, Zhuokai Zhao, and et al. Halc: Object hallucination reduction via adaptive focal-contrast decoding. arXiv preprint arXiv:2403.00425, 2024.

[25] Zhe Chen, Weiyun Wang, and et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. arXiv preprint arXiv:2404.16821, 2024.

[26] Wei-Lin Chiang and Zhuohan et al Li. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See https://vicuna. lmsys. org (accessed 14 April 2023), 2023.

[27] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In International Conference on Learning Representations, 2024.

[28] Inseop Chung, Seonguk Park, Jangho Kim, and Nojun Kwak. Feature-map-level online adversarial knowledge distillation. In International Conference on Machine Learning, 2020.

[29] Wenliang Dai and Junnan Li et al. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

[30] Yinpei Dai, Run Peng, Sikai Li, and Joyce Chai. Think, act, and ask: Open-world interactive personalized robot navigation. In 2024 IEEE International Conference on Robotics and Automation (ICRA), pages 3296–3303. IEEE, 2024.

[31] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(7):3366–3385, 2022.

[32] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv:2305.15712, 2018.

[33] Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, and Hang Zhao. On uni-modal feature learning in supervised multi-modal learning. In International Conference on Machine Learning, 2023.

[34] Mostafa Elhoushi, Akshat Shrivastava, and et al. Layer skip: Enabling early exit inference and self-speculative decoding. arXiv preprint arXiv:2404.16710, 2024.

[35] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. arXiv preprint arXiv:1805.04833, 2018.

[36] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Fushuo Huo, Jinyu Chen, and Song Guo. Overcome modal bias in multi-modal federated learning via balanced modality selection. In European Conference on Computer Vision, pages 178–195. Springer, 2024.

[37] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. Pmr: Prototypical modal rebalance for multimodal learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

[38] Alessandro Favero, Luca Zancato, and et al. Multi-modal hallucination control by visual information grounding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14303–14312, June 2024.

[39] Shangbin Feng, Weijia Shi, and et al. Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. arXiv preprint arXiv:2402.00367, 2024.

[40] Chaoyou Fu, Peixian Chen, and et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394, 2023.

[41] Zigang Geng, Binxin Yang, and et al. Instructdiffusion: A generalist modeling interface for vision tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12709–12720, June 2024.

[42] Sreyan Ghosh, Chandra Kiran Reddy Evuru, Sonal Kumar, Utkarsh Tyagi, Oriol Nieto, Zeyu Jin, and Dinesh Manocha. Visual description grounding reduces hallucinations and boosts reasoning in lvlms. In International Conference on Learning Representations, 2024.

[43] Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge distillation: A survey. International Journal of Computer Vision, 2021.

[44] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In L. Saul, Y. Weiss, and L. Bottou, editors, Advances in Neural Information Processing Systems, volume 17, 2004.

[45] Alex Graves. Sequence transduction with recurrent neural networks. arXiv preprint arXiv:1211.3711, 2012.

[46] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, and et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. In Advances in Neural Information Processing Systems, volume 36, pages 15890–15902. Curran Associates, Inc., 2023.

[47] Tianrui Guan, Fuxiao Liu, and et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024.

[48] Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. Proceedings of the AAAI conference on artificial intelligence, 2024.

[49] Jingcai Guo, Song Guo, Qihua Zhou, Ziming Liu, Xiaocheng Lu, and Fushuo Huo. Graph knows unknowns: Reformulate zero-shot learning as sample-level graph recognition. Proceedings of the AAAI Conference on Artificial Intelligence, 37(6):7775–7783, Jun. 2023.

[50] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016.

[51] Shaozhe Hao, Kai Han, and Kwan-Yee K. Wong. Learning attention as disentangler for compositional zero-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15315–15324, June 2023.

[52] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 2016.

[53] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.

[54] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. arXiv:1503.02531, 2015.

[55] Shaul Hochstein and Merav Ahissar. View from the top: Hierarchies and reverse hierarchies in the visual system. Neuron, 36(5):791–804, 2002.

[56] D.D. Hoffman and W.A. Richards. Parts of recognition. Cognition, 18(1):65–96, 1984.

[57] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In International Conference on Learning Representations, 2020.

[58] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, and et al. Lora: Low-rank adaptation of large language models. In International Conference on Learning Representations, 2022.

[59] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 2018.

[60] Xiaoming Hu and Zilei Wang. Leveraging sub-class discimination for compositional zero-shot learning. Proceedings of the AAAI conference on artificial intelligence, pages 890–898, Jun. 2023.

[61] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13418–13427, June 2024.

[62] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. In Advances in Neural Information Processing Systems, 2022.

[63] Tao Huang, Yuan Zhang, Mingkai Zheng, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge Diffusion for Distillation. arXiv:2305.15712, 2023.

[64] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6700–6709, 2019.

[65] Fushuo Huo, Bingheng Li, and Xuegui Zhu. Efficient wavelet boost learning-based multi-stage progressive refinement network for underwater image enhancement. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, pages 1944–1952, October 2021.

[66] Fushuo Huo, Ziming Liu, Jingcai Guo, Wenchao Xu, and Song Guo. Utdnet: A unified triplet decoder network for multimodal salient object detection. Neural Networks, 170:521–534, 2024.

[67] Fushuo Huo, Wenchao Xu, Jingcai Guo, Haozhao Wang, and Yunfeng Fan. Non-exemplar online class-incremental continual learning via dual-prototype self-augment and refinement. Proceedings of the AAAI Conference on Artificial Intelligence, 38(11):12698–12707, Mar. 2024.

[68] Fushuo Huo, Wenchao Xu, Jingcai Guo, Haozhao Wang, Yunfeng Fan, and Song Guo. Offline-online class-incremental continual learning via dual-prototype self-augment and refinement. arXiv preprint arXiv:2303.10891, 7, 2023.

[69] Fushuo Huo, Xuegui Zhu, and Bingheng Li. Three-stream interaction decoder network for rgb-thermal salient object detection. Knowledge-Based Systems, 258:110007, 2022.

[70] Fushuo Huo, Xuegui Zhu, Hongjiang Zeng, Qifeng Liu, and Jian Qiu. Fast fusion-based dehazing with histogram modification and improved atmospheric illumination prior. IEEE Sensors Journal, 21(4):5259–5270, 2021.

[71] Fushuo Huo, Xuegui Zhu, Lei Zhang, Qifeng Liu, and Yu Shu. Efficient context-guided stacked refinement network for rgb-t salient object detection. IEEE Transactions on Circuits and Systems for Video Technology, 32(5):3111–3124, 2022.

[72] Fushuo Huo, Xuegui Zhu, Qian Zhang, Ziming Liu, and Wenchao Yu. Real-time one-stream semantic-guided refinement network for rgb-thermal salient object

detection. IEEE Transactions on Instrumentation and Measurement, 71:1–12, 2022.

[73] Dat Huynh and Ehsan Elhamifar. Fine-grained generalized zero-shot learning via dense attribute-based attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 2020.

[74] Phillip Isola, Joseph J. Lim, and Edward H. Adelson. Discovering states and transformations in image collections. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 2015.

[75] Ziwei Ji, Nayeon Lee, and et al. Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12):1–38, 2023.

[76] Chaoya Jiang, Haiyang Xu, and et al. Hallucination augmented contrastive learning for multimodal large language model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 27036–27046, 2024.

[77] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361, 2020.

[78] Shyamgopal Karthik, Massimiliano Mancini, and Zeynep Akata. Revisiting visual product for compositional zero-shot learning. In Advances in Neural Information Processing Systems, 2021.

[79] Shyamgopal Karthik, Massimiliano Mancini, and Zeynep Akata. Kg-sp: Knowledge guided simple primitives for open world compositional zero-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9336–9345, June 2022.

[80] Mayank Kejriwal, Eric Kildebeck, Robert Steininger, and Abhinav Shrivastava. Challenges, evaluation and opportunities for open-world learning. Nature Machine Intelligence, 6(6):580–588, 2024.

[81] M. G. Kendall. Rank correlation methods. In Griffin, Oxford, England, 1948.

[82] Muhammad Gul Zain Ali Khan, Muhammad Ferjad Naeem, Luc Van Gool, Alain Pagani, Didier Stricker, and Muhammad Zeshan Afzal. Learning attention propagation for compositional zero-shot learning. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 3828–3837, January 2023.

[83] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, Youngjin Yoon, and In So Kweon. Disjoint multi-task learning between heterogeneous human-centric tasks. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 1699–1708, 2018.

[84] Taehyeon Kim, Joonkee Kim, Gihun Lee, and Se-Young Yun. Instructive decoding: Instruction-tuned large language models are self-refiner from noisy instructions. In International Conference on Learning Representations, 2024.

[85] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2015.

[86] Takeshi Kojima, Shixiang Shane Gu, and et al. Large language models are zero-shot reasoners. Advances in Neural Information Processing Systems, 35:22199–22213, 2022.

[87] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In International Conference on Machine Learning, International Conference on Machine Learning, 2019.

[88] Ranjay Krishna, Yuke Zhu, and et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision, 123:32–73, 2017.

[89] Xin Lai, Zhuotao Tian, and et al. Lisa: Reasoning segmentation via large language model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9579–9589, June 2024.

[90] Xu Lan, Xiatian Zhu, and Shaogang Gong. Knowledge distillation by on-the-fly native ensemble. In Advances in Neural Information Processing Systems, 2018.

[91] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In International Conference on Machine Learning Workshop, 2013.

[92] Nayeon Lee, Wei Ping, and et al. Factuality enhanced language models for open-ended text generation. Advances in Neural Information Processing Systems, 35:34586–34599, 2022.

[93] Pilhyeon Lee, Taeoh Kim, Minho Shim, Dongyoon Wee, and Hyeran Byun. Decomposed cross-modal distillation for rgb-based temporal action detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

[94] Sicong Leng, Hang Zhang, and et al. 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, pages 13872–13882, June 2024.

[95] Bingheng Li and Fushuo Huo. Reqa: Coarse-to-fine assessment of image quality to alleviate the range effect. Journal of Visual Communication and Image Representation, 98:104043, 2024.

[96] Bo Li, Kaichen Zhang, and et al. Llava-next: Stronger llms supercharge multimodal capabilities in the wild, May 2024.

[97] Bo Li, Yuanhan Zhang, and et al. Mimic-it: Multi-modal in-context instruction tuning. arXiv preprint arXiv:2306.05425, 2023.

[98] Chunyuan Li, Cliff Wong, and et al. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In Advances in Neural Information Processing Systems, volume 36, pages 28541–28564, 2023.

[99] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In International Conference on Machine Learning, 2023.

[100] Lujun Li. Self-regulated feature learning via teacher-free feature distillation. In European Conference on Computer Vision, 2022.

[101] Lujun Li and Zhe Jin. Shadow knowledge distillation: Bridging offline and online knowledge transfer. In Advances in Neural Information Processing Systems, 2022.

[102] Wei-Hong Li, Xialei Liu, and Hakan Bilen. Learning multiple dense prediction tasks from partially annotated data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18879–18889, June 2022.

[103] Xiang Lisa Li, Ari Holtzman, and et al. Contrastive decoding: Open-ended text generation as optimization. arXiv preprint arXiv:2210.15097, 2022.

[104] Xiangyu Li, Zhe Xu, Kun Wei, and Cheng Deng. Generalized zero-shot learning via disentangled representation. Proceedings of the AAAI conference on artificial intelligence, 35(3):1966–1974, May 2021.

[105] Xiangyu Li, Xu Yang, Kun Wei, Cheng Deng, and Muli Yang. Siamese contrastive embedding network for compositional zero-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9326–9335, June 2022.

[106] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. arXiv preprint arXiv:2305.10355, 2023.

[107] Yong-Lu Li, Yue Xu, Xiaohan Mao, and Cewu Lu. Symmetry and group in attribute-object compositions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 2020.

[108] Jiajun Liang, Linze Li, Zhaodong Bing, Borui Zhao, Yao Tang, Bo Lin, and Haoqiang Fan. Efficient one pass self-distillation with zipf's label smoothing. In European Conference on Computer Vision, 2022.

[109] Tsung-Yi Lin, Michael Maire, and et al. Microsoft coco: Common objects in context. In European Conference on Computer Vision, pages 740–755, 2014.

[110] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In International Conference on Learning Representations, 2024.

[111] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 26296–26306, June 2024.

[112] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In Advances in Neural Information Processing Systems, volume 36, pages 34892–34916, 2023.

[113] Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. arXiv preprint arXiv:2407.21771, 2024.

[114] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, and et al. Dora: Weight-decomposed low-rank adaptation. In International Conference on Machine Learning, 2024.

[115] Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, et al. Llava-plus: Learning to use tools for creating multimodal agents. In European Conference on Computer Vision, pages 126–142. Springer, 2024.

[116] Shilong Liu, Zhaoyang Zeng, and et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. European Conference on Computer Vision, 2024.

[117] Yuan Liu, Haodong Duan, and et al. Mmbench: Is your multi-modal model an all-around player? arXiv preprint arXiv:2307.06281, 2023.

[118] Ziming Liu, Song Guo, Jingcai Guo, Yuanyuan Xu, and Fushuo Huo. Towards unbiased multi-label zero-shot learning with pyramid and semantic attention. IEEE Transactions on Multimedia, 25:7441–7455, 2023.

[119] Ziming Liu, Song Guo, Xiaocheng Lu, Jingcai Guo, Jiewei Zhang, Yue Zeng, and Fushuo Huo. (ml)\$ˆ2\$p-encoder: On exploration of channel-class correlation for multi-label zero-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 23859–23868, June 2023.

[120] David M. W. Powers. Applications and explanations of zipf's law. In New methods in language processing and computational natural language learning, 1998.

[121] Fan Ma, Xiaojie Jin, Heng Wang, Yuchen Xian, Jiashi Feng, and Yi Yang. Vista-llama: Reducing hallucination in video language models via equal distance to visual tokens. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024.

[122] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In European Conference on Computer Vision, 2018.

[123] Potsawee Manakul, Adian Liusie, and Mark JF Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. arXiv preprint arXiv:2303.08896, 2023.

[124] Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Open world compositional zero-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5222–5230, June 2021.

[125] Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Learning graph embeddings for open world compositional zero-shot learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, pages 1–1, 2022.

[126] AI Meta. Introducing meta llama 3: The most capable openly available llm to date. Meta AI, 2024.

[127] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems, volume 26, 2013.

[128] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. In Advances in Neural Information Processing Systems, 2023.

[129] Ishan Misra, Abhinav Gupta, and Martial Hebert. From red wine to red tomato: Composition with context. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, July 2017.

[130] Muhammad Ferjad Naeem, Yongqin Xian, Federico Tombari, and Zeynep Akata. Learning graph embeddings for compositional zero-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 953–962, June 2021.

[131] Tushar Nagarajan and Kristen Grauman. Attributes as operators: Factorizing unseen attribute-object compositions. In European Conference on Computer Vision, September 2018.

[132] Vinod Nair and Geoffrey Hinton. Rectified linear units improve restricted boltzmann machines. In International Conference on Machine Learning, 2010.

[133] Vladimir Nekrasov, Thanuja Dharmasiri, Andrew Spek, Tom Drummond, Chunhua Shen, and Ian Reid. Real-time joint semantic segmentation and depth estimation using asymmetric annotations. In 2019 International Conference on Robotics and Automation (ICRA), pages 7101–7107, 2019.

[134] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.

[135] Jitendra Parmar, Satyendra Chouhan, Vaskar Raychoudhury, and Santosh Rathore. Open-world machine learning: applications, challenges, and opportunities. ACM Computing Surveys, 55(10):1–37, 2023.

[136] Karl Pearson. Vii. mathematical contributions to the theory of evolution.—iii. regression, heredity, and panmixia. In Philosophical Transactions of the Royal Society of London, 1896.

[137] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

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

[139] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. Proceedings of the AAAI conference on artificial intelligence, 32(1), 2018.

[140] Senthil Purushwalkam, Maximilian Nickel, Abhinav Gupta, and Marc'Aurelio Ranzato. Task-driven modular networks for zero-shot compositional learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, October 2019.

[141] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning, volume 139, pages 8748–8763, 18–24 Jul 2021.

[142] Yongming Rao, Wenliang Zhao, and et al. Dynamicvit: Efficient vision transformers with dynamic token sparsification. Advances in Neural Information Processing Systems, 34:13937–13949, 2021.

[143] Siddharth Roheda, Benjamin S. Riggan, Hamid Krim, and Liyi Dai. Cross-modality distillation: A case for conditional generative adversarial networks. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018.

[144] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. arXiv preprint arXiv:1809.02156, 2018.

[145] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets, 2015.

[146] Frank Ruis, Gertjan Burghouts, and Doina Bucur. Independent prototype propagation for zero-shot compositionality. In Advances in Neural Information Processing Systems, volume 34, 2021.

[147] Nirat Saini, Khoi Pham, and Abhinav Shrivastava. Disentangling visual embeddings for attributes and objects. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13658–13667, June 2022.

[148] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[149] Tal Schuster, Adam Fisch, and et al. Confident adaptive language modeling. In Advances in Neural Information Processing Systems, 2022.

[150] Dustin Schwenk, Apoorv Khandelwal, and et al. A-okvqa: A benchmark for visual question answering using world knowledge. In European Conference on Computer Vision, pages 146–162, 2022.

[151] Yuzhang Shang, Mu Cai, and et al. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. arXiv preprint arXiv:2403.15388, 2024.

[152] Yiqing Shen, Liwu Xu, Yuzhe Yang, Yaqian Li, and Yandong Guo. Self-distillation from the last mini-batch for consistency regularization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[153] Inkyu Shin, Yi-Hsuan Tsai, Bingbing Zhuang, Samuel Schulter, Buyu Liu, Sparsh Garg, In So Kweon, and Kuk-Jin Yoon. Mm-tta: multi-modal test-time adaptation for 3d semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16928–16937, 2022.

[154] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In European Conference on Computer Vision, 2012.

[155] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In Proceedings of the AAAI conference on artificial intelligence, page 4444–4451, 2017.

[156] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 15:1929–1958, JUN 2014.

[157] Rohan Taori, Ishaan Gulrajani, and et al. Stanford alpaca: an instruction-following llama model (2023). URL https://github.com/tatsu-lab/stanford_alpaca, 1(9), 2023.

[158] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In European Conference on Computer Vision, 2018.

[159] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In International Conference on Learning Representations, 2020.

[160] Hugo Touvron, Thibaut Lavril, and et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023.

[161] Hugo Touvron, Louis Martin, and et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023.

[162] Simon Vandenhende, Stamatios Georgoulis, Wouter Van Gansbeke, Marc Proesmans, Dengxin Dai, and Luc Van Gool. Multi-task learning for dense prediction tasks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(7):3614–3633, 2022.

[163] Ashish Vaswani, Noam Shazeer, and et al. Attention is all you need. Advances in Neural Information Processing Systems, 30, 2017.

[164] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In International Conference on Learning Representations, 2021.

[165] Lin Wang and Kuk-Jin Yoon. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.

[166] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11531–11539, 2020.

[167] Qingsheng Wang, Lingqiao Liu, Chenchen Jing, Hao Chen, Guoqiang Liang, Peng Wang, and Chunhua Shen. Learning conditional attributes for compositional zero-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11197–11206, June 2023.

[168] Sheng Wang, Zihao Zhao, and et al. Chatcad: Interactive computer-aided diagnosis on medical image using large language models. arXiv preprint arXiv:2302.07257, 2023.

[169] Shuai Wang, Zipei Yan, Daoan Zhang, Haining Wei, Zhongsen Li, and Rui Li. Prototype knowledge distillation for medical segmentation with missing modality. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5. IEEE, 2023.

[170] Xintong Wang, Jingheng Pan, and et al. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. arXiv preprint arXiv:2403.18715, 2024.

[171] Xinyi Wang, Wanrong Zhu, and et al. Large language models are latent variable models: Explaining and finding good demonstrations for in-context learning. Advances in Neural Information Processing Systems, 36, 2024.

[172] Sangmin Woo, Donguk Kim, and et al. Don't miss the forest for the trees: Attentional vision calibration for large vision language models. arXiv preprint arXiv:2405.17820, 2024.

[173] Junfei Wu, Qiang Liu, and et al. Logical closed loop: Uncovering object hallucinations in large vision-language models. arXiv preprint arXiv:2402.11622, 2024.

[174] Mingrui Wu, Jiayi Ji, and et al. Evaluating and analyzing relationship hallucinations in lvlms. arXiv preprint arXiv:2406.16449, 2024.

[175] Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(9):2251–2265, 2019.

[176] Junlin Xie, Zhihong Chen, Ruifei Zhang, Xiang Wan, and Guanbin Li. Large multimodal agents: A survey. arXiv preprint arXiv:2402.15116, 2024.

[177] Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(10):12113–12132, 2023.

[178] Ziwei Xu, Guangzhi Wang, Yongkang Wong, and Mohan S. Kankanhalli. Relation-aware compositional zero-shot learning for attribute-object pair recognition. IEEE Transactions on Multimedia, 24:3652–3664, 2022.

[179] Zihui Xue, Zhengqi Gao, Sucheng Ren, and Hang Zhao. The modality focusing hypothesis: Towards understanding crossmodal knowledge distillation. In International Conference on Learning Representations, 2023.

[180] Zihui Xue, Sucheng Ren, Zhengqi Gao, and Hang Zhao. Multimodal knowledge expansion. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.

[181] Chuanguang Yang, Zhulin An, Helong Zhou, Linhang Cai, Xiang Zhi, Jiwen Wu, Yongjun Xu, and Qian Zhang. Mixskd: Self-knowledge distillation from mixup for image recognition. In European Conference on Computer Vision, 2022.

[182] Chuanguang Yang, Helong Zhou, Zhulin An, Xue Jiang, Yongjun Xu, and Qian Zhang. Cross-image relational knowledge distillation for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[183] Jing Yang, Brais Martinez, Adrian Bulat, and Georgios Tzimiropoulos. Knowledge distillation via softmax regression representation learning. In International Conference on Learning Representations, 2021.

[184] Mouxing Yang, Yunfan Li, Zhang Changqing, Peng Hu, and Xi Peng. Test-time adaption against multi-modal reliability bias. In International Conference on Learning Representations, May 2024.

[185] Muli Yang, Cheng Deng, Junchi Yan, Xianglong Liu, and Dacheng Tao. Learning unseen concepts via hierarchical decomposition and composition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 2020.

[186] Muli Yang, Chenghao Xu, Aming Wu, and Cheng Deng. A decomposable causal view of compositional zero-shot learning. IEEE Transactions on Multimedia, pages 1–11, 2022.

[187] Penghui Yang, Ming-Kun Xie, Chen-Chen Zong, Lei Feng, Gang Niu, Masashi Sugiyama, and Sheng-Jun Huang. Multi-label knowledge distillation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023.

[188] Zhendong Yang, Ailing Zeng, Zhe Li, Tianke Zhang, Chun Yuan, and Yu Li. From knowledge distillation to self-knowledge distillation: A unified approach with normalized loss and customized soft labels. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023.

[189] Qinghao Ye, Haiyang Xu, and et al. mplug-owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178, 2023.

[190] Shukang Yin, Chaoyou Fu, and et al. Woodpecker: Hallucination correction for multimodal large language models. arXiv preprint arXiv:2310.16045, 2023.

[191] Alex Young, Bei Chen, and et al. Yi: Open foundation models by 01. ai. arXiv preprint arXiv:2403.04652, 2024.

[192] Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 2014.

[193] Qifan Yu, Juncheng Li, and et al. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12944–12953, June 2024.

[194] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In Advances in Neural Information Processing Systems, 2020.

[195] Tianyu Yu, Yuan Yao, and et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13807–13816, June 2024.

[196] Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[197] Zihao Yue, Liang Zhang, and Qin Jin. Less is more: Mitigating multimodal hallucination from an eos decision perspective. arXiv preprint arXiv:2402.14545, 2024.

[198] Heeseung Yun, Joonil Na, and Gunhee Kim. Dense 2d-3d indoor prediction with sound via aligned cross-modal distillation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023.

[199] Chi Zhang, Zhao Yang, Jiaxuan Liu, Yucheng Han, Xin Chen, Zebiao Huang, Bin Fu, and Gang Yu. Appagent: Multimodal agents as smartphone users. arXiv preprint arXiv:2312.13771, 2023.

[200] Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. How language model hallucinations can snowball. In International Conference on Machine Learning, 2024.

[201] Shilong Zhang, Peize Sun, and et al. Gpt4roi: Instruction tuning large language model on region-of-interest. arXiv preprint arXiv:2307.03601, 2023.

[202] Tianlu Zhang, Hongyuan Guo, Qiang Jiao, Qiang Zhang, and Jungong Han. Efficient rgb-t tracking via cross-modality distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

[203] Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. Deep mutual learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[204] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[205] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In European Conference on Computer Vision, 2018.

[206] Zhiyuan Zhao, Bin Wang, and et al. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. arXiv preprint arXiv:2311.16839, 2024.

[207] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 2016.

[208] Yiyang Zhou, Chenhang Cui, and et al. Analyzing and mitigating object hallucination in large vision-language models. In International Conference on Learning Representations, 2024.

[209] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592, 2023.

[210] Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding. arXiv preprint arXiv:2402.18476, 2024.

[211] Yichen Zhu, Ning Liu, Zhiyuan Xu, Xin Liu, Weibin Meng, Louis Wang, Zhicai Ou, and Jian Tang. Teach less, learn more: On the undistillable classes in knowledge distillation. In Advances in Neural Information Processing Systems, 2022.

[212] Yichen Zhu and Yi Wang. Student customized knowledge distillation: Bridging the gap between student and teacher. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.