SMART: Sink-based Modality-Aware Redistribution of Transformer Attention

Anonymous ACL submission

Abstract

Multimodal Large Language Models (MLLMs) often exhibit *modality bias*, where predictions over-rely on one modality while underutilizing others. However, the mechanism underlying modality bias remains poorly understood. In this work, we first introduce **Banana-Counting**, a controlled diagnostic dataset for measuring modality preference under cross-modal conflicts. Evaluations across diverse MLLMs reveal pervasive modality imbalance. Then, through attention analysis, we show that modality bias is strongly associated with imbalanced cross-modal attention allocation and modality-specific attention sinks that absorb redundant attention mass. Motivated by these findings, we propose **SMART**, an inference-time method that mitigates modality bias by redistributing redundant sink attention toward under-attended modalities. SMART consistently improves multimodal balance across model families. On Banana-Counting, it reduces the modality gap from 27.73 to 0.66 on LLaVA-v1.5-7B while improving balanced multimodal accuracy. Moreover, SMART generalizes effectively to downstream multimodal reasoning tasks without retraining or parameter updates.

1 Introduction

Multimodal Large Language Models (MLLMs) have achieved remarkable progress on vision–language tasks such as visual question answering, image captioning, and multimodal reasoning by integrating pretrained large language models with visual encoders (Liu et al., 2024a; OpenAI et al., 2023; Wang et al., 2023a). Despite these advances, recent studies suggest that MLLMs often fail to utilize multimodal information in a balanced manner (Chen et al., 2024b; Guo et al., 2023). Instead, predictions may disproportionately rely on a single modality while underutilizing information from others, leading to what is commonly referred to as *modality bias*.

Modality bias poses a fundamental challenge for multimodal reasoning. When one modality dominates generation, MLLMs become vulnerable to noisy, misleading, or conflicting cross-modal signals, reducing robustness and limiting effective multimodal integration (Park et al., 2025; Wu et al., 2025; Zhang et al., 2023).

A major obstacle is the lack of controlled diagnostic datasets for isolating modality preference from unimodal capability. Existing multimodal datasets primarily evaluate task accuracy or reasoning performance, but they do not explicitly measure which modality a model prefers when conflicting evidence simultaneously exists across modalities.

To address this limitation, we introduce **Banana-Counting**, a controlled diagnostic dataset for probing modality bias in MLLMs. The core design principle is to introduce intentionally conflicting numerical cues across textual and visual modalities, thereby enabling a fine-grained analysis of modality preference during generation. Across a diverse set of MLLMs, we observe pervasive modality imbalance, with many models exhibiting large accuracy gaps between textual and visual information despite strong unimodal capability.

To understand this phenomenon, we analyze cross-modal attention dynamics during autoregressive decoding. Our analysis shows that modality bias is closely associated with imbalanced attention allocation across modalities, where attention progressively concentrates on the dominant modality. We further find that a substantial portion of attention is absorbed by low-information *attention sink* tokens, which act as redundant reservoirs and further suppress effective cross-modal interaction.

Motivated by these findings, we propose **SMART** (**S**ink-based **M**odality-**A**ware

Redistribution of **T**ransformer Attention), an inference-time method for mitigating modality bias in MLLMs. SMART identifies modality-specific sink regions and redistributes redundant attention mass toward informative tokens in the under-attended modality. Importantly, SMART is a training-free method that operates entirely during inference, requiring neither retraining nor parameter updates.

Extensive experiments demonstrate that SMART consistently reduces modality imbalance across different MLLM families while improving balanced multimodal performance. On the Banana-Counting dataset, SMART reduces the modality gap from 27.73 to 0.66 on LLaVA-v1.5-7B. Furthermore, SMART generalizes effectively to downstream multimodal reasoning tasks while introducing only modest inference overhead.

Our contributions are summarized as follows:

- We introduce **Banana-Counting**, a controlled diagnostic dataset for quantifying modality bias under cross-modal conflicts.
- We provide a mechanistic analysis showing that modality bias is strongly associated with progressively amplified cross-modal attention imbalance and modality-specific attention sinks.
- We propose **SMART**, an inference-time attention redistribution method that mitigates modality bias, while consistently improving multimodal balance and transferring effectively to downstream multimodal reasoning benchmarks.

2 Related Work

Modality Bias in Multimodal Models. Modality bias arises when a model overly relies on one modality while under-utilizing others, manifesting as either *text bias*, where textual priors dominate reasoning, or *visual bias*, where visual cues overshadow textual information (Wang et al., 2020; Huang et al., 2022; Lin et al., 2023). This phenomenon weakens generalization and can lead to hallucination-like failures (Zhang et al., 2025). Prior studies have extensively explored modality bias in VQA tasks (Niu et al., 2021; Chen et al., 2024b; Guo et al., 2023), or specifically investigated visual prior or textual priors. Existing debiasing approaches mainly rely on training-time strategies, including contrastive learning (Luo et al., 2024), self-improving data augmentation (Lee et al., 2025), and preference optimization (Wang et al., 2024; Zhang et al., 2025). These methods can reduce modality-specific biases but often require costly retraining, task-specific datasets, or complex optimization, limiting their general applicability to large-scale MLLMs. In contrast, our work is the first to systematically evaluate modality bias in MLLMs and to propose a lightweight, inference-time approach that mitigates bias across models and tasks without requiring retraining or task-specific datasets.

Attention Sink Phenomenon. Attention sinks, first observed in LLMs, are low-information tokens (e.g., BOS, punctuation, or newlines) that consistently receive disproportionately high attention despite contributing little to predictions (Xiao et al., 2023; Ferrando and Voita, 2024; Kobayashi et al., 2020; Bondarenko et al., 2023). This effect arises from abnormally large activations in specific hidden dimensions (Sun et al., 2024; Cancedda, 2024), and has been exploited in LLMs to recalibrate attention and improve accuracy (Yu et al., 2024). Similar behavior has been found in vision transformers, where background or irrelevant visual patches attract excessive attention (Darcet et al., 2023). Therefore, attention sink is a common phenomenon across different modalities. Methods like VAR (Kang et al., 2025) reuse these surplus attention resources to enhance visual reasoning without retraining. However, no studies leverage attention sinks to mitigate *modality bias*. Our approach fills this gap by redistributing attention from sinks to under-attended modalities, addressing the root cause of bias in MLLMs.

3 Diagnosing Modality Bias

To study modality bias in MLLMs, we require a controlled benchmark that measures modality preference under conflicting multimodal signals. Existing multimodal datasets mainly evaluate reasoning, perception, or hallucination, but they do not explicitly measure modality preference when inconsistent information simultaneously exists across modalities.

To address this limitation, we introduce **Banana-Counting**, a diagnostic dataset designed to quantify modality bias under controlled multimodal conflicts. Specifically, we construct paired cross-modal contradictions by injecting inconsistent numerical needle information into textual and visual inputs, thereby disentangling modality preference from intrinsic unimodal capability and enabling systematic analysis of model behavior un-

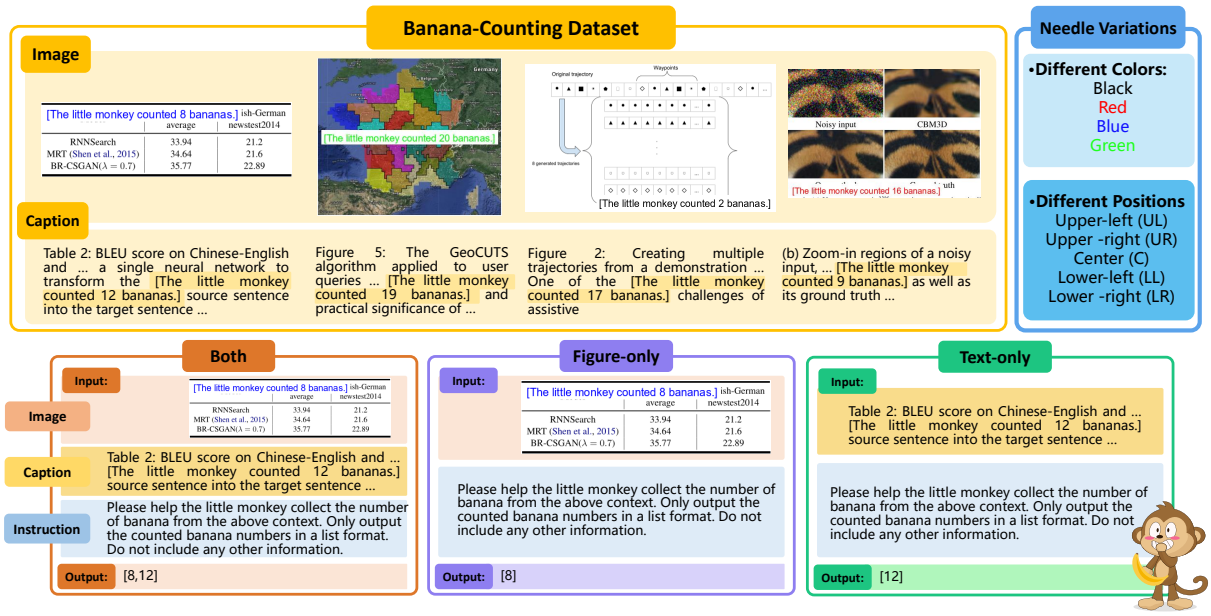

Figure 1: Overview of the Banana-Counting Dataset. We extracted figures and captions from the SPIQA dataset and inserted needle phrases into both modalities. The needles color and position in figures were randomly assigned while in text, it was inserted at random positions to avoid biases. To evaluate modality bias, we designed three settings: (1) Both: The model receives both figure and caption, (2) Figure-only: The model receives only the figure, and (3) Text-only: The model receives only the caption. This setup examines whether MLLMs effectively integrate multimodal information or favor one modality over the other.

der conflicting multimodal evidence.

3.1 Diagnostic Dataset Construction

Banana-Counting is constructed based on SPIQA (Pramanick et al., 2024), a scientific figure question-answering benchmark containing figures, captions, and associated textual context from research papers.

For each sample, we insert conflicting numerical needle phrases into both the image and text modalities using the template:

The little monkey counted {number} bananas.

where the number is randomly sampled from 1 to 20 independently for each modality.

To avoid superficial shortcuts, the insertion position and visual appearance of the needle are randomized. Construction details are provided in Appendix A. The final dataset contains 1026 evaluation samples. Figure 1 illustrates the overall construction pipeline.

3.2 Experimental Setup

We evaluate models under three settings:

(1) **Both**: both image and text are provided;
(2) **Figure-only**: only the image is provided;
(3) **Text-only**: only the text is provided.

The Figure-only and Text-only settings evaluate unimodal capability, while the Both setting serves as the core diagnostic task. Since conflicting banana counts simultaneously exist in both modalities, the model's prediction directly reveals its modality preference.

We quantify modality bias using the results in **Both** setting: (1) textual accuracy ACC_{text}, (2) visual accuracy ACC_{fig}, (3) modality gap

$$\Delta = ACC_{\text{text}} - ACC_{\text{fig}},$$

and (4) balanced accuracy

$$ACC_{\text{bal}} = \min(ACC_{\text{text}}, ACC_{\text{fig}}).$$

We evaluate representative MLLMs including MiniCPM-V (Yao et al., 2024), Qwen2-VL (Bai et al., 2023), CogVLM (Wang et al., 2023b), GPT-4o-mini (OpenAI et al., 2023), LLaVA series (Liu et al., 2024a), MoE-LLaVA (Lin et al., 2026), DeepSeek-VL2 (Wu et al., 2024), Yi-VL (Young et al., 2024) and InternVL2 (Chen et al., 2024c).

3.3 Results and Discussion

Table 1 presents the evaluation results on Banana-Counting. We observe pervasive modality bias across modern MLLMs. Most models strongly prefer visual information when both modalities are

Model	Model Size #Parameters (B)	Both		Figure only	Text only
		$\mathbf{ACC_{text}}$	$\mathbf{ACC_{fig}}$	$\mathbf{ACC_{fig}}$	$\mathbf{ACC_{text}}$
MiniCPM-V-2_6 (Yao et al., 2024)	8.10	67.17	**74.93**	92.14	98.48
Qwen2-VL-72B (Bai et al., 2023)	72.00	77.17	**89.96**	99.51	100.00
Qwen2-VL-Instruct (Bai et al., 2023)	7.00	51.66	**72.32**	98.05	99.90
Cogvlm2-Llama3-chat (Wang et al., 2023b)	19.00	**67.82**	58.77	96.80	91.29
GPT-4o-mini (OpenAI et al., 2023)	-	81.52	**87.15**	99.51	100.00
Llava-v1.5-7B (Liu et al., 2024a)	7.00	27.79	**55.52**	58.87	83.82
Llava-v1.5-13B (Liu et al., 2024a)	13.00	**33.48**	19.57	38.01	81.60
Llava-v1.6-vicuna-13B (Liu et al., 2024a)	13.00	30.52	**53.46**	68.13	95.32
Llava-next-Llama3 (Liu et al., 2024a)	8.00	**47.95**	33.72	42.50	93.86
MoE-LLaVA-Phi2-2.7B-4e (Lin et al., 2026)	5.61	22.71	**96.20**	52.05	88.11
MoE-LLaVA-Phi2-2.7B-4e-384 (Lin et al., 2026)	5.73	38.15	**82.07**	92.88	78.65
Deepseek-vl2-small (Wu et al., 2024)	16.10	8.00	**78.17**	90.55	98.83
Yi-VL-6B (Young et al., 2024)	6.00	**77.78**	27.17	48.54	90.55
InternVL2-8B (Chen et al., 2024c)	8.00	58.87	**78.26**	89.67	99.90

Table 1: Performance of different LLMs on the Banana-Counting dataset. The table reports the accuracy of various models in identifying the number of bananas under three settings: (1) **Both** (2) **Image only**, and (3) **Text only**. **#Parameters** denotes the model size in billions. $\mathbf{ACC_{text}}$ and $\mathbf{ACC_{fig}}$ represent the accuracy of extracting the banana count from text and images, respectively. The results highlight significant **modality bias**, where most models favor image-based information ($\mathbf{ACC_{fig}}$) while often overlooking text-based cues ($\mathbf{ACC_{text}}$).

available, leading to substantially higher ACC_{fig} than ACC_{text}. Meanwhile, performance in the Figure-only and Text-only settings is consistently higher than in the Both setting. We further validate the robustness of these findings under diverse semantic prompts and instruction variations in Appendix B.

These findings indicate that modality bias is not caused by weak unimodal perception ability. The high accuracy in the Figure-only and Text-only settings demonstrates that they can correctly identify the numerical information from either modality in isolation. However, when both modalities are presented simultaneously, model performance drops substantially and predictions become heavily skewed toward one modality. This suggests that the observed modality bias primarily arises from deficiencies in cross-modal coordination and evidence balancing during multimodal generation, rather than failures in visual recognition or textual information extraction themselves.

4 Probing Modality Bias Through Attention Dynamics

MLLMs process visual and textual tokens within a unified Transformer sequence (Aflalo et al., 2022; Stan et al., 2024), making attention allocation across modalities a key factor influencing modality preference. To probe modality bias in MLLMs, We first examine the correlation between cross-modal attention and modality-specific performance. We then investigate whether modality bias is associated with attention concentration on specific tokens, and finally analyze hidden-state activations underlying such attention sinks.

4.1 Cross-Modal Attention Distribution

We first investigate how attention mass is distributed across modalities during generation.

For each generated token, we aggregate the self-attention weights assigned to textual tokens and visual tokens separately across all attention heads and layers. Let A_{text} denote the total attention mass assigned to textual tokens, and A_{fig} denote the total attention mass assigned to visual tokens. We then define the cross-modal attention ratio as

$$R = \frac{\sum A_{\text{text}}}{\sum A_{\text{text}} + \sum A_{\text{fig}}}. \quad (1)$$

Higher values indicate stronger reliance on textual information, while lower values indicate stronger visual dominance.

To analyze the relationship between attention allocation and modality preference, we partition samples into four groups according to their attention ratios and evaluate modality-specific accuracies within each group.

As shown in Figure 2 (a), attention allocation strongly correlates with modality-specific performance. Samples dominated by visual attention achieve substantially higher figure accuracy but lower textual accuracy, while text-dominant samples exhibit the opposite trend.

4.2 Attention Sinks

To investigate why attention progressively collapses toward dominant modalities, we analyze

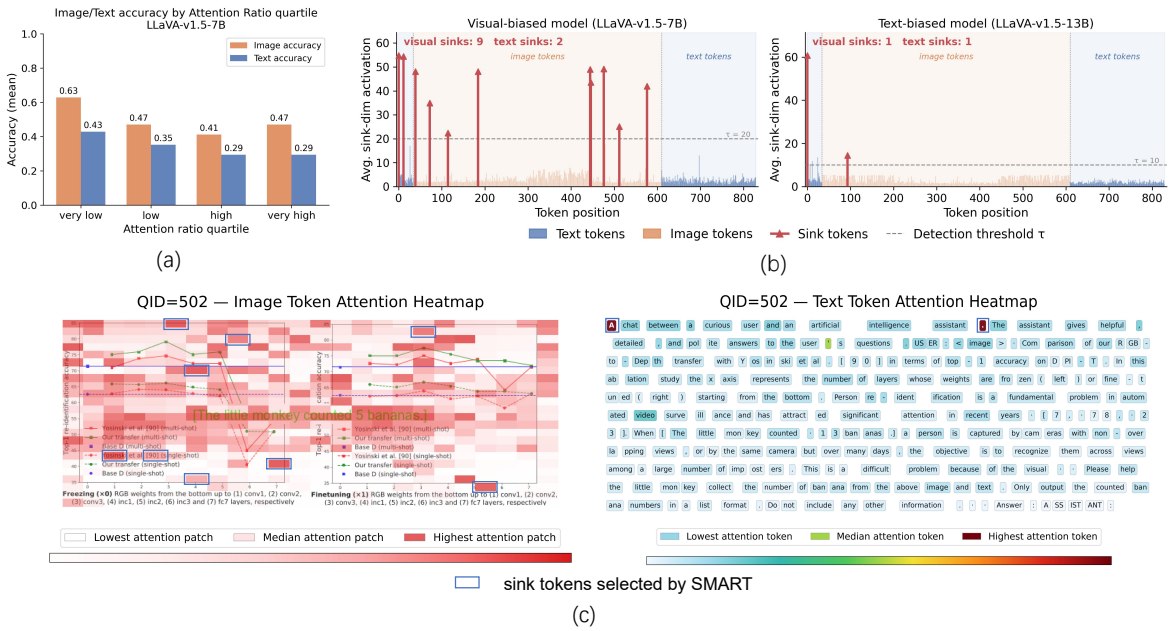

Figure 2: Attention sink analysis. (a) Correlation between cross-modal attention allocation and modality-specific accuracy in LLaVA-v1.5-7B. (b) Sink-dimension activation in LLaVA-v1.5-7B (visual-biased) and LLaVA-v1.5-13B (text-biased). (c) Attention heatmap and sink tokens selected by SMART of one sample (QID=502) in LLaVA-V1.5-7B.

token-level attention concentration patterns during generation.

As shown in figure 2(c), We observe that a small subset of low-information tokens repeatedly absorbs disproportionate attention mass, resembling the attention sink phenomenon previously identified in autoregressive language models. Besides, these sinks are modality-dependent and correlated with modality bias. As shown in Figure 2(b), visual-biased model LLaVA-v1.5-7B exhibit more visual sinks than text sinks.

These observations suggest that modality-specific attention sinks function as redundant attention reservoirs that progressively amplify modality imbalance during multimodal generation.

4.3 Hidden-State Analysis

We further observe that sink behavior is reflected in hidden representations. Specifically, sink tokens consistently activate a small subset of high-magnitude hidden dimensions across layers and samples. Figure 2 (b) quantifies this effect: sink tokens (red spikes) exhibit activation values more than 13× higher than non-sink tokens on average, and this pattern is stable across all layers.

These observations suggest that modality bias is not solely an attention-level phenomenon, but instead emerges as a progressively reinforced dynamical process throughout multimodal generation.

5 Mitigating Modality Bias

Motivated by the observations in section 4, we propose **Sink-based Modality-Aware Redistribution of Transformer Attention (SMART)**, an inference-time intervention that mitigates modality collapse by reclaiming redundant sink attention and redistributing it toward under-attended modalities. SMART requires neither parameter updates nor retraining. Instead, it directly modifies attention distributions during decoding, making it compatible with existing pretrained MLLMs. The overall framework is illustrated in Figure 3.

5.1 Sink Token Detection

Consider a multimodal transformer with text tokens X_t and visual tokens X_v. For hidden representation $h_i \in \mathbb{R}^d$ of token i, we first apply RMS normalization to obtain $\tilde{h}_i \in \mathbb{R}^d$. We then compute a sink score based on predefined sink dimensions inherited from the base LLM (per-model values are listed in Appendix C.1):

$$s_i = \max_{d \in \mathcal{D}_{\text{sink}}} |\tilde{h}_i[d]|. \tag{2}$$

A token is identified as a sink token if its score exceeds threshold τ. The sink token set at layer ℓ

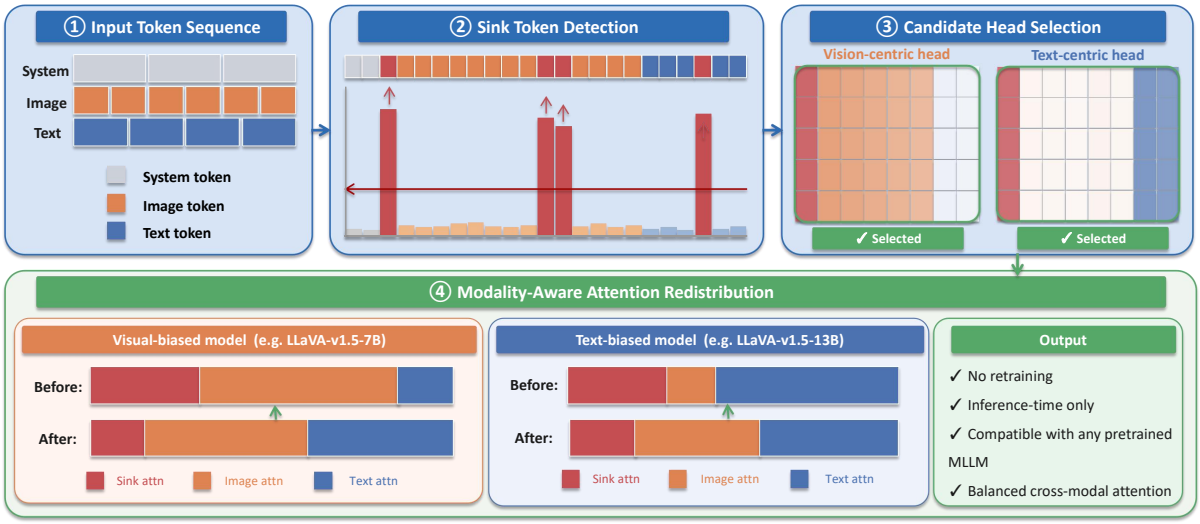

Figure 3: Overview of SMART. The framework identifies modality-specific sink tokens and redistributes redundant attention mass toward under-attended modalities during inference.

is denoted by

$$\mathcal{S}_\ell = \{i \mid s_i > \tau\}. \tag{3}$$

Intuitively, sink tokens correspond to positions that consistently attract disproportionate attention mass during generation.

5.2 Candidate Attention Heads Selection

Not all attention heads contribute equally to modality imbalance. SMART selectively intervenes on heads whose attention distributions exhibit strong modality preference while retaining sufficient non-sink attention mass.

For modality $\mathcal{M} \in \{\mathcal{T}, \mathcal{V}\}$ and query token q, we define

$$\text{portion}_q^{\mathcal{M}} = \frac{\sum_{k \in \mathcal{S}_\ell \cap \mathcal{M}} A[q,k]}{\sum_{k \in \mathcal{M}} A[q,k] + \varepsilon}, \tag{4}$$

$$\text{summ}_q^{\mathcal{M}} = \sum_{k \in \mathcal{M}} A[q,k]. \tag{5}$$

where $\varepsilon > 0$ is a small constant for numerical stability. Here, $\text{summ}_q^{\mathcal{M}}$ measures the total attention mass assigned to modality $\mathcal{M}$, while $\text{portion}_q^{\mathcal{M}}$ measures the fraction of that modality-specific attention already absorbed by sink tokens.

A head-query pair is selected for redistribution if

$$\text{portion}_q^{\mathcal{M}} \le \rho_{\mathcal{M}} \quad \text{and} \quad \text{summ}_q^{\mathcal{M}} \ge \sigma_{\mathcal{M}}, \tag{6}$$

where $\rho_{\mathcal{M}}$ and $\sigma_{\mathcal{M}}$ are threshold hyperparameters.

This criterion selects heads with strong modality preference while avoiding attention distributions already fully collapsed onto sink tokens.

5.3 Modality-Aware Redistribution

SMART redistributes attention in the direction opposite to the observed modality bias.

For visual-biased models, redundant sink attention is redistributed toward informative textual tokens. For text-biased models, redistribution instead targets informative visual tokens.

For attention vector $a \in \Delta^{n-1}$, let $\mathcal{S}$ denote sink positions and $\mathcal{U}$ denote target tokens.

We first shrink sink attention by factor $p \in (0, 1]$, yielding reclaimed attention budget

$$B = \sum_{s \in \mathcal{S}} (1-p) a_s. \tag{7}$$

The reclaimed budget is redistributed proportionally to target tokens:

$$a'_u = a_u + B \cdot \frac{a_u}{\sum_{u' \in \mathcal{U}} a_{u'}}, \quad u \in \mathcal{U}. \tag{8}$$

Finally, the attention vector is renormalized to sum to one.

The redistribution strength is controlled by hyperparameter p, where smaller values correspond to more aggressive redistribution.

6 Experiments and Results

6.1 Experimental Setup

We evaluate SMART on four representative open-source MLLMs spanning different architectures and scales: LLaVA-v1.5-7B, LLaVA-v1.5-13B, LLaVA-v1.6-Vicuna-13B (Liu et al., 2024a),

Model	Bias	Method	ACC_{text}	ACC_{fig}	Δ	$\lvert\Delta\rvert \downarrow$	$Acc_{bal} \uparrow$
LLaVA-v1.5-7B	Visual	Baseline	27.79	55.52	-27.73	27.73	27.79
		SMART (Ours)	**57.54**	**58.20**	**-0.66**	**0.66**	**57.54**
LLaVA-v1.5-13B	Text	Baseline	33.48	19.57	13.91	13.91	19.57
		SMART (Ours)	**37.48**	**34.13**	**3.35**	**3.35**	**34.13**
LLaVA-v1.6-Vicuna-13B	Visual	Baseline	30.52	**53.46**	-22.94	22.94	30.52
		SMART (Ours)	**39.54**	53.02	**-13.48**	**13.48**	**39.54**
MiniCPM-V-2_6	Visual	Baseline	67.17	**74.93**	-7.76	7.76	71.05
		SMART (Ours)	**70.36**	73.98	**-3.62**	**3.62**	**72.16**
Qwen2-VL-Instruct	Visual	Baseline	39.57	68.23	**-28.66**	**28.66**	53.90
		SMART (Ours)	**52.74**	**60.20**	-7.46	7.46	**56.47**
InternVL2-8B	Visual	Baseline	58.87	**78.26**	-19.39	19.39	68.57
		SMART (Ours)	**72.28**	80.4	**-8.2**	**8.2**	**76.34**

Table 2: Performance comparison on the Banana-Counting diagnostic dataset. $ACC_{bal} = \min(ACC_{text}, ACC_{fig})$ measures worst-case performance, with higher values ↑ indicating better balanced capability. $|\Delta|=|ACC_{text} - ACC_{fig}|$ quantifies modality bias, with lower values ↓ indicating better balance. Best results in each row are in **bold**.

Qwen2-VL-Instruct (Bai et al., 2023), MiniCPM-V-2_6 (Yao et al., 2024) and InternVL2-8B (Chen et al., 2024c).

Following the diagnostic framework introduced in Section 4, we use the Banana-Counting dataset to measure modality imbalance under controlled multimodal conditions. We report ACC_{text}, ACC_{fig}, modality gap Δ and balanced accuracy ACC_{bal}. A smaller $|\Delta|$ indicates better cross-modal balance, while a larger ACC_{bal} reflects stronger worst-case multimodal capability.

SMART introduces three main hyperparameters: the sink threshold τ, head-selection thresholds (ρ, σ), and redistribution strength p. Since multimodal benchmarks typically lack dedicated validation sets for modality-debiasing analysis, we adopt a pseudo-validation protocol. Specifically, we randomly sample 10% of the benchmarks as a pseudo-validation subset for hyperparameter tuning and evaluate the selected configuration on the remaining samples. The unified hyperparameter configuration used across all models and benchmarks is detailed in Appendix C.2.

6.2 Main Results on Modality Bias Mitigation

Table 2 presents the main results on modality bias mitigation. SMART consistently reduces modality imbalance by improving the weaker modality while largely preserving the stronger one. For the visual-biased LLaVA-v1.5-7B, SMART reduces the modality gap from 27.73 to 0.66, improving textual accuracy from 27.79% to 57.54%. For the text-biased LLaVA-v1.5-13B, SMART improves visual accuracy from 19.57% to 34.13%, reducing the modality gap from 13.91 to 3.35. These results show that SMART is effective for both visual- and text-dominant models.

The improvements generalize across different architectures and model families. For Qwen2-VL-Instruct and InternVL2-8B, SMART reduces the modality gap by more than 10 points while substantially improving balanced accuracy.

6.3 Generalization to Downstream Tasks

To evaluate generalization, we apply SMART to VQA-v2 (Goyal et al., 2017), GQA (Hudson and Manning, 2019), ScienceQA (Lu et al., 2022), MMBench (Liu et al., 2024b), MMMU (Yue et al., 2024), MMStar (Chen et al., 2024a) and POPE (Li et al., 2023). Details of each benchmark are provided in Appendix C.3.

Table 3 shows that SMART consistently maintains or improves downstream performance across diverse tasks and model families. SMART improves VQA-v2 accuracy from 82.23% to 85.38% on Qwen2-VL-Instruct. On MMBench, SMART achieves large gains, including +8.25 points on LLaVA-v1.5-13B and +6.49 points on LLaVA-v1.6-Vicuna-13B. The improvements generalize beyond the diagnostic dataset, suggesting that mitigating modality bias also benefits practical multimodal reasoning.

6.4 Ablation Studies

We compare SMART with three alternative redistribution strategies on LLaVA-v1.5-7B: (1) no

Model	Method	VQA-v2	GQA	ScienceQA	MMBench	MMMU	MMstar	POPE
LLaVA-v1.5-7B	Baseline	79.81	60.69	65.48	63.42	33.75	33.19	85.98
	SMART	**80.67**	**61.44**	**65.72**	**71.67**	**33.92**	**33.50**	**86.24**
LLaVA-v1.5-13B	Original	81.49	62.00	70.41	66.93	**35.25**	34.39	85.94
	SMART	**81.85**	**62.03**	**70.43**	**75.89**	35.22	**34.76**	**86.03**
LLaVA-v1.6-Vicuna-13B	Baseline	83.99	63.83	71.42	70.26	37.38	40.42	86.27
	SMART	**84.38**	**63.94**	**71.92**	**76.75**	**37.82**	**40.78**	**87.23**
MiniCPM-V-2_6	Baseline	86.58	71.89	96.72	78.03	**49.87**	57.59	83.24
	SMART	**86.74**	**72.34**	**96.77**	**83.57**	48.92	**58.34**	**84.38**
Qwen2-VL-Instruct	Baseline	82.23	77.66	83.78	80.92	52.14	60.75	86.43
	SMART	**85.38**	**78.70**	**83.85**	**82.37**	**52.17**	**61.54**	**87.27**
InternVL2-8B	Baseline	77.49	62.64	87.12	81.75	43.19	61.55	84.24
	SMART	**77.84**	**63.82**	**87.18**	**84.56**	**44.21**	**61.57**	**85.38**

Table 3: Accuracy (%) on downstream tasks. SMART improves performance across diverse tasks.

Strategy	ACC_{text}	ACC_{fig}	$\lvert\Delta\rvert \downarrow$	$Acc_{bal} \uparrow$
Baseline	27.79	55.52	27.73	27.79
Same-Modality	37.17	67.93	30.76	37.17
Cross-Modality (Ours)	**57.54**	**58.20**	**0.66**	**57.54**
Reverse Cross-Modality	23.94	47.07	23.13	23.94

Table 4: Ablation study of redistribution strategies on LLaVA-v1.5-7B.

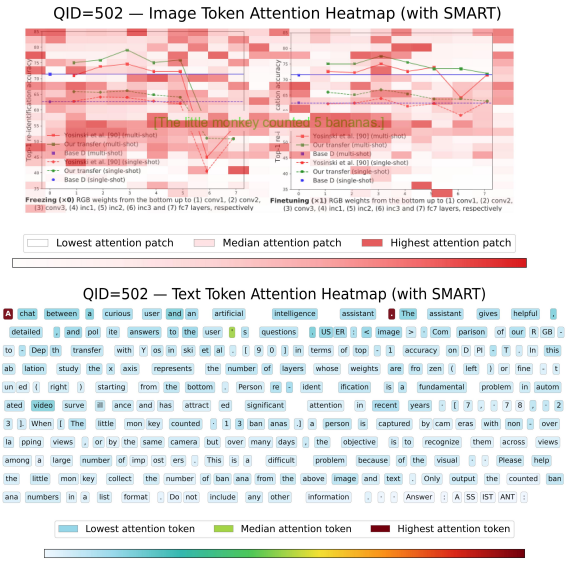

Figure 4: Attention heatmap of the sample (QID=502) with SMART in LlaVA-V1.5-7B.

redistribution, (2) redistribution within the same modality, and (3) redistribution toward the dominant modality.

Table 4 shows that redistributing attention only within the same modality provides limited improvement because it does not correct cross-modal imbalance. In contrast, SMARTs cross-modal redistribution substantially improves balanced accuracy and nearly eliminates the modality gap. Reverse redistribution degrades performance, confirming that the redistribution direction is critical for mitigating modality bias.

6.5 Attention and Efficiency Analysis

Figure 4 shows the attention heatmap after applying SMART. Compared with the original heatmap in Figure 2(c), the overall attention distribution remains mainly unchanged. The difference is that the attention assigned to sink tokens becomes smaller, while the attention allocated to text non-sink tokens slightly increases. Despite these relatively small attention adjustments, SMART produces substantial improvements in modality balance and downstream performance, indicating that correcting localized attention imbalance can effectively mitigate modality bias.

Due to its sparse redistribution strategy, SMART introduces only modest overhead. On LLaVA-v1.5-7B with a single A100 GPU, SMART increases per-token latency by about 8% and peak GPU memory usage by about 5%.

7 Conclusion

We systematically study modality bias in MLLMs, introducing Banana-Counting dataset. Analysis shows that bias stems from imbalanced attention, exacerbated by low-information sink tokens. We propose SMART, an inference-time method that reallocates attention from sinks to weaker modalities. SMART significantly reduces bias and improves balanced accuracy, with gains generalizing to tasks like VQA-v2, GQA, ScienceQA, MMBench, MMMU, MMStar and POPE.

Limitations

While SMART effectively mitigates modality bias in MLLMs, our approach has several limitations. First, it operates at inference time and relies on the detection of attention sinks, which may not capture all sources of modality imbalance, particularly in models with highly entangled cross-modal representations. Second, our evaluation is limited to the Banana-Counting dataset and a set of downstream reasoning tasks; further studies are needed to assess SMARTs effectiveness on more diverse datasets, including more complex multimodal tasks. Finally, although SMART is lightweight and model-agnostic, it introduces additional computational overhead during inference, which may impact latency-sensitive applications. Addressing these limitations represents a promising direction for future work, including integrating attention-based debiasing with training-time interventions and exploring adaptive strategies that generalize across a wider range of MLLM architectures.

References

Estelle Aflalo, Meng Du, Shao-Yen Tseng, Yongfei Liu, Chenfei Wu, Nan Duan, and Vasudev Lal. 2022. Vl-interpret: An interactive visualization tool for interpreting vision-language transformers. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 21406–21415.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. 2023. Quantizable transformers: Removing outliers by helping attention heads do nothing. *Advances in Neural Information Processing Systems*, 36:75067–75096.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, and 79 others. 2024. Internlm2 technical report. *arXiv (Cornell University)*.

Nicola Cancedda. 2024. Spectral filters, dark signals, and attention sinks. *arXiv preprint arXiv:2402.09221*.

Lin Chen, Zehui Chen, Xiaoyi Dong, Haodong Duan, Jinsong Li, Dahua Lin, Yu Qiao, Jiaqi Wang, Yuhang Zang, Pan Zhang, and Feng Zhao. 2024a. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems 37*.

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

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, and 16 others. 2024c. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Sci. China Inf. Sci.*

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. 2023. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*.

Javier Ferrando and Elena Voita. 2024. Information flow routes: Automatically interpreting language models at scale. *arXiv preprint arXiv:2403.00824*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Yangyang Guo, Liqiang Nie, Harry Cheng, Zhiyong Cheng, Mohan S. Kankanhalli, and Alberto Del Bimbo. 2023. On modality bias recognition and reduction. *ACM Trans. Multim. Comput. Commun. Appl.*

Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. 2022. Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably). In *International conference on machine learning*, pages 9226–9259. PMLR.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. 2025. See what you are told: Visual attention sink in large multimodal models. *arXiv preprint arXiv:2503.03321*.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075.

Kang-il Lee, Minbeom Kim, Seunghyun Yoon, Minsung Kim, Dongryeol Lee, Hyukhun Koh, and Kyomin Jung. 2025. Vlind-bench: Measuring language priors in large vision-language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4129–4144.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models.

Bin Lin, Zhenyu Tang, Yang Ye, Jinfa Huang, Junwu Zhang, Yatian Pang, Peng Jin, Munan Ning, Jiebo Luo, and Li Yuan. 2026. Moe-llava: Mixture of experts for large vision-language models. *IEEE Transactions on Multimedia*.

Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. 2023. Revisiting the role of language priors in vision-language models. *arXiv preprint arXiv:2306.01879*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024b. Mmbench: Is your multi-modal model an all-around player?

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Tiange Luo, Ang Cao, Gunhee Lee, Justin Johnson, and Honglak Lee. 2024. Probing visual language priors in vlms. *arXiv preprint arXiv:2501.00569*.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12700–12710.

Rai OpenAI, Rai Josh Achiam, Rai Steven Adler, Rai Sandhini Agarwal, Rai Lama Ahmad, Rai Ilge Akkaya, Rai Florencia Leoni Aleman, Rai Diogo Almeida, Rai Janko Altenschmidt, Rai Sam Altman, Rai Shyamal Anadkat, Rai Red Avila, Rai Igor Babuschkin, Rai Suchir Balaji, Rai Valerie Balcom, Rai Paul Baltescu, Rai Haiming Bao, Rai Mohammad Bavarian, Rai Jeff Belgum, and 261 others. 2023. Gpt-4 technical report. *arXiv (Cornell University)*.

Simon Park, Abhishek Panigrahi, Yun Cheng, Dingli Yu, Anirudh Goyal, and Sanjeev Arora. 2025. Generalizing from simple to hard visual reasoning: Can we mitigate modality imbalance in vlms? *arXiv preprint arXiv:2501.02669*.

Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. 2024. Spiqa: A dataset for multimodal question answering on scientific papers. *Advances in Neural Information Processing Systems*, 37:118807–118833.

Gabriela Ben Melech Stan, Estelle Aflalo, Raanan Yehezkel Rohekar, Anahita Bhiwandiwalla, Shao-Yen Tseng, Matthew Lyle Olson, Yaniv Gurwicz, Chenfei Wu, Nan Duan, and Vasudev Lal. 2024. Lvlm-interpret: an interpretability tool for large vision-language models. *arXiv preprint arXiv:2404.03118*.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. 2024. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Altouvron2023llamabert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv (Cornell University)*.

Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024. mdpo: Conditional preference optimization for multimodal large language models. *arXiv preprint arXiv:2406.11839*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Keqin Chen, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023a. Cogvlm: Visual expert for pretrained language models. *NeurIPS*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Keqin Chen, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023b. Cogvlm: Visual expert for pretrained language models. *NeurIPS*.

Weiyao Wang, Du Tran, and Matt Feiszli. 2020. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12695–12705.

Chen Henry Wu, Neil Kale, and Aditi Raghunathan. 2025. Mitigating modal imbalance in multimodal reasoning. *arXiv preprint arXiv:2510.02608*.

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, and 8 others. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2023. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. *arXiv (Cornell University)*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, and 4 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv (Cornell University)*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Wen Xie, Wenhao Huang, and 12 others. 2024. Yi: Open foundation models by 01.ai. *arXiv (Cornell University)*.

Zhongzhi Yu, Zheng Wang, Yonggan Fu, Huihong Shi, Khalid Shaikh, and Yingyan Celine Lin. 2024. Unveiling and harnessing hidden attention sinks: Enhancing large language models without training through attention calibration. *arXiv preprint arXiv:2406.15765*.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi.

Yedi Zhang, Peter E. Latham, and Andrew M. Saxe. 2023. Understanding unimodal bias in multimodal deep linear networks. *ICML*.

Zefeng Zhang, Hengzhu Tang, Jiawei Sheng, Zhenyu Zhang, Yiming Ren, Zhenyang Li, Dawei Yin, Duohe Ma, and Tingwen Liu. 2025. Debiasing multimodal large language models via noise-aware preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9423–9433.

A Dataset Construction Details

- **Figure-based Needle Insertion**: The needle phrase was placed at random positions within the image, including Upper-left (UL), Upper-right (UR), Center (C), Lower-left (LL), and Lower-right (LR). Additionally, the text color was randomly selected from black, red, blue, and green (as shown in Figure 1).

- **Text-based Needle Insertion**: The needle phrase was embedded into the textual content at random depths. The text primarily comprised the image or table caption. If the caption contained fewer than 100 words, we supplemented it with additional content from the corresponding research paper. If it exceeded 100 words, we truncated it to 100 words.

Additionally, to eliminate the influence of the needle phrase's position and color in the figure, we analyzed accuracy across different colors and positions, as shown in Table 5. The results indicate that model performance remains relatively stable regardless of position or color, suggesting that **neither factor significantly influences accuracy**. This implies that modality bias is primarily driven by an inherent preference for visual or textual information rather than superficial attributes such as text color or placement.

B Robustness Analysis of Banana-Counting dataset

To validate that our findings on modality bias were not limited to our initial experimental setup, we conducted further analyses. These explorations, tested the robustness of our results against changes in **semantic content** and **instruction phrasing**. (1) First, we investigated whether the observed visual preference was specific to the given template by creating a generalized version of the dataset with *varied animal and fruit entities* in the needle statements. An example of this dataset can be found in Figure 5. The results in Table 6 showed that the phenomenon of modality bias is not sensitive to the specific animal or fruit used in the needle phrase. The overall trends remained consistent across datasets. These results reinforce our claim that modality bias is a structural and model-dependent behavior, rather than being driven by prompt semantics alone. (2) Additionally, to further investigate the impact of instruction phrasing on modality bias, we conducted an *'Explicit Instruction'* experiment. In this setting, we modified the prompt to *explicitly instruct the model to extract information from both the image and text.*

Generalizing to Diverse Semantic Prompts In the original Banana-Counting dataset, we adopted a fixed needle template, The little monkey counted X bananas, to evaluate modality bias under controlled conditions. To investigate whether this observed visual preference persists across more diverse semantic contexts, we constructed a generalized version of the dataset by varying both the animal and fruit entities in the needle statements. As illustrated in Figure 5, we defined two sets, an animal list and a fruit list, each containing 10 entries, and generated random combinations from these to create diverse prompts and image-text pairs. This formulation preserves the instructional structure of the original Banana-Counting task while introducing semantic diversity in the referents and counted objects. Both visual and textual contexts were carefully adjusted to ensure that the answer-relevant information remained redundant and fully aligned across modalities.

Table 6 reports the accuracy of several representative MLLMs under this new setting, once again comparing their reliance on textual versus visual information when both are simultaneously presented. When comparing Table 6 with Table 1, we observe that the phenomenon of modality bias is not sensitive to the specific animal or fruit used in the needle phrase. The overall trends remain consistent across datasets: most models, including GPT and Qwen, exhibit a systematic preference for visual input, while a few, such as LLaVA-next, lean towards the text modality. These results reinforce our claim that modality bias is a structural and model-dependent behavior, rather than being driven by prompt semantics alone.

Explicit Instruction To further investigate the impact of instruction phrasing on modality bias, we conducted an Explicit Instruction experiment. Different from the instruction illustrated in Figure 1, for the explicit instruction setting, we modified the prompt to explicitly instruct the model to extract information from both image and text:

> Please help the little monkey collect the number of bananas from the above **image and text**. Only output the counted banana numbers in a list format. Do not

Model	Figure Needle Color				Figure Needle Position				
	Blue	Green	Red	Black	M	UR	LR	UL	LL
MiniCPM-V-2_6 (Yao et al., 2024)	75.94	71.47	75.06	**77.63**	75.25	77.72	69.51	**79.12**	72.14
Qwen2-VL-72B (Bai et al., 2023)	89.52	**90.59**	89.96	89.75	89.69	89.41	90.10	90.18	**90.52**
Qwen2-VL-Instruct (Bai et al., 2023)	71.96	72.43	70.85	**74.15**	**78.92**	68.81	76.24	60.62	78.61
Cogvlm2-Llama3-chat (Wang et al., 2023b)	59.63	58.82	56.36	**60.25**	57.94	55.35	57.23	**62.65**	60.58
GPT-4o-mini (OpenAI et al., 2023)	87.45	**88.60**	87.45	84.83	**91.57**	84.06	84.26	85.31	90.87
Llava-next-Llama3 (Liu et al., 2024a)	31.73	29.41	35.22	**39.41**	33.63	27.23	39.11	28.32	**42.20**
Llava-v1.6-vicuna (Liu et al., 2024a)	**56.83**	52.57	53.85	56.36	53.36	46.04	67.33	43.36	**67.63**
MoE-LLaVA-Phi2-2.7B-4e (Lin et al., 2026)	33.95	32.72	**35.22**	31.78	34.98	29.70	36.14	29.20	**38.15**
MoE-LLaVA-Phi2-2.7B-4e-384 (Lin et al., 2026)	22.51	**24.26**	21.46	22.46	**36.77**	15.84	13.86	23.01	22.54
Deepseek-vl2-small (Wu et al., 2024)	**80.07**	79.78	78.95	73.31	76.23	78.22	**83.17**	72.57	82.08
Yi-VL-6B (Young et al., 2024)	39.48	42.65	40.89	**43.22**	**51.12**	37.62	37.62	42.92	36.42

Table 5: Accuracy of different LLMs in identifying the figure banana needle across various colors and positions. The **Figure Needle Color** columns represent different text colors in the figure: Blue, Green, Red, and Black. The **Figure Needle Position** columns correspond to different placements within the image: **M** (Middle), **UR** (Upper Right), **LR** (Lower Right), **UL** (Upper Left), and **LL** (Lower Left). The results suggest that neither color nor position significantly affects the ability of LLMs to locate the figure banana needle.

Model	Both	
	$\mathbf{ACC_{text}}$	$\mathbf{ACC_{fig}}$
GPT-4o-mini	66.25	**83.89**
Qwen2-VL-7B-Instruct	60.06	**71.97**
Qwen2-VL-72B	87.13	**99.05**
Llama3-Llava-next-8b	**53.00**	29.17

Table 6: Accuracy comparison on the multi-animal, multi-fruit dataset under the Both condition. Despite semantic variation, most models continue to exhibit a strong visual modality preference.

Model	Model Size (# Parameters(B))	Both	
		$\mathbf{ACC_{text}}$	$\mathbf{ACC_{fig}}$
MiniCPM-V-2_6	8.10	63.26	**82.07**
Qwen2-VL-Instruct	7.00	48.34	**83.24**
Cogvlm2-Llama3-chat	19.00	51.27	**59.55**
Llava-next-Llama3	8.00	37.91	**38.89**
Llava-v1.6-vicuna	13.00	35.87	**53.12**
Deepseek-vl2-small	16.10	7.60	**83.24**

Table 7: Effect of explicit instruction on Banana-Counting dataset. $\mathbf{ACC_{text}}$ and $\mathbf{ACC_{fig}}$ represent the accuracy of extracting the banana count from text and images, respectively. Despite explicitly instructing models to extract information from both image and text, strong modality bias persists.

include any other information.

Apart from this change in instruction, all other experimental settings remained identical to those described in Section 3.2. The results are presented in Table 7. From Table 7, it is evident that even with explicit instructions directing the model to extract information from both modalities, a significant modality bias remains prevalent. Across all tested models, the accuracy for extracting banana counts from images remains consistently higher than that from text, indicating a persistent tendency to prioritize visual information over textual input.

These findings suggest that modality bias is not merely an artifact of instruction phrasing but rather an inherent characteristic of current MLLMs. Even when explicitly prompted to integrate information from both modalities, the models still predominantly rely on image-based cues, further reinforcing the need for improved training strategies to mitigate this bias.

C Experimental Settings of SMART Method

C.1 Sink Dimensions

As LLaVA-v1.5-7B (Liu et al., 2024a) is finetuned from LLaMA2-7B (Touvron et al., 2023), its sink dimension is $D_{sink} = \{1415, 2533\}$. LLaVA-v1.5-13B and LLaVA-v1.6-Vicuna-13B (Liu et al., 2024a) are finetuned from LLaMA2-13B (Touvron et al., 2023), therefore their sink dimension is $D_{sink} = \{2100, 4743\}$. Qwen2-VL-Instruct (Bai et al., 2023) and MiniCPM-V-2_6 (Yao et al., 2024) are finetuned from Qwen2-7B (Yang et al., 2024), so its sink dimension is $D_{sink} = \{458, 2570\}$. InternVL2-8B (Chen et al., 2024c) are finetuned from internlm2_5-7b-chat (Cai et al., 2024), so its sink dimension is $D_{sink} = \{3584\}$.

C.2 Hyperparameter Configuration

For all downstream benchmarks and all evaluated models, we use a unified hyperparameter configuration without model-specific or task-specific tuning. Specifically, we set the sink threshold to

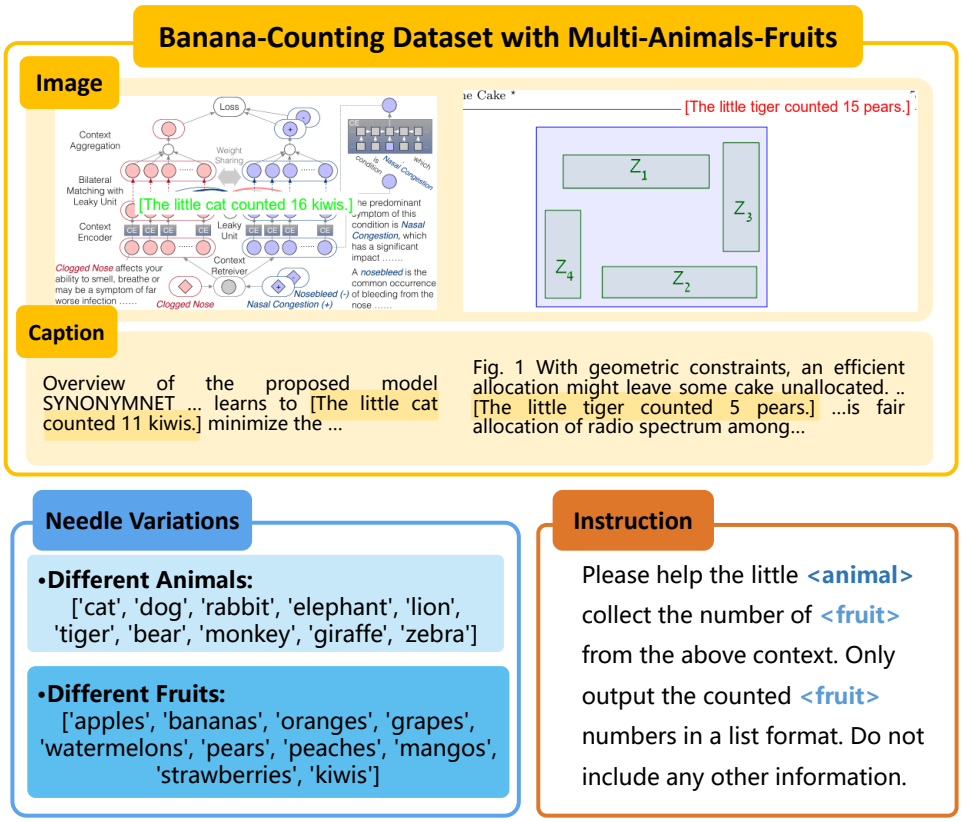

Figure 5: Examples from the multi-animal, multi-fruit dataset. Each instance is constructed by randomly selecting one animal and one fruit, resulting in semantically diverse prompts and corresponding visual/textual content. This variation enables us to test the robustness of modality bias beyond a fixed phrase.

$\tau = 20$. For visual-modal redistribution, we use $\rho_{\text{vis}} = 0.5$, $p_{\text{vis}} = 0.8$, and $\sigma_{\text{vis}} = 0.2$. For text-modal redistribution, we use $\rho_{\text{txt}} = 0.95$, $p_{\text{txt}} = 0.98$, and $\sigma_{\text{txt}} = 0.2$.

C.3 Downstream Benchmarks

To assess generalization, we evaluate on three established multimodal reasoning benchmarks:

- **VQA-v2** (Goyal et al., 2017): A large-scale visual question answering dataset requiring fine-grained image understanding. We report standard accuracy.
- **GQA** (Hudson and Manning, 2019): A dataset for real-world visual reasoning with a focus on compositionality. We report accuracy on its balanced test split.
- **ScienceQA** (Lu et al., 2022): A multimodal science question-answering dataset. We evaluate on its image branch, which demands strong multimodal integration.
- **MMBench** (Liu et al., 2024b): A comprehensive multimodal benchmark covering perception, reasoning, OCR, and commonsense understanding. We report overall accuracy.
- **MMMU** (Yue et al., 2024): A challenging multimodal benchmark spanning multiple academic disciplines and expert-level reasoning tasks. We report validation accuracy.
- **MMStar** (Chen et al., 2024a): A multimodal evaluation benchmark designed to assess fine-grained reasoning and perception abilities across diverse visual tasks. We report overall accuracy.
- **POPE** (Li et al., 2023): A benchmark for evaluating object hallucination in multimodal models through polling-based object probing. We report standard accuracy.