MLLMs ARE DEEPLY AFFECTED BY MODALITY BIAS

Xu Zheng[1,3,*] **Chenfei Liao**[1,*] **Yuqian Fu**[3] **Kaiyu Lei**[4] **Yuanhuiyi Lyu**[1]
Lutao Jiang[1] **Bin Ren**[5,6,3] **Jialei Chen**[7] **Jiawen Wang**[8] **Chengxin Li**[9,10]
Linfeng Zhang[11] **Danda Pani Paudel**[3] **Xuanjing Huang**[12] **Yu-Gang Jiang**[12]
Nicu Sebe[6] **Dacheng Tao**[13] **Luc Van Gool**[3] **Xuming Hu**[1,2,†]
[1]HKUST(GZ) [2]CSE, HKUST [3]INSAIT, Sofia University "St. Kliment Ohridski"
[4]Xi'an Jiaotong University [5]University of Pisa, IT [6]University of Trento, IT
[7]Nagoya University [8]China University of Mining & Technology, Beijing
[9]Tongji University [10]SPIC Energy Science and Technology Research Institute
[11]Shanghai Jiao Tong University [12]Fudan University
[13]College of Computing & Data Science, Nanyang Technological University

ABSTRACT

Recent advances in Multimodal Large Language Models (MLLMs) have shown promising results in integrating diverse modalities such as texts and images. MLLMs are heavily influenced by modality bias, often relying on language while under-utilizing other modalities like visual inputs. This position paper ***argues that MLLMs are deeply affected by modality bias***. Firstly, we diagnose the current state of modality bias, highlighting its manifestations across various tasks. Secondly, we propose a systematic research road-map related to modality bias in MLLMs. Thirdly, we identify key factors of modality bias in MLLMs, including data characteristics, imbalanced backbone capabilities, and training objectives, offering actionable suggestions for future research to mitigate it. These findings highlight the need for balanced training strategies and model architectures to better integrate multiple modalities in MLLMs. We call for interdisciplinary efforts to tackle these challenges and drive innovation in MLLM research. Our work provides a fresh perspective on modality bias in MLLMs and offers insights for developing more robust and generalizable multimodal systems—advancing progress toward Artificial General Intelligence.

1 INTRODUCTION

1.1 BACKGROUND

Multimodal Large Language Models (MLLMs) (Bai et al., 2025; Chen et al., 2024c) have revolutionized the ability to handle diverse modalities, including text, image, audio, video, and other emerging modalities (tactile (Dahiya et al., 2009), event (Gallego et al., 2020; Rebecq et al., 2019), panoramic image (Zhong et al., 2025), *etc*). This expansion into the multimodal domain typically involves pretraining with multimodal data pairs or fine-tuning on specialized multimodal instruction datasets (Liu et al., 2024e; Fang et al., 2024). MLLMs excel at understanding complex multimodal patterns and translating them into a coherent language representation space (Wu et al., 2024). Despite significant advancements, challenges remain, where one major issue is **"Modality Bias"**. As in (Zhang et al., 2024), MLLMs often generate content that is disproportionately influenced by the underlying language model used during pretraining, rather than the input images or other modalities. In cases where images are noisy or even absent, MLLMs still confidently generate answers, highlighting a clear bias towards learned language patterns over multimodal integration (Park et al., 2025; Tong et al., 2024).

An ideal MLLM should be modality-balanced, effectively integrating useful information from all modalities to provide reliable, accurate, and comprehensive answers (Chen et al., 2024a). Achieving

[*]These authors have equal contributions.
[†]Corresponding author.

this balance is crucial for overcoming modality bias and ensuring that the model can leverage the full potential of each modality in multimodal tasks (Chen et al., 2024b; Yue et al., 2024b).

1.2 Modality Bias Phenomenon

Multimodal learning improves neural networks' cross-modal comprehension by fusing heterogeneous data modalities (*e.g.*, images, text, audio, and video), thereby enabling better world modeling (Xu et al., 2023; Brödermann et al., 2025). It is widely assumed that leveraging multiple input modalities will lead to improved model performance (Manzoor et al., 2023). However, research has demonstrated that these modalities are not always utilized to their full potential (Wei et al., 2024; Zheng et al., 2024a;b). Despite achieving superior performance over unimodal models, multimodal models still fail to fully exploit the capabilities of each modality (Peng et al., 2022).

As noted by (Alabdulmohsin et al.), multimodal models, particularly those employing contrastive learning techniques such as CLIP (Radford et al., 2021), learn representations from multimodal data but may inadvertently inherit biases due to the imbalanced and biased nature of the training data. As discussed in (Xu et al., 2025), this issue severely hinders the effectiveness of multimodal learning. In detail, such a condition occurs when certain modalities dominate the training process while others remain underrepresented, constraining the model's capacity to capture the comprehensive information embedded in multimodal data distributions. Such dominance can lead to models' over-reliance on the dominant modality, which impedes their ability to generalize effectively to unseen data or situations where the dominant modality is absent.

Moreover, in multimodal systems, modality bias can manifest in several other ways. For instance, if a model is primarily trained on image-text pairs, but audio or video data is only sparsely represented, the model may learn representations that are disproportionately influenced by the image-text modalities, while neglecting the rich information available from the audio or video inputs. This type of imbalance leads to a model that may perform well under normal circumstances but struggles to generalize when the dominant modality is absent.

From a model learning perspective, (Yang et al., 2024) identifies the differing convergence rates of modalities as a core cause of modality bias. The varying levels of difficulty in fitting category labels across different modalities contribute to this disparity. Some modalities may align more easily with the target labels, leading to an unequal contribution to the final learned representations. This uneven convergence further exacerbates the problem, reinforcing the bias towards certain modalities.

Recent studies in multimodal scene understanding have further highlighted this issue. For example, research by (Zheng et al., 2024a) and (Liao et al., 2025) shows that multimodal segmenters often over-rely on certain modalities, resulting in significant performance degradation when these dominant modalities are missing or unavailable. These findings underscore the need to address modality imbalance to ensure all modalities contribute effectively to the learning process.

1.3 Our Position

In the context of Multimodal Large Language Models (MLLMs), the presence of modality bias is also evident. For instance, empirical results in (Zhang et al., 2024) reveal that MLLMs exhibit modality bias in the generated content. Specifically, the output of MLLMs is often primarily influenced by the language model's prior knowledge, rather than by the input images. That is, MLLMs frequently produce confident responses even when relevant images are absent or when incongruent visual input is provided. This phenomenon is also validated by our experiments in Sec. 4. Moreover, work (Leng et al., 2024) further confirms that the modality bias in MLLMs stems from the complex interactions between multiple modalities, which complicates the multimodal debiasing process. Thus, we propose the position that **MLLMs are deeply affected by modality bias**. Firstly, we offer the definition of modality bias in Sec. 2. Secondly, we review the research roadmap about modality bias in MLLMs in Sec. 3.1. Thirdly, we analysis the key factors of modality bias in MLLMs in Sec. 3.2, accompanied with a case study in Sec. 4. Finally, we conclude further the targeted solutions of modality bias in MLLMs, including current works and future directions in Sec. 5 and Fig. 3.

2 PROPOSED DEFINITION OF MODALITY BIAS

Modality bias arises when certain modalities dominate the learning process, while others are underutilized or contribute less effectively (Guo et al., 2023). This imbalance often results in a model that is biased towards the dominant modality, thus failing to fully leverage the potential of the underrepresented modalities (Vosoughi et al., 2024). As a result, the model's performance can degrade significantly when the dominant modality is missing, unavailable, or unreliable.

To mathematically describe this imbalance, let us define the contribution of each modality M_i as $C(M_i)$, where $i \in \{1, 2, \ldots, n\}$ represents the different modalities (e.g., image, text, audio). The total contribution of all modalities is given by the sum of these individual contributions:

$$C_{\text{total}} = \sum_{i=1}^{n} C(M_i). \tag{1}$$

If certain modalities dominate, the distribution of contributions becomes imbalanced, such that $C(M_i) \gg \{C(M_x), C(M_y), \cdots, C(M_z)\}$ for some $i \neq \{x, y, \cdots, z\}$, as shown in Fig. 1. This imbalance can lead to several issues:

① Over-reliance on dominant modalities: The model may become overly sensitive to the dominant modality M_i, resulting in biased predictions that fail to incorporate the full diversity of information from the multimodal data.

② Underutilization of certain modalities: Modalities that are underrepresented in the training data, such as audio or video, contribute less to the learned representations, leading to models that lack robustness when these modalities are needed.

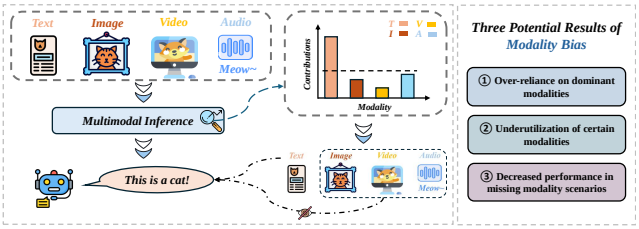

Figure 1: Further definition of modality bias and three potential results.

③ Decreased performance in missing modality scenarios: When a dominant modality is missing during inference (for example, if an image is unavailable), the model's performance can drastically drop, as it has not sufficiently learned how to balance the different modalities during training.

To capture the extent of modality bias, we can define a relative measure of imbalance, known as the modality imbalance ratio Δ_{modality}, as the ratio of the contribution of the dominant modality to the underutilized modality[1]:

$$\Delta_{\text{modality}} = \frac{C(M_{\text{dominant}})}{C(M_{\text{underutilized}})}. \tag{2}$$

This ratio quantifies the disparity between the contributions of the modalities and can serve as a diagnostic tool to identify and address modality bias. A high value of Δ_{modality} indicates a strong bias towards the dominant modality, which can hinder the model's ability to generalize effectively.

In conclusion, modality bias is a fundamental issue in multimodal learning that arises from the unequal contributions of different modalities. It leads to suboptimal learning outcomes and impairs the model's ability to generalize, especially when certain modalities are missing or unavailable. Addressing modality bias involves ensuring that all modalities are effectively utilized and contribute in a balanced manner, thereby improving the robustness and performance of multimodal systems.

[1]The definition is for better illustration of modality bias, not for calculating.

3 How are MLLMs Deeply Affected by Modality Bias?

3.1 Research Road-map

The exploration process of modality bias in MLLMs can be divided into three directions: *(a) How to prove the bias? (b) How to solve the bias through datasets? (c) How to solve the bias through methods?* These three directions are defined by their different focuses, including bias/debias, datasets/methods, collaborating to highlight and solve modality bias in MLLMs.

① How to prove the bias?

With the modality bias in MLLMs emerging gradually as research focus, several datasets and benchmarks have been proposed to measure the modality bias in MLLMs (Park et al., 2025; Tong et al., 2024; Lee et al., 2024; Leng et al., 2024; Liu et al., 2024d). Park *et al*. (Park et al., 2025) directly proposed a metric named Modality Importance Score (MIS) to measure each modality's contribution in the video question answering task. Based on a comprehensive benchmark, the modality imbalance in current multimodal datasets is proven. Lee *et al*. (Lee et al., 2024) and Leng *et al*. (Leng et al., 2024) mainly emphasized the modality prior, which is a key reason for modality bias in MLLMs. Specifically, Lee *et al*. (Lee et al., 2024) introduced counterfactual images in VLind-Bench to measure the language priors of LVLMs, proving LVLMs have a great over-reliance on language priors. Leng *et al*. (Leng et al., 2024) proposed a more comprehensive benchmark, namely Curse of Multi-Modalities (CMM), including three modalities: language, visual, and audio. The results of CMM further explain the contributors to hallucinations, where the over-reliance on unimodal priors plays an important role. Liu *et al*. (Liu et al., 2024d) explored the bias from the perspective of vision-knowledge conflicts, proving the over-reliance of MLLMs on texts. Moreover, Tong *et al*. (Tong et al., 2024) proposed the Multimodal Visual Patterns (MMVP) benchmark, further exploring the contrastive language-image pre-training (CLIP)'s weaknesses, which lead to MLLMs' failures in understanding visual information.

② How to solve the bias through datasets?

With the modality bias proven to be a common phenomenon in datasets, which is the foundation of training and validation for MLLMs, researchers set their sights on how to solve the bias in datasets (Chen et al., 2024a;b; Yue et al., 2024b). Chen *et al*. (Chen et al., 2024b) proposed MORE, a VQA dataset that requires multi-hop reasoning and overcoming unimodal biases, providing counterexample data to drive the LVLMs to overcome modality bias. Meanwhile, several works focus on decreasing the modality bias in multimodal datasets. Chen *et al*. (Chen et al., 2024a) proposed MMStar, a meticulously designed multimodal benchmark, of which each sample shows visual dependency, avoiding the modality bias in datasets. Yue *et al*. (Yue et al., 2024b) built a robust benchmark MMMU-Pro based on MMMU (Yue et al., 2024a). Through steps such as making questions embedded in images, MMMU-Pro is equipped with the ability to force MLLMs to both "see" and "read".

③ How to solve the bias through methods?

Besides datasets, applying specific methods to reduce the modality bias in MLLMs is another tendency (Zhang et al., 2024; Tong et al., 2024; Zhao et al., 2024a; Pi et al., 2024; Liu et al., 2024d;c; Zhao et al., 2024b; Zhang et al., 2025b; Li et al., 2025). Typically, Pi *et al*. (Pi et al., 2024) and Zhang *et al*. (Zhang et al., 2025b) introduced preference learning methods, such as Bootstrapped Preference Optimization (BPO) and Noise-Aware Preference Optimization (NaPO), solving the modality bias problem based on building negative response datasets. Meanwhile, Zhang *et al*. (Zhang et al., 2024), Liu *et al*. (Liu et al., 2024d), and Tong *et al*. (Tong et al., 2024) proposed frameworks and methods to "force" MLLMs to pay more attention to images and boost MLLMs' visual understanding abilities. Moreover, Li *et al*. (Li et al., 2025) focused on the Multimodal Reward Models (MM-RMs) for MLLMs, proposing a shortcut-aware MM-RM learning algorithm, decreasing MLLMs' reliance on unimodal spurious correlations. Most above works consider unimodal dependency, especially on textual modality, as the key reason for modality bias. Thus, the boosting of visual modality gradually turns into a major research topic.

3.2 KEY FACTORS OF MODALITY BIAS IN MLLMS

Based on Sec 3.1, the key factors of modality bias in MLLMs can be concluded as follows: dataset imbalances, varying modal capabilities, training objectives, and the interactions between modalities. These factors contribute to the unequal utilization of modalities during training, leading to biases towards certain modalities and suboptimal performance. We summarize three key factors as follows:

① **Dataset Imbalances:** The training dataset composition significantly influences modality utilization. Datasets often have imbalanced modality distributions, where modalities, such as text or images, are more abundant or have different information density (Chen et al., 2024a;b; Yue et al., 2024b). This imbalance leads to models learning representations biased towards the more frequent modalities, while under-utilizing the less represented ones, even when multiple modalities are available. In addition, textual data is more semantically dense or informative than visual data in certain tasks, due to its structured and explicit nature. As a result, models tend to prioritize textual inputs during learning, treating accompanying modalities such as images merely as auxiliary cues, further amplifying the reliance on dominant modalities.

② **Asymmetric Modal Backbone Capabilities:** Different modalities vary in complexity and in the architectural designs used to process them. Language models often benefit from mature and highly optimized transformer-based architectures (Liu et al., 2024a; Bi et al., 2024; Naveed et al., 2023), which are not only effective but also backed by extensive research and industrial-scale pretraining. In contrast, processing visual or acoustic data typically requires more diverse and specialized backbones (Ren et al., 2023; Han et al., 2022; Ren et al., 2024b; Liu et al., 2021; Ren et al., 2024a; Huang et al., 2024) and may not benefit from equally massive pretraining corpora. Moreover, the rapid advancement of language models, fueled by large-scale datasets and sustained community focus, has further widened the performance gap across modalities. As a result, multimodal models with strong language backbones tend to over-rely on text inputs, under-utilizing other modalities, particularly those that demand more complex or less mature processing pipelines.

③ **Training Objectives:** The choice of training objectives fundamentally shapes how multimodal models utilize different modalities and often exacerbates modality bias. Pretraining strategies in many state-of-the-art multimodal models—such as CLIP-style contrastive learning, image–text matching (ITM), masked language modeling (MLM), or caption generation—tend to prioritize text–image alignment due to the abundance of paired data and the relative ease of textual supervision. These objectives implicitly encourage the model to rely heavily on language as the semantic anchor, such as LanguageBind (Zhu et al., 2024) and UniBind (Lyu et al., 2024b). Consequently, modalities like audio, video, point clouds, or thermal data—which are harder to align, less semantically rich in isolation, or lack large-scale supervision—are under-optimized during pretraining. Furthermore, most objectives do not explicitly encourage consistent cross-modal alignment or robust fusion across diverse modalities, resulting in imbalanced feature representations and limited generalization to underrepresented input types.

Additionally, two other factors contribute to modality bias:

④ **Differences in Convergence Rates:** Each modality converges at different rates during training. Some modalities, like images and text, are more easily aligned with target labels due to their structure and high information density, while others, such as audio or video, require more complex processing. This disparity results in certain modalities being more influential in the model's final learned representation, amplifying modality bias.

⑤ **Modal Interactions and Integrations:** The interaction between modalities also affects modality bias. If relationships between modalities are not explicitly learned, the model may favor the more easily processed modality, like text, over others. The complexity of integrating multimodal information can exacerbate bias, as the model may struggle to effectively combine all modalities, resulting in predictions that under-utilize available data.

In summary, modality bias is driven by factors such as dataset imbalances, differences in modal capabilities, training objectives, and the interactions between modalities. Addressing these factors is essential to mitigating modality bias and improving multimodal model performance. Strategies to balance modality contributions during training, optimize multimodal integration, and address dataset imbalances are critical for building fairer and more robust multimodal systems.

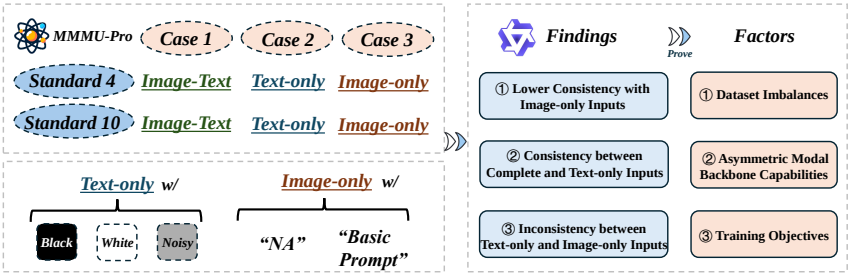

Figure 2: Case study for exploring modality bias in MLLMs. Dataset: MMMU-Pro, MLLM: Qwen2.5VL. Based on this case study, the three main factors proposed in Sec. 3.2 are further illustrated and proved. "white" means the image pixels are all set to 255. "black" means the image pixels are all set to 0. Results come from a single run.

4 Case Study

Table 1: Directly applying missing modality evaluation with MVLMs on MMMU-pro dataset. Basic prompt (B-P) for Image only is: "Based on the provided images, please answer the question."

Model	Inference	Standard 4			Standard 10		
		Image-Text	Text /w white	Image	Image-Text	Text w/ white	Image w/ B-P
Qwen2.5-VL-7B-I	Direct	48.32	34.91	28.73	37.57	21.73	14.91
		-	-13.41 ↓	-19.59 ↓	-	-15.84 ↓	-22.66 ↓
	CoT	52.14	35.20	28.38	39.94	21.39	14.97
		-	-16.94 ↓	-23.76 ↓	-	-18.55 ↓	-24.97 ↓
Qwen2.5-VL-32B-I	Direct	57.80	40.17	27.23	43.94	28.32	15.38
		-	-17.63 ↓	-30.57 ↓	-	-15.62 ↓	-28.56 ↓
	CoT	59.88	40.12	26.36	48.55	26.82	14.34
		-	-19.76 ↓	-33.52 ↓	-	-21.73 ↓	-34.21 ↓
Model	Inference	Standard 4			Standard 10		
		Image-Text	Text /w black	Image	Image-Text	Text w/ black	Image w/ B-P
Qwen2.5-VL-7B-I	Direct	48.32	34.86	29.02	37.63	21.62	14.51
		-	-13.46 ↓	-19.30 ↓	-	-16.01 ↓	-23.12 ↓
	CoT	52.08	21.45	28.38	39.94	21.45	14.86
		-	-30.63 ↓	-23.70 ↓	-	-18.49 ↓	-25.08 ↓
Qwen2.5-VL-32B-I	Direct	57.80	40.17	26.99	43.87	28.21	15.03
		-	-17.63 ↓	-30.81 ↓	-	-15.66 ↓	-28.84 ↓
	CoT	59.83	40.35	26.65	48.44	26.82	14.86
		-	-19.48 ↓	-33.18 ↓	-	-21.62 ↓	-33.58 ↓
Model	Inference	Standard 4			Standard 10		
		Image-Text	Text w/ noise	Image	Image-Text	Text w/ noise	Image w/ B-P
Qwen2.5-VL-7B-I	Direct	48.32	34.97	26.99	37.57	21.56	14.86
		-	-13.35 ↓	-21.33 ↓	-	-16.01 ↓	-22.71 ↓
	CoT	51.73	35.66	27.40	40.00	21.27	13.35
		-	-16.07 ↓	-24.33 ↓	-	-18.73 ↓	-26.65 ↓
Qwen2.5-VL-32B-I	Direct	57.75	40.23	26.13	43.70	28.27	14.28
		-	-17.52 ↓	-31.62 ↓	-	-15.43 ↓	-29.42 ↓
	CoT	60.12	40.12	28.03	48.03	27.17	13.99
		-	-20.00 ↓	-32.09 ↓	-	-20.86 ↓	-34.04 ↓

The results presented in Tab. 1 and Tab. 2 reveal several key insights regarding the performance of the multimodal large model when tested with different input combinations across the MMMU-pro (Yue et al., 2024b) dataset. The process of case study is shown in Fig. 2. These insights can be linked to the three key factors identified in our analysis of modality bias in Multimodal Large Language Models (MLLMs): ① dataset imbalances, ② asymmetric modal backbone capabilities, and ③ training objectives.

Lower Consistency with Image-only Inputs 27.17% (Complete & Image-only, Direct) and 28.21% (Complete & Image-only, CoT): The relatively low consistency between the complete input and image-only input suggests that the image modality alone is not sufficient for the model to make consistent predictions. When the model only has access to visual data, its predictions tend to be less reliable, underscoring the inadequacies of the model in processing visual data in isolation. This result supports the factor of ① dataset imbalances, where the richness and complexity of image data, compared to more compact textual data, pose challenges for the model. Although images provide

Table 2: Prediction consistency analysis of Qwen2.5-VL-7B-Instruct model on MMMU-pro. Complete means both images and text are used as input; Direct and CoT refer to the inference techniques.

Choices		Standard 4 (Direct)		Standard 4 (CoT)		Standard 10 (Direct)		Standard 10 (CoT)	
		Num.	Percentage	Num.	Percentage	Num.	Percentage	Num.	Percentage
All Samples									
Consistent	Complete & Text-only	978	56.53%	755	43.64%	821	47.46%	568	32.83%
	Complete & Image-only	470	27.17%	488	28.21%	262	15.14%	260	15.03%
	Text-only & Image-only	463	26.76%	449	25.95%	234	13.53%	232	13.41%
	All	267	15.43%	220	12.72%	112	6.47%	87	5.03%
Inconsistent	Complete & Text-only	752	43.47%	975	56.36%	909	52.54%	1162	67.17%
	Complete & Image-only	1260	72.83%	1242	71.70%	1468	84.86%	1470	84.97%
	Text-only & Image-only	1267	73.24%	1281	74.05%	1496	86.47%	1498	86.59%
	All	353	20.40%	478	27.63%	637	36.82%	844	48.79%
Correct Samples									
Consistent	Complete & Text-only	456	26.32%	417	24.08%	284	16.42%	231	13.36%
	Complete & Image-only	254	14.67%	264	15.26%	126	7.28%	135	7.80%
	Text-only & Image-only	185	10.68%	174	10.05%	61	3.53%	56	3.24%
	All	142	8.20%	125	7.22%	49	2.83%	42	2.43%
Wrong Samples									
Consistent	Complete & Text-only	522	30.15%	338	19.51%	537	31.04%	337	19.46%
	Complete & Image-only	216	12.49%	224	12.94%	136	7.86%	125	7.22%
	Text-only & Image-only	278	16.06%	275	15.91%	173	9.99%	176	10.16%
	All	125	7.22%	95	5.49%	63	3.64%	45	2.60%

important visual cues, the model struggles to effectively utilize the image modality alone, indicating that the image modality is underutilized in the absence of complementary text data.

Consistency between Complete and Text-only Inputs 56.53% (Complete & Text-only, Direct) and 43.64% (Complete & Text-only, CoT): The finding that over half of the samples show consistency between the complete (both image and text) and text-only inputs across both inference techniques (Direct and CoT) is significant. It suggests that textual information alone is a strong foundation for the model's predictions, and in many cases, the image modality does not substantially alter the model's output. This highlights the dominance of the language modality, which is particularly advantageous due to its well-established processing capabilities. This result is consistent with the factor of ② asymmetric modal backbone capabilities, where models with stronger language backbones, such as this one, tend to perform better on language tasks, often overshadowing the visual modality and limiting the model's ability to effectively integrate multimodal information.

Inconsistency between Text-only and Image-only Inputs 26.76% (Text-only & Image-only, Direct) and 25.95% (Text-only & Image-only, CoT): The low consistency between text-only and image-only inputs highlights the challenge the model faces when dealing with these two distinct modalities separately. This discrepancy suggests that both text and image provide complementary yet crucial information for accurate predictions. Textual data offers rich semantic context, nuances, and detail that images alone cannot convey, while images provide visual cues and spatial relationships that text cannot fully express. The low consistency between these two modalities, especially in the CoT setting, where reasoning and integration are critical, points to the challenge of combining these modalities effectively. This underscores the factor of ③ training objectives, where existing training strategies often fail to adequately balance multimodal learning, leading to modality-specific shortcuts. In the case of this model, the failure to effectively combine text and image information results in inconsistent predictions, especially when reasoning across modalities is required.

Our findings underscore the importance of balanced training strategies and model architectures to address modality bias and improve multimodal integration. This also highlights the need for future research aimed at developing MLLMs that can more effectively process and combine diverse sources of information, mitigating the impact of modality bias.

5 Targeted Solutions

5.1 Current Works

① **Enhance visual modality's contribution in datasets:** With the in-depth exploration in modality bias, especially in the vision-language modality combination, visual information tends to be proven

to be ignored, resulting in MLLMs' over-reliance on the textual modality (Zhang et al., 2024). Thus, researchers naturally attempt to enhance visual modality's contribution in datasets to balance the information from different modalities. Typical cases include MMStar (Chen et al., 2024a) and MMMU-Pro (Yue et al., 2024b), where MMStar carefully selects visually dependent samples and MMMU-Pro not only filters out visually independent samples but also embeds questions into images. *Such works provide an optimization direction for current multimodal datasets. While few works contain a systematic index to evaluate the modality bias in datasets (Park et al., 2025), the others tend to prove the datasets' necessity based on MLLMs' disappointing results. More modality bias evaluation methods need to be explored to construct a better debias multimodal datasets.* Importantly, the feedback from MLLMs on these datasets should also be considered, as their performance can inform how datasets are optimized, offering valuable insights for future improvements.

② Turn the focus of MLLMs from textual information into visual information: Considering the ignorance of visual modality in the inference of MLLMs, it is an intuitive approach to force MLLMs to lay more emphasis on visual modality. Works such as (Liu et al., 2024c; Zhang et al., 2024) apply strategies, mostly training-free, to guide MLLMs towards visual modality. While Zhao *et al*.(Zhao et al., 2024b) proposed a novel framework to help MLLMs compress the influence of textual bias, enhancing visual modality across the model. *However, due to the excessive focus on the visual modality in such works, there's difficulty in applying them to broader modality bias situations. The real-world application requires the exploration of bias in richer modality combination.*

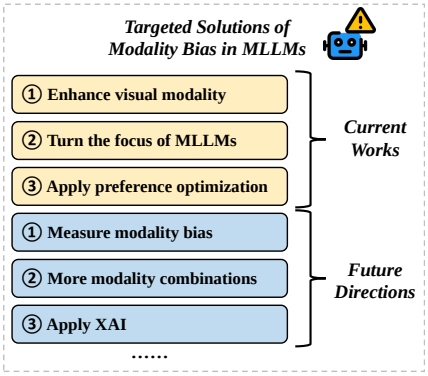

Figure 3: Targeted solutions of modality bias in MLLMs, including current works and future directions.

③ Apply preference optimization strategies: Besides the adjustments of multimodal datasets' content and MLLMs' focus, another popular method is to use a preference optimization strategy to correct modality bias internally (Pi et al., 2024; Zhang et al., 2025b). Pi *et al*. (Pi et al., 2024) built a preference dataset containing samples reflecting modality bias generated from the pretraining process. Zhang *et al*. (Zhang et al., 2025b) forced MLLMs to generate answers according to a specific modality through adding noise, thus creating the preference dataset. *Considering solving modality bias as a preference optimization goal is a creative and reasonable idea that brings new insights to researchers. However, existing preference optimization methods only generate modality bias samples in limited ways, while the real causes of modality bias are complex and multi-stage, which await further exploration.*

5.2 Future Directions

① Measure modality bias in MLLMs: The exploration of an objective and systematic metric to measure modality bias is crucial for the development of related research. For example, for dataset construction, a metric is needed as a flag that offers researchers a clear direction to make progress. Fields like semantic segmentation (Rahman & Wang, 2016; Rezatofighi et al., 2019) and image restoration (Hu et al., 2020) have seen a huge development with the existence and optimization of evaluation metrics, where modality bias in MLLMs still remains almost blank. Therefore, more research works are being called for regarding measuring modality bias in MLLMs.

② Explore modality bias in more modality combinations: Despite several works attempting to address the modality bias problem, the research focus is mainly set on the modality bias in LVLMs, which is part of the MLLMs. Although the textual information and visual information show great importance for world understanding (Xu et al., 2024), modalities like audio and tactile also matter (Liu et al., 2024b; Dave et al., 2024; Lyu et al., 2024a). As to the robotics field (Zhang et al., 2025a; Kirschner et al., 2025; Agarwal et al., 2025), tactile information is indispensable for robots to understand environments and handle downstream tasks such as dexterous manipulation (Gbagbe et al., 2024). Due to the modality limitation of current debiasing methods, it is hard for them to be applied in broader situations, hindering their applications in the real world. Thus, more generalized debiasing

strategies are required in real-world applications to handle conditions that are more complex and have more modalities besides images and texts.

③ **Apply XAI for modality bias in MLLMs:** Last but not least, finding the causes of modality bias in MLLMs and visualizing them will have a great positive influence on future works. Even though current works attempt to dig out the reasons for modality bias in MLLMs, they propose opinions from the phenomenon level. The internal mechanism of modality bias still lacks exploration, which is theoretical evidence and guidance to support future works. Thus, explainable AI (Bennetot et al., 2024; Dwivedi et al., 2023) is required here, such as visualizing the interaction process between modalities, to deeply analyze the theoretical causes and working mechanism of modality bias in MLLMs, which can be a more solid inspiration for future works.

6 Conclusion

This paper aims to highlight the phenomenon of modality bias in MLLMs and call for research work targeted at better integrating multiple modalities. Our position is that **MLLMs are deeply affected by modality bias**, which is proved and explored by both the theoretical analysis and case study. Moreover, we offer an in-depth discussion targeted at modality bias in MLLMs, including the key factors, potential results, and targeted solutions, hoping to bring new insights to the development of more robust and generalizable multimodal systems.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No.62506318); Guangdong Provincial Department of Education Project (Grant No.2024KQNCX028); CAAI-Ant Group Research Fund; Scientific Research Projects for the Higher-educational Institutions (Grant No.2024312096), Education Bureau of Guangzhou Municipality; Guangzhou-HKUST(GZ) Joint Funding Program (Grant No.2025A03J3957), Education Bureau of Guangzhou Municipality.

References

Arpit Agarwal, Achu Wilson, Timothy Man, Edward Adelson, Ioannis Gkioulekas, and Wenzhen Yuan. Vision-based tactile sensor design using physically based rendering. *Communications Engineering*, 4(1):21, 2025.

Ibrahim Alabdulmohsin, Xiao Wang, Andreas Peter Steiner, Priya Goyal, Alexander D'Amour, and Xiaohua Zhai. Clip the bias: How useful is balancing data in multimodal learning? In *The Twelfth International Conference on Learning Representations*.

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

Adrien Bennetot, Ivan Donadello, Ayoub El Qadi El Haouari, Mauro Dragoni, Thomas Frossard, Benedikt Wagner, Anna Sarranti, Silvia Tulli, Maria Trocan, Raja Chatila, et al. A practical tutorial on explainable ai techniques. *ACM Computing Surveys*, 57(2):1–44, 2024.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.

Tim Brödermann, Christos Sakaridis, Yuqian Fu, and Luc Van Gool. Cafuser: Condition-aware multimodal fusion for robust semantic perception of driving scenes. *IEEE Robotics and Automation Letters*, 2025.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024a.

Meiqi Chen, Yixin Cao, Yan Zhang, and Chaochao Lu. Quantifying and mitigating unimodal biases in multimodal large language models: A causal perspective. *arXiv preprint arXiv:2403.18346*, 2024b.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024c.

Ravinder S Dahiya, Giorgio Metta, Maurizio Valle, and Giulio Sandini. Tactile sensing—from humans to humanoids. *IEEE transactions on robotics*, 26(1):1–20, 2009.

Vedant Dave, Fotios Lygerakis, and Elmar Rueckert. Multimodal visual-tactile representation learning through self-supervised contrastive pre-training. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8013–8020. IEEE, 2024.

Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, et al. Explainable ai (xai): Core ideas, techniques, and solutions. *ACM Computing Surveys*, 55(9):1–33, 2023.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *Advances in Neural Information Processing Systems*, 37:89098–89124, 2024.

Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020.

Koffivi Fidèle Gbagbe, Miguel Altamirano Cabrera, Ali Alabbas, Oussama Alyunes, Artem Lykov, and Dzmitry Tsetserukou. Bi-vla: Vision-language-action model-based system for bimanual robotic dexterous manipulations. In *2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 2864–2869. IEEE, 2024.

Yangyang Guo, Liqiang Nie, Harry Cheng, Zhiyong Cheng, Mohan Kankanhalli, and Alberto Del Bimbo. On modality bias recognition and reduction. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(3):1–22, 2023.

Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022.

Bo Hu, Leida Li, Jinjian Wu, and Jiansheng Qian. Subjective and objective quality assessment for image restoration: A critical survey. *Signal Processing: Image Communication*, 85:115839, 2020.

Siteng Huang, Biao Gong, Yutong Feng, Xi Chen, Yuqian Fu, Yu Liu, and Donglin Wang. Learning disentangled identifiers for action-customized text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7797–7806, 2024.

Robin Jeanne Kirschner, Kübra Karacan, Alessandro Melone, and Sami Haddadin. Categorizing robots by performance fitness into the tree of robots. *Nature Machine Intelligence*, pp. 1–12, 2025.

Kang-il Lee, Minbeom Kim, Seunghyun Yoon, Minsung Kim, Dongryeol Lee, Hyukhun Koh, and Kyomin Jung. Vlind-bench: Measuring language priors in large vision-language models. *arXiv preprint arXiv:2406.08702*, 2024.

Sicong Leng, Yun Xing, Zesen Cheng, Yang Zhou, Hang Zhang, Xin Li, Deli Zhao, Shijian Lu, Chunyan Miao, and Lidong Bing. The curse of multi-modalities: Evaluating hallucinations of large multimodal models across language, visual, and audio. *arXiv preprint arXiv:2410.12787*, 2024.

Zichao Li, Xueru Wen, Jie Lou, Yuqiu Ji, Yaojie Lu, Xianpei Han, Debing Zhang, and Le Sun. The devil is in the details: Tackling unimodal spurious correlations for generalizable multimodal reward models. *arXiv preprint arXiv:2503.03122*, 2025.

Chenfei Liao, Kaiyu Lei, Xu Zheng, Junha Moon, Zhixiong Wang, Yixuan Wang, Danda Pani Paudel, Luc Van Gool, and Xuming Hu. Benchmarking multi-modal semantic segmentation under sensor failures: Missing and noisy modality robustness. *arXiv preprint arXiv:2503.18445*, 2025.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a.

Jing Liu, Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, and Jinhui Tang. Valor: Vision-audio-language omni-perception pretraining model and dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024b.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, pp. 125–140. Springer, 2024c.

Xiaoyuan Liu, Wenxuan Wang, Youliang Yuan, Jen-tse Huang, Qiuzhi Liu, Pinjia He, and Zhaopeng Tu. Insight over sight? exploring the vision-knowledge conflicts in multimodal llms. *arXiv preprint arXiv:2410.08145*, 2024d.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024e.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Yuanhuiyi Lyu, Xu Zheng, Dahun Kim, and Lin Wang. Omnibind: Teach to build unequal-scale modality interaction for omni-bind of all. *arXiv preprint arXiv:2405.16108*, 2024a.

Yuanhuiyi Lyu, Xu Zheng, Jiazhou Zhou, and Lin Wang. Unibind: Llm-augmented unified and balanced representation space to bind them all. In *CVPR*, pp. 26742–26752. IEEE, 2024b.

Muhammad Arslan Manzoor, Sarah Albarri, Ziting Xian, Zaiqiao Meng, Preslav Nakov, and Shangsong Liang. Multimodality representation learning: A survey on evolution, pretraining and its applications. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20 (3):1–34, 2023.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2023.

Jean Park, Kuk Jin Jang, Basam Alasaly, Sriharsha Mopidevi, Andrew Zolensky, Eric Eaton, Insup Lee, and Kevin Johnson. Assessing modality bias in video question answering benchmarks with multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 19821–19829, 2025.

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*, pp. 8238–8247, 2022.

Renjie Pi, Tianyang Han, Wei Xiong, Jipeng Zhang, Runtao Liu, Rui Pan, and Tong Zhang. Strengthening multimodal large language model with bootstrapped preference optimization. In *European Conference on Computer Vision*, pp. 382–398. Springer, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Md Atiqur Rahman and Yang Wang. Optimizing intersection-over-union in deep neural networks for image segmentation. In *International symposium on visual computing*, pp. 234–244. Springer, 2016.

Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):1964–1980, 2019.

Bin Ren, Yahui Liu, Yue Song, Wei Bi, Rita Cucchiara, Nicu Sebe, and Wei Wang. Masked jigsaw puzzle: A versatile position embedding for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20382–20391, 2023.

Bin Ren, Yawei Li, Jingyun Liang, Rakesh Ranjan, Mengyuan Liu, Rita Cucchiara, Luc V Gool, Ming-Hsuan Yang, and Nicu Sebe. Sharing key semantics in transformer makes efficient image restoration. *Advances in Neural Information Processing Systems*, 37:7427–7463, 2024a.

Bin Ren, Guofeng Mei, Danda Pani Paudel, Weijie Wang, Yawei Li, Mengyuan Liu, Rita Cucchiara, Luc Van Gool, and Nicu Sebe. Bringing masked autoencoders explicit contrastive properties for point cloud self-supervised learning. In *Proceedings of the Asian Conference on Computer Vision*, pp. 2034–2052, 2024b.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 658–666, 2019.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024.

Ali Vosoughi, Shijian Deng, Songyang Zhang, Yapeng Tian, Chenliang Xu, and Jiebo Luo. Cross modality bias in visual question answering: A causal view with possible worlds vqa. *IEEE Transactions on Multimedia*, 2024.

Yake Wei, Ruoxuan Feng, Zihe Wang, and Di Hu. Enhancing multimodal cooperation via sample-level modality valuation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27338–27347, 2024.

Junda Wu, Zhehao Zhang, Yu Xia, Xintong Li, Zhaoyang Xia, Aaron Chang, Tong Yu, Sungchul Kim, Ryan A Rossi, Ruiyi Zhang, et al. Visual prompting in multimodal large language models: A survey. *arXiv preprint arXiv:2409.15310*, 2024.

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.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Shaoxuan Xu, Menglu Cui, Chengxiang Huang, Hongfa Wang, and Di Hu. Balancebenchmark: A survey for multimodal imbalance learning. *arXiv preprint arXiv:2502.10816*, 2025.

Yang Yang, Fengqiang Wan, Qing-Yuan Jiang, and Yi Xu. Facilitating multimodal classification via dynamically learning modality gap. *Advances in Neural Information Processing Systems*, 37: 62108–62122, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024a.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024b.

Ningbin Zhang, Jieji Ren, Yueshi Dong, Xinyu Yang, Rong Bian, Jinhao Li, Guoying Gu, and Xiangyang Zhu. Soft robotic hand with tactile palm-finger coordination. *Nature Communications*, 16(1):2395, 2025a.

Yi-Fan Zhang, Weichen Yu, Qingsong Wen, Xue Wang, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. Debiasing multimodal large language models. *arXiv preprint arXiv:2403.05262*, 2024.

Zefeng Zhang, Hengzhu Tang, Jiawei Sheng, Zhenyu Zhang, Yiming Ren, Zhenyang Li, Dawei Yin, Duohe Ma, and Tingwen Liu. Debiasing multimodal large language models via noise-aware preference optimization. *arXiv preprint arXiv:2503.17928*, 2025b.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. Mmicl: Empowering vision-language model with multi-modal in-context learning. In *ICLR*, 2024a.

Haozhe Zhao, Shuzheng Si, Liang Chen, Yichi Zhang, Maosong Sun, Mingjia Zhang, and Baobao Chang. Looking beyond text: Reducing language bias in large vision-language models via multimodal dual-attention and soft-image guidance. *arXiv preprint arXiv:2411.14279*, 2024b.

Xu Zheng, Yuanhuiyi Lyu, Jiazhou Zhou, and Lin Wang. Centering the value of every modality: Towards efficient and resilient modality-agnostic semantic segmentation. In *European Conference on Computer Vision*, pp. 192–212. Springer, 2024a.

Xu Zheng, Haiwei Xue, Jialei Chen, Yibo Yan, Lutao Jiang, Yuanhuiyi Lyu, Kailun Yang, Linfeng Zhang, and Xuming Hu. Learning robust anymodal segmentor with unimodal and cross-modal distillation. *arXiv preprint arXiv:2411.17141*, 2024b.

Ding Zhong, Xu Zheng, Chenfei Liao, Yuanhuiyi Lyu, Jialei Chen, Shengyang Wu, Linfeng Zhang, and Xuming Hu. Omnisam: Omnidirectional segment anything model for uda in panoramic semantic segmentation. *arXiv preprint arXiv:2503.07098*, 2025.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, Hongfa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Caiwan Zhang, Zhifeng Li, Wei Liu, and Li Yuan. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *ICLR*. OpenReview.net, 2024.