Debias-DAPO: Mitigating Unimodal Bias in Multi-modal Large Language Models via Attention-Based Reward and Modality-Balanced Learning

Anonymous CVPR submission

Paper ID *****

Abstract

Multi-modal Large Language Models (MLLMs) have made significant progress in various applications but still suffer from unimodal bias, particularly "textual supremacy and visual neglect," leading to multimodal hallucinations. To address this challenge, we propose an enhancement to the DAPO algorithm, Debias-DAPO, which introduces attention-anchored hard negative mining and preference-shaping term that operates only over high-quality responses to correct answers that excessively rely on text. By calculating the ratio of per-token attention between text and visual modalities, we quantify unimodal bias. The reward function incentivizes responses grounded in vision while penalizing excessive text dependence, directly addressing harmful text biases. Meanwhile, by token-weighted advantage distribution using attention-derived evidence weights, optimization is refocused on key output tokens grounded in visual evidence. This ensures gradient updates prioritize visual information, encouraging the model to base its reasoning on images. Additionally, we introduce the first modality-balanced dataset, designed to overcome limitations in existing multimodal datasets. This dataset includes multimodal data where the multimodal response is correct, but unimodal responses are incorrect, along with manually verified contradictory image-text pairs and Chain of Thought (CoT) data. The goal is to encourage models to reason using both modalities effectively. Extensive experiments validate Debias-DAPO, demonstrating that it maintains visual grounding while effectively suppressing text bias drift, outperforming strong DAPO baselines. ß

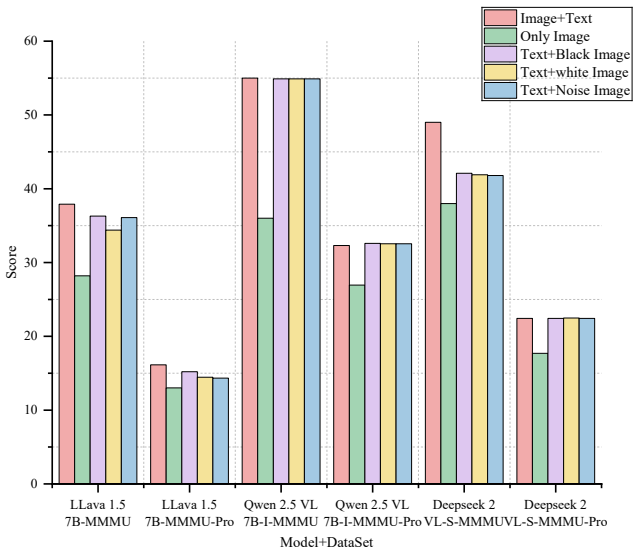
1. Introduction

Recent advancements in Multimodal Large Language Models (MLLMs) have led to significant progress across various tasks [**?**]. By integrating multiple modalities, such as text, audio, and images, these models have proven effective across diverse domains. However, a persistent challenge remains: MLLMs often exhibit unimodal bias, favoring one modality—typically text—while neglecting others. This bias can lead to over-reliance on a single modality, causing the model to generate responses that seem confident but are inaccurate or biased when information from less dominant modalities is missing or incorrect.

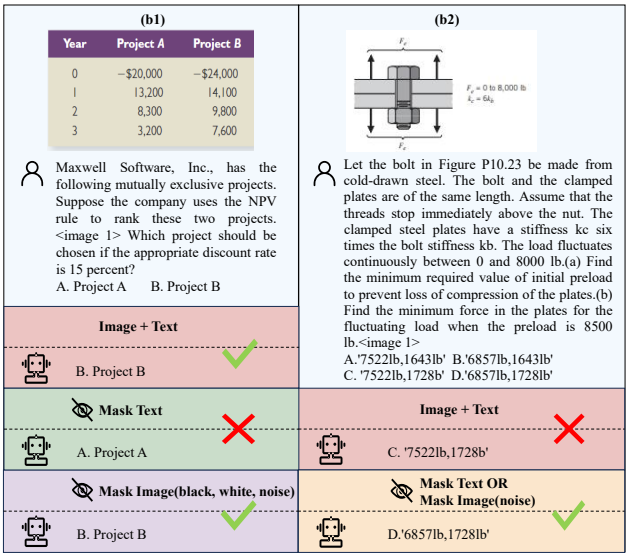
As illustrated in Fig. **??**(a), MLLMs commonly demonstrate uneven reliance on input modalities, with a clear text-dominant bias. For example, models like LLaVA ([11]), Qwen-VL ([2]), and DeepSeek-VL ([26]) produce correct answers with text-image inputs, but fail when only visual inputs are provided. In more extreme cases, models generate incorrect answers with text-image pairs but succeed with text-only inputs, ignoring the visual modality entirely (Fig. **??**(b2)). This paradox underscores the issue: the model relies excessively on text, sidelining visual information even when it is crucial for accuracy. These biases prevent the model from fully exploiting multimodal information and lead to overconfident errors when non-dominant modalities are absent or degraded. The ideal MLLM should integrate information from all modalities to ensure robust, accurate responses in a variety of conditions.

To mitigate this issue, existing approaches have attempted to balance modality usage by altering dataset distributions [32], applying causal interventions [3], and debiasing during training [16]. However, these solutions often require large-scale supervised fine-tuning and show limited generalization. To address these limitations, we introduce the **first dedicated unimodal bias dataset**, designed to address current benchmarks' shortcomings and facilitate better multimodal integration.

Building on existing datasets like A-OKVQA [20], VisRAG-Ret-Test-ArxivQA [27], and TextVQA [22], we develop a novel data construction methodology. Our approach introduces modal conflicts by perturbing complementary modalities, forcing models to reason effectively in scenarios involving unimodal bias and conflict. The resulting **Debiased Multimodal Dataset** includes two crucial subsets: multimodal correct data and modality-conflicting samples. The multimodal correct subset retains only sam-



(a) Dataset vs. accuracy



(b) Ablation study

Figure 1. Overall caption for the two panels.

ples that require simultaneous use of both text and image information to generate correct answers. The modality-conflicting subset consists of data where the visual and textual information contradict each other, manually verified for quality.

To further tackle unimodal bias, we extend the DAPO framework ([**?**]) by introducing a bias-specific reward function. This reward function integrates modality-balance and task-accuracy rewards, encouraging models to focus on both visual and textual information. Additionally, we introduce **Chain-of-Thought (CoT)** reasoning to guide the model's stepwise decision-making, improving robustness in challenging tasks.

We evaluate our approach using models such as Qwen2.5-VL-7B-Instruct [2], LLaVA-1.5 [11], and DeepSeek-VL2-small [26]. We train the models using the generated Debiased Multimodal Dataset and assess performance on both VLind-Bench [9] and standard hallucination benchmarks, including MMHalBench [23], AMBER [24], and CHAIR [18]. Our method significantly outperforms baseline models like NAPO ([29]) in terms of accuracy and hallucination reduction, demonstrating the effectiveness of the **Debiased Multimodal Dataset** and the enhanced DAPO algorithm.

2. Related work

2.1. Unimodal bias in MLLMs

In recent years, significant advancements have been made in MLLMs([6]),[4],[8]. Unimodal bias occurs when models prioritize simple patterns within a single modality while neglecting the strengths of other modalities, leading to performance degradation and hallucinations when modalities are missing during inference [33]. For instance, prompt text significantly influences the performance of many state-of-the-art multimodal models. Model performance improves substantially when textual prompts are included, but degrades drastically without this modality. In recent years, significant advancements have been made in MLLMs([6]),[4],[8]. Unimodal bias occurs when models prioritize simple patterns within a single modality while neglecting the strengths of other modalities, leading to performance degradation and hallucinations when modalities are missing during inference [33]. For instance, prompt text significantly influences the performance of many state-of-the-art multimodal models. Model performance improves substantially when textual prompts are included, but degrades drastically without this modality. Various methods have been proposed to quantify and mitigate unimodal bias in VQA tasks, with a focus on balancing datasets [32] and complex training strategies [7] [31] [30]. Recent research includes modality bias evaluation benchmarks [9], datasets for modality bias [15], weight adjustments [12], prompting strategies [5], Preference Optimization [31], [30], and decoding configurations to reduce language priors [14] [28]. Preference Optimization fine-tunes models with human feedback to align outputs with human expectations. Recent work, like Noise-robust Alignment with Preference Optimization (NAPO) [29], incorporates noise-robust mechanisms into the Direct Preference Optimization (DPO) framework [17], which corrects errors but does not fundamentally change the model's capabilities. The Shortcut-aware MM-RM algorithm [**?**] mitigates out-of-distribution

generalization deficits caused by over-reliance on unimodal textual correlations. *By contrast, we introduce a dataset for determining whether the modality is balanced, while optimizing the reward model and sampling mechanism of the **GRPO** [21] algorithm to mitigate unimodal bias and hallucination.*

2.2. MLLMs Post Training with RL

Reinforcement learning (RL) develops intelligent agents by learning optimal policies through trial-and-error interactions to maximize rewards. Recently, RL has seen significant progress in multimodal large models. Proximal Policy Optimization (PPO) ([19]) is a policy-gradient algorithm that stabilizes training by constraining policy update magnitudes. ReMax is an off-policy RL algorithm that optimizes the policy via a "relative policy improvement" mechanism, avoiding the complex and unstable evaluation steps of conventional methods [10]. More recently, Grouped Relative Policy Optimization (GRPO) [21] improves policy optimization efficiency by using grouped updates and relative entropy constraints, reducing repetitive calculations and balancing exploration and exploitation. Despite these advancements, a critical issue remains: modality imbalance, which can significantly impact the contribution of each modality to the reward function.

3. Method

3.1. Preliminary

3.1.1. DAPO

DAPO (Decoupled Clip and Dynamic Sampling Policy Optimization) [**?**] is an open-source reinforcement learning system and engineering framework designed to support efficient, reproducible, and scalable complex reasoning tasks with large language models (LLMs). The efficacy of the proposed approach is comparable to or superior to conventional methods in terms of performance, whilst concomitantly enhancing the efficiency of training and reducing computational expenses. It is noteworthy that the DAPO model outperforms the GRPO model despite utilising a mere half of the training steps.

While maintaining GRPO's utilisation of intra-group relative rewards and a clipped objective, DAPO incorporates four pivotal engineering enhancements: The following techniques are employed: clip-higher dynamic sampling, token-level loss, and overlong reward shaping. The KL penalty term is also removed. These innovations effectively address several limitations of GRPO, including entropy collapse, gradient wastage, and weight dilution in long-horizon reasoning tasks, leading to more stable and efficient training.

As a result, DAPO delivers superior performance, accelerated training convergence, enhanced stability, and improved reproducibility. Its efficacy has been demonstrated through successful implementation in a range of large-scale model deployments [**?**]. The objective function of DAPO is given by:

$$\mathcal{J}_{\text{DAPO}}(\theta) = \mathbb{E}\big[(q,a)\sim\mathcal{D},\ \{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)\big] \left[\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\min\Big(r_{i,t}(\theta)\hat{A}_{i,t},\ \text{clip}\Big(r_{i,t}(\theta),1-\varepsilon_{\text{low}},1+\varepsilon_{\text{high}}\Big)\hat{A}_{i,t}\Big)\right] \tag{1}$$

$$\text{s.t.}\ \ 0 < \Big|\{o_i \mid \texttt{is\_equivalent}(a,o_i)\}\Big| < G$$

where

$$r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t}\mid q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}\mid q,o_{i,<t})}, \quad \hat{A}_{i,t} = \frac{R_i - \text{mean}(\{R_i\}_{i=1}^{G})}{\text{std}(\{R_i\}_{i=1}^{G})} \tag{2}$$

Randomly sample a question-answer pair (q, a) from the training dataset $\mathcal{D}$. Using the old policy model $\pi_{\theta_{\text{old}}}$, generate a set of G responses for question q. $\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}$ indicates we will compute the loss for each token in every response, then normalize by the total number of tokens across all responses. $r_{i,t}$ measures the change in token selection probability between the new and old policies. The advantage function $\hat{A}_{i,t}$ measures how well the decision to generate this token performs relative to the average, given the question and historical context. The $\min\Big(r_{i,t}(\theta)\hat{A}_{i,t}, \text{clip}\Big(r_{i,t}(\theta),1-\varepsilon_{\text{low}},1+\varepsilon_{\text{high}}\Big)\hat{A}_{i,t}\Big)$ operation decouples and prunes the model. By actively encouraging exploration with a larger upper bound ε_{low} while strictly constraining the maximum step size of policy updates with a more conservative lower bound ε_{low}, it achieves a critical balance between improving performance and preventing training collapse, ensuring the stability of the training process.

3.1.2. Reward Function

DAPO employs a fully rule-driven reward function, moving away from the often unreliable neural network-based reward models typically used in traditional reinforcement learning. The fundamental principle of the method is to assign a binary reward of positive one or negative one, based solely on the final correctness of the generated answer. This straightforward reward scheme ensures strong alignment between the optimization objective and the task goal, effectively circumventing reward hacking. In order to regulate the output length, DAPO incorporates an adaptive length penalty mechanism. This mechanism progressively penalises responses that exceed a predefined threshold, thereby encouraging the model to generate concise answers without interfering with valid reasoning processes.

During the forward pass, DAPO deploys a meticulously structured batch sampling and training procedure. For each query, multiple responses are generated dynamically, and sample groups that cannot contribute meaningful gradient signals are filtered out strategically; this may include those that are entirely correct or entirely incorrect. Consequently, it is assured that every parameter update is informed by data of significant value. The advantage for each response is then computed using group-wise reward normalization, thereby eliminating the need for a complex value function. In the subsequent loss computation phase, DAPO employs a novel approach by implementing token-level policy optimisation and utilising decoupled clipping on importance sampling ratios. The amalgamation of these design choices ensures elevated stability and sample efficiency throughout the course of large-scale reinforcement learning training.

3.2. Debiased Multimodal Datasets

3.2.1. Overview

In this paper, we propose the first open-source resource designed dataset for achieving modality balance, namely Debiased Multi-Modal dataset **(DMM)**.

Recent research has revealed that unimodal bias arises from a model's over-reliance on one modality (textual or visual) while neglecting the other. To address this problem, we focus on enhancing inter-modal correlations and incorporates scenarios where modalities conflict. Extending beyond conventional visual question answering, our DMM dataset includes samples that require both visual and textual information for correct responses — cases that a single modality cannot adequately address. Meanwhile, our DMM dataset introduces conflicting samples where the textual and visual content diverge, thereby establishing a foundation for reward mechanisms based on modal equilibrium.

For each bimodal data sample and its corresponding conflicting sample in the training set, Chain-of-Thought (COT) [25] data is generated using GPT-4.1 [1]. The Debias dataset builds upon A-OKVQA [20], VisRAG-Ret-Test-ArxivQA[27], and TextVQA [22], covering a wide range of topics, including commonsense reasoning, encyclopedic knowledge, diagram interpretation, and physical principles. The dataset is partitioned into training and test sets containing 12,000 and 1,200 visual question–answer pairs respectively. Images are provided in JPG and PNG formats with resolutions ranging from 448 to 1024 pixels. In total, the dataset consists of 13,200 visual question-answer pairs. 91.67% of these represent unbiased examples. The remaining 8.33% comprise modality-conflicting instances.

3.2.2. Dataset Construction

Data Design for Enhanced Modality Complementarity. We introduce a filtering process based on the OK-VQA ([13]), A-OKVQA ([20]), VisRAG-Ret-Test-ArxivQA[27]

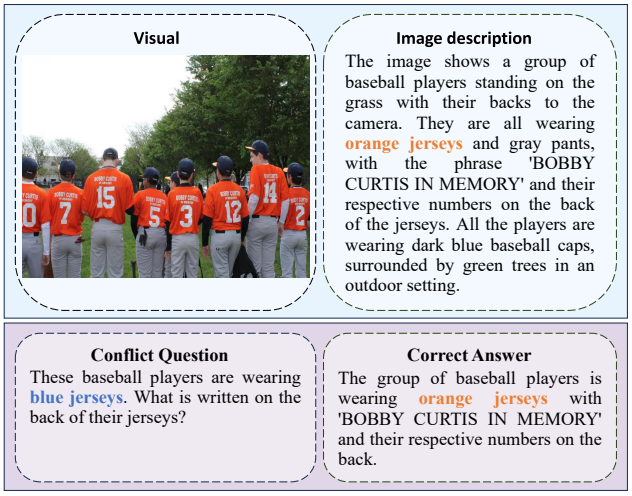


Figure 2. Put your overall caption here.

datasets. First, we utilize the original images and text questions to identify data points with correct answers. Next, we test the model using only the images. The original text questions were replaced with a simple prompt such as 'Please answer the question based on the image.' We then tested using only the text. The original images were replaced with black images. This process removed data that could be answered correctly using only one modality (either images or text alone). We specifically target data requiring combined information from both modalities. For example, a text question might ask, 'Is the rabbit on the chair?', when the image clearly shows the rabbit on a table. Answering this correctly requires using the image to correct the text's premise. Ultimately, 3,310 data points resulted from the filtering.

Modality-Conflict Samples. Through initial experiments, we observe that the reward mechanism fails to converge due to the input of biased data. This results in the model's performance fluctuating unpredictably during training. To encourage large multi-modal models to rely more consistently on reliable modalities, we create a specialized set of samples. Based on the TextVQA dataset [22], we employ **GPT 4.1** to generate conflict samples in which the textual questions contradict the visual content. The aim is to enable multi-modal models to detect such inconsistencies through visual cues, allowing them to correct textual misinformation based on visual evidence rather than simply adhering to the text. The visual attributes in these samples are derived solely from objective image information and serve as the ground truth for identifying contradictions. The textual questions are designed to introduce conflicts centered around these samples, including attribute contradictions, object existence conflicts and state or action discrepancies. For instance, for an image showing a group of baseball players standing on the grass, all wearing orange

jerseys with the text 'BOBBY CURTIS IN MEMORY' and their respective numbers on the back, the corresponding question might be: 'These baseball players are wearing blue jerseys. What is written on the back of their jerseys?' A robust model should recognise the discrepancy between the textual reference to 'blue jerseys' and the orange jerseys evident in the image, and respond accurately based on the image. The expected answer would be: 'The group of baseball players is wearing orange jerseys with 'BOBBY CURTIS IN MEMORY' and their respective numbers on the back.' All generated conflict samples are textbf manually verified to ensure that the contradictions could be identified and the answers corrected visually. This process results in a high-quality conflict sample set consisting of 1,000 training samples and 100 test samples.

3.3. Debiased Framework

Debias-DAPO augments DAPO with two complementary components: (i) an attention-anchored hard negative mining and preference-shaping term that operates only over high-quality responses, and (ii) token-weighted advantage distribution using attention-derived evidence weights. We use attention patterns as a heuristic proxy for grounding strength, acknowledging that they reflect correlation rather than causation. The complete training objective combines both components with the base DAPO loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{DAPO}}(\theta; \tilde{A}_{i,t}) + \lambda \mathcal{L}_{\text{pref}}(\theta), \tag{3}$$

where $\mathcal{L}_{\text{DAPO}}$ uses token-weighted advantages $\tilde{A}_{i,t}$ (Eq. 8) instead of uniform sequence-level advantages, and λ controls the strength of preference regularization.

3.3.1. Attention-Anchored Hard Negative Mining and Preference-Shaping Term

Visual Grounding Ratio During rollout generation, we compute a Visual Grounding Ratio (VGR) using all layers and heads while filtering out special/system tokens (BOS, system/role markers, PAD/EOS), so the prompt side contains only visual tokens (V) and question tokens (T). For each layer–head (ℓ, h), take the generation×prompt attention slice and row-normalize it within $(V \cup T)$: $\tilde{\alpha}^{(\ell,h)}_{t\to k} = \alpha^{(\ell,h)}_{t\to k} / \sum_{k' \in V\cup T} \alpha^{(\ell,h)}_{t\to k'}$. Let G be the set of generated tokens. We aggregate across all (ℓ, h, t) compactly via a mean operator to obtain modality masses:

$$\widehat{A}_V = \operatorname*{mean}_{\ell,h,t}\left(\sum_{k\in V} \tilde{\alpha}^{(\ell,h)}_{t\to k}\right), \qquad \widehat{A}_T = \operatorname*{mean}_{\ell,h,t}\left(\sum_{k\in T} \tilde{\alpha}^{(\ell,h)}_{t\to k}\right), \tag{4}$$

and define the VGR by normalizing with token counts:

$$\text{VGR} = \frac{\widehat{A}_V/|V|}{\widehat{A}_T/|T|}. \tag{5}$$

To reduce confounds from sequence length and question difficulty, we apply a **per-question monotone normalization** to raw VGR scores. Specifically, within each question's response group (the G responses generated for that question), we apply rank-based quantile transformation to map VGR values to a standard distribution, producing the normalized score s_i used in Eq. 6. This monotone transformation preserves the relative ordering of responses while controlling for question-specific variations in attention patterns.

Within-positive preference shaping. Within each question's high-quality response set $\mathcal{P} = \{y_i\}$ (responses receiving positive rewards in DAPO's rule-based reward function, i.e., correct answers with valid formatting), we use the group-normalized VGR score s_i to identify text-biased positives as hard negatives and to enforce a margin that favors vision-grounded positives—without relabeling. *While these responses are correct in terms of final answers, they are undesirable from a grounding perspective—achieving accuracy through textual shortcuts rather than genuine multimodal reasoning. We thus treat them as negatives in the preference ranking to discourage shortcut learning.* Concretely, we form ordered pairs by sampling (i) from the top-K highest s (more vision-grounded) and (j) from the bottom-K lowest s (more text-biased), and optimize a pairwise logistic-margin loss:

$$\mathcal{L}_{\text{pref}} = \mathbb{E}_{(i,j)\sim\mathcal{M}(\mathcal{P})}\Big[-\log\sigma\Big(\beta\big((s_i - s_j) - m\big)\Big)\Big], \tag{6}$$

where $\mathcal{M}(\mathcal{P})$ denotes the within-positive pairing policy, $m \geq 0$ is a margin, and β is a temperature. This attention-anchored hard-negative preference-shaping term operates only within high-quality responses, complements the base DAPO objective, and directly suppresses textual shortcuts by ensuring that vision-grounded positives consistently outrank text-biased ones. (An optional listwise variant uses $\tilde{p}_i = \exp(\lambda s_i)/\sum_{y\in\mathcal{P}}\exp(\lambda s_y)$ as a soft target and minimizes cross-entropy against the policy over $\mathcal{P}$.)

3.3.2. Token-Weighted Advantage Distribution using Attention-Derived Evidence Weights

To further refine gradient allocation, we distribute the sequence-level advantage $\hat{A}_i$ obtained from DAPO to individual output tokens according to normalized attention-derived evidence weights. For each generated token t, we compute an attention-based evidence weight using the row-normalized attention $\tilde{\alpha}_{t\to k}$ (averaged across layers and heads) that measures its grounding in the visual context:

$$w_{i,t} = \frac{\sum_{k\in V}\bar{\alpha}_{t\to k}}{\sum_{k\in V\cup T}\bar{\alpha}_{t\to k}}, \tag{7}$$

where $\bar{\alpha}_{t\to k} = \frac{1}{LH}\sum_{\ell,h}\tilde{\alpha}^{(\ell,h)}_{t\to k}$ denotes the layer-head averaged normalized attention from token t to prompt token

k. These weights are normalized within each sequence:

$$\tilde{A}_{i,t} = \frac{w_{i,t}}{\sum_{t'} w_{i,t'}} \cdot \hat{A}_i. \tag{8}$$

The two-stage normalization serves distinct purposes: $w_{i,t}$ in Eq. 7 measures each token's proportional reliance on visual context relative to all prompt tokens, while the sequence-level renormalization in Eq. 8 ensures that token-level advantages sum to the original sequence advantage $\hat{A}_i$, preserving DAPO's advantage scaling properties. This token-weighted advantage distribution routes gradients toward evidence-bearing tokens that contribute to grounded reasoning, enhancing fine-grained credit assignment and complementing the within-positive preference-shaping term.

4. Experiments

4.1. Datasets

4.2. Training Details

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 4

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. 1, 2

[3] Meiqi Chen, Yixin Cao, Yan Zhang, and Chaochao Lu. Quantifying and mitigating unimodal biases in multimodal large language models: A causal perspective, 2024. 1

[4] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks, 2024. 2

[5] Yusheng Dai, Hang Chen, Jun Du, Ruoyu Wang, Shihao Chen, Haotian Wang, and Chin-Hui Lee. A study of dropout-induced modality bias on robustness to missing video frames for audio-visual speech recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27445–27455, 2024. 2

[6] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. 2

[7] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Fushuo Huo, Jinyu Chen, and Song Guo. Overcome modal bias in multimodal federated learning via balanced modality selection. In *European Conference on Computer Vision*, pages 178–195. Springer, 2024. 2

[8] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino,

Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun

Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. 2

[9] 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, 2025. 2

[10] Ziniu Li, Tian Xu, Yushun Zhang, Zhihang Lin, Yang Yu, Ruoyu Sun, and Zhi-Quan Luo. Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models, 2024. 3

[11] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1, 2

[12] 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*, pages 125–140. Springer, 2024. 2

[13] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4

[14] Kyungmin Min, Minbeom Kim, Kang-il Lee, Dongryeol Lee, and Kyomin Jung. Mitigating hallucinations in large vision-language models via summary-guided decoding. *arXiv preprint arXiv:2410.13321*, 2024. 2

[15] 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*, pages 19821–19829, 2025. 2

[16] Vaidehi Patil, Adyasha Maharana, and Mohit Bansal. Debiasing multimodal models via causal information minimization, 2023. 1

[17] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023. 2

[18] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning, 2019. 2

[19] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 3

[20] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer, 2022. 1, 4

[21] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 3

[22] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 1, 4

[23] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented rlhf, 2023. 2

[24] Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation, 2024. 2

[25] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 4

[26] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024. 1, 2

[27] Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, et al. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*, 2024. 1, 4

[28] 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. 2

[29] 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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9423–9433, 2025. 2

[30] Xu Zheng, Yuanhuiyi Lyu, and Lin Wang. Learning modality-agnostic representation for semantic segmentation from any modalities. In *European Conference on Computer Vision*, pages 146–165. Springer, 2024. 2

[31] 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*, pages 192–212. Springer, 2024. 2

[32] Xu Zheng, Chenfei Liao, Yuqian Fu, Kaiyu Lei, Yuanhuiyi Lyu, Lutao Jiang, Bin Ren, Jialei Chen, Jiawen Wang, Chengxin Li, Linfeng Zhang, Danda Pani Paudel, Xuanjing Huang, Yu-Gang Jiang, Nicu Sebe, Dacheng Tao, Luc Van Gool, and Xuming Hu. Mllms are deeply affected by modality bias, 2025. 1, 2

[33] Xu Zheng, Yuanhuiyi Lyu, Lutao Jiang, Danda Pani Paudel, Luc Van Gool, and Xuming Hu. Reducing unimodal bias in multi-modal semantic segmentation with multi-scale functional entropy regularization. *arXiv preprint arXiv:2505.06635*, 2025. 2