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

Anonymous ACL submission

Abstract

Multimodal Large Language Models (MLLMs) integrate visual and textual information, yet often exhibit modality bias, where predictions over-rely on one modality while underutilizing the other. Through analysis, we find that modality bias in MLLMs arises from imbalanced Transformer attention distribution. The dominant modality tends to receive disproportionately high attention. Moreover, low-information sink tokens absorb redundant attention that could otherwise be allocated to the under-attended modality. Motivated by this, we propose **SMART (Sink-based Modality-Aware Redistribution of Transformer Attention)**, an inference-time method that detects modality-specific attention sinks and redistributes excessive attention to the under-attended modality. To better quantitatively assess modality bias, we construct **Banana-Counting**, a diagnostic dataset of 1,026 instances with mirrored information across visual and textual modalities. Our evaluation across ten MLLMs reveals severe modality bias, with some models exhibiting over 20-point accuracy gaps between visual and textual data. SMART effectively reduces the modality bias gap from 27.73 to 0.66 and improves balanced accuracy by up to 29.75%. Moreover, these gains consistently generalize to downstream tasks including VQA-v2, GQA, and ScienceQA, indicating that mitigating modality bias improves both robustness and generalization.

1 Introduction

The rapid evolution of Large Language Models (LLMs) has been a major driving force in the pursuit of Artificial General Intelligence (AGI). To meet the increasing demands of real-world applications, LLMs have advanced beyond single-modality processing into Multimodal LLMs (MLLMs). By integrating LLMs with visual encoders (Radford et al., 2021), MLLMs have demonstrated remarkable capabilities across a wide range of vision–language tasks, including visual question answering, image captioning, and multimodal reasoning (Liu et al., 2024; Achiam et al., 2023; Wang et al., 2024b).

Despite their impressive performance, recent studies have revealed that MLLMs often exhibit *modality bias*, which refers to an over-reliance on one modality while neglecting information from the others (Chen et al., 2024; Guo et al., 2023). Such bias can impair multimodal reasoning and make models vulnerable to irrelevant or misleading inputs, particularly when one modality is noisy or paired data are scarce (Park et al., 2025; Wu et al., 2025; Zhang et al., 2023). Consequently, MLLMs may fail to effectively integrate complementary information across modalities, reducing reliability in real-world scenarios that require balanced multimodal understanding.

To address these challenges, we first introduce **Banana-Counting**, a controlled diagnostic benchmark designed to systematically identify and quantify modality bias in MLLMs. As far as we know, our dataset is the first to feature a balanced multimodal design. Evaluation on Banana-Counting shows that most modern MLLMs achieve highly asymmetric accuracy across modalities, with large accuracy gaps between textual and visual data.

Modern MLLMs rely on Transformer-based unified self-attention, where visual and textual tokens compete for attention during generation (Aflalo et al., 2022; Stan et al., 2024; Vaswani et al., 2017). Through analysis on the Banana-Counting benchmark, we observe that this competition is often skewed, with a dominant modality consistently receiving excessive attention, leading to modality bias. Meanwhile, recent studies have identified the phenomenon of attention sinks, in which low-information tokens, such as special tokens or background patches, absorb a disproportionate amount of attention (Xiao et al., 2023; Fer-

rando and Voita, 2024).

Motivated by these observations, we propose **SMART** (**S**ink-based **M**odality-**A**ware **R**edistribution of **T**ransformer Attention), an inference-time method that mitigates modality bias by identifying bias-inducing attention heads and their associated sinks, and redistributing excessive attention mass to information-rich regions in the under-attended modality. Through comprehensive empirical analysis, we demonstrate that our method **SMART** successfully reduces modality bias while simultaneously enhancing model robustness and maintaining or improving performance on general reasoning benchmarks. By addressing modality bias, we provide insights into how MLLMs can be improved for more reliable multimodal reasoning.

The main contributions of this work are summarized as follows: (1) We present the first systematic study that identifies and quantifies modality bias in MLLMs, revealing its prevalence and impact on multimodal reasoning. (2) We introduce **Banana-Counting** dataset, a novel and reproducible diagnostic benchmark specifically designed to quantify modality preference in a controlled manner. (3) We uncover a strong connection between modality bias and imbalanced cross-modal attention allocation, and demonstrate that attention sinks provide exploitable attention redundancy for bias mitigation. (4) We propose **SMART**, an inference-time method that mitigates modality bias by redistributing attention from bias-inducing sinks to under-attended modalities, without modifying model parameters. (5) We provide a comprehensive empirical analysis showing that SMART consistently reduces modality bias and improves robustness, with gains that generalize to diverse downstream multimodal reasoning tasks.

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., 2024). Prior studies have extensively explored modality bias in VQA tasks (Niu et al., 2021; Chen et al., 2024; 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., 2024a; 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 multimodal LLMs.

3 Quantifying modality bias

3.1 Banana-Counting Dataset Construction

Banana-Counting dataset is based on SPIQA (Scientific Paper Image Question Answering) (Pramanick et al., 2024), a large-scale QA dataset designed to interpret complex figures and tables within the context of scientific research articles across various domains of computer science. We constructed our dataset using the test-A split from SPIQA, selecting images, their corresponding cap-

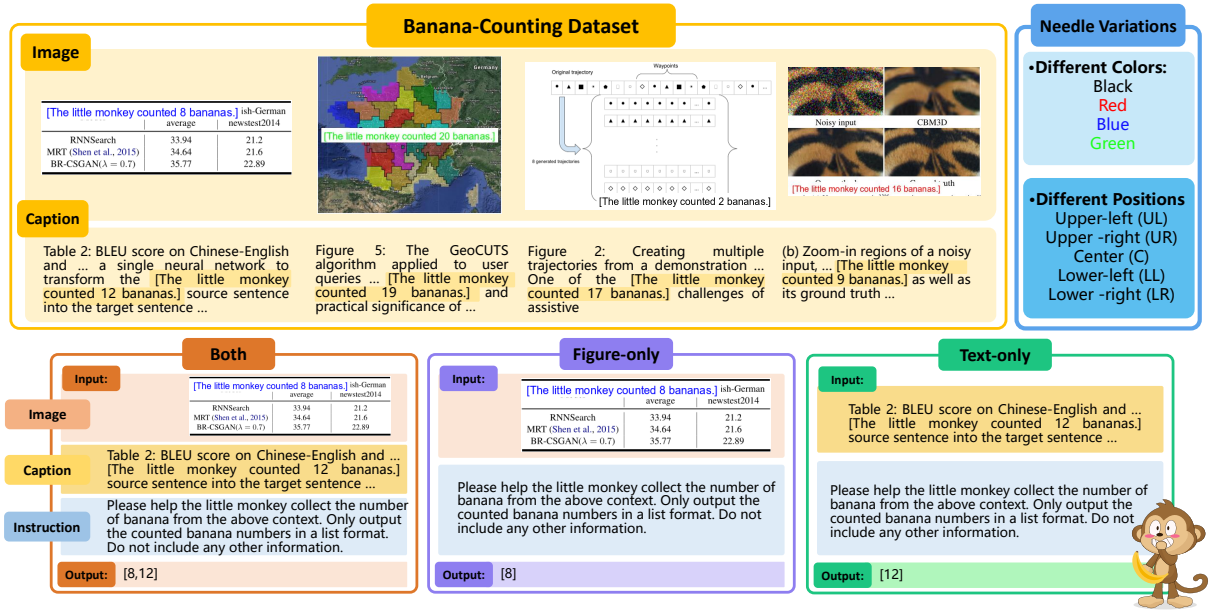

Figure 1: Overview of the Banana-Counting Dataset. We extracted figures and captions from the SPIQA dataset and inserted needle phrases into both modalities. The needles color and position in figures were randomly assigned while in text, it was inserted at random positions to avoid biases. To evaluate 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.

tions, and associated text. We then inserted a needle phrase into both text and images in the format:

The little monkey counted {number} bananas.

where the number of bananas in each needle was randomly generated within the range of 1 to 20. The Banana-Counting dataset contains a total of 1026 instances for evaluation. The overview of the dataset is illustrated in Figure 1. To eliminate potential biases caused by superficial cues, the **needles color and position in the figure were randomly assigned**, while in the text, the needle was inserted at random positions. Specifically:

- **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.

3.2 Experiments

During the inference phase, we provided input with predefined instructions, as shown in Figure 1, to guide MLLMs in extracting the banana count from the context. We use the above constructed Banana-Counting Dataset to evaluate whether LLMs could **identify and extract the banana count from the given context**. In each instance, different needle phrases are embedded in both the image and the text. This setup is used to test models under three conditions:

1. Both: The model receives both the figure and caption as input;
2. Figure-only: The model receives only the figure as input;
3. Text-only: The model receives only the caption as input.

In the Both setting, we aim to measure different MLLMs' preferences for specific modalities. In the single-modality settings (Figure-only and Text-only), we further investigated the accuracy of MLLMs when restricted to a single

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., 2023)	19.00	**67.82**	58.77	96.80	91.29
GPT-4o-mini (Achiam et al., 2023)	-	81.52	**87.15**	99.51	100.00
Llava-v1.5-7B (Liu et al., 2024)	7.00	27.79	**55.52**	58.87	83.82
Llava-v1.5-13B (Liu et al., 2024)	13.00	**33.48**	19.57	38.01	81.60
Llava-v1.6-vicuna-13B (Liu et al., 2024)	13.00	30.52	**53.46**	68.13	95.32
Llava-next-Llama3 (Liu et al., 2024)	8.00	**47.95**	33.72	42.50	93.86
MoE-LLaVA-Phi2-2.7B-4e (Lin et al., 2024)	5.61	22.71	**96.20**	52.05	88.11
MoE-LLaVA-Phi2-2.7B-4e-384 (Lin et al., 2024)	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
NVILA-8B (Liu et al., 2025)	8.00	94.93	**95.07**	95.32	96.47

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 ($\mathbf{ACC_{text}}$).

modality. This evaluation serves to demonstrate that MLLMs possess sufficient capability to process each modality independently, ensuring that modality bias is not merely a result of inadequate unimodal processing ability, but rather an inherent preference for one modality over the other. We tested ten different MLLMs on the Banana-Counting dataset including, MiniCPM-V-2_6 (Yao et al., 2024), Qwen2-VL-7B-Instruct, Qwen2-VL-72B (Bai et al., 2023), Cogvlm2-Llama3-chat-19B (Wang et al., 2023), GPT-4o-mini (Achiam et al., 2023), Llava-v1.5-7B, Llava-v1.5-13B, Llava-v1.6-vicuna-13B (Liu et al., 2024), MoE-LLaVA-Phi2-2.7B-4e, MoE-LLaVA-Phi2-2.7B-4e-384 (Lin et al., 2024), Deepseek-vl2-small (Wu et al., 2024), YiVL-6B (Young et al., 2024) and NVILA-8B (Liu et al., 2025).

3.3 Results

The experimental results are presented in Table 1. The results reveal clear modality bias across most MLLMs: (1) When both modalities are available, most models prioritize visual information, extracting the banana count primarily from the figure while overlooking the textual banana count. This results in significantly higher $\mathbf{ACC_{fig}}$ values compared to $\mathbf{ACC_{text}}$. (2) In the Figure-only and Text-only settings, all models demonstrate higher accuracy in extracting banana counts compared to the Both setting. This indicates that when information must be combined across modalities, models struggle to effectively integrate textual and visual cues. (3) Among the models exhibiting an inverse modality bias, Cogvlm, Yi, Llava-v1.5-13B and Llava-next-Llama3 show distinct trends. Cogvlm displays a stronger preference for textual information, leading to lower $\mathrm{ACC_{fig}}$ compared to other models. Meanwhile, Yi and Llava-next-Llama3 demonstrates extremely low accuracy in the Figure-only setting, suggesting that it has poor image text recognition capabilities, causing it to ignore visual information and rely more heavily on text. (4) NVILA-15B stands out as nearly balanced, with both accuracies approaching 95%, suggesting effective multimodal integration in its base configuration.

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 2. 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.

3.4 Attention Analysis

The results in Section 3.3 reveal consistent modality bias across a wide range of MLLMs. To understand its underlying cause, we analyze models' internal attention patterns during generation.

In modern MLLMs, visual inputs are encoded as visual tokens and concatenated with textual tokens into a unified sequence processed by a Trans-

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., 2023)	59.63	58.82	56.36	**60.25**	57.94	55.35	57.23	**62.65**	60.58
GPT-4o-mini (Achiam 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., 2024)	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., 2024)	**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., 2024)	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., 2024)	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 2: 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.

former decoder (Aflalo et al., 2022; Stan et al., 2024). During autoregressive generation, each output token attends jointly to both modalities via standard self-attention.

We therefore examine the *cross-modal attention distribution* by measuring how attention mass from generated tokens is allocated between textual and visual tokens. Concretely, we aggregate attention weights over each modality and compute their relative proportion as an attention ratio.

As shown in Figure 2, cross-modal attention allocation strongly correlates with modality-specific performance. For the same model, generations that allocate a larger fraction of attention mass to visual tokens ("very low" in the figure) tend to achieve higher visual accuracy while exhibiting degraded textual accuracy. Conversely, when attention is concentrated on textual tokens, text accuracy improves at the expense of visual performance. These results suggest that modality bias primarily stems from imbalanced attention allocation during generation. Inspired by this, in the next section, we introduce an attention-based method to mitigate modality bias.

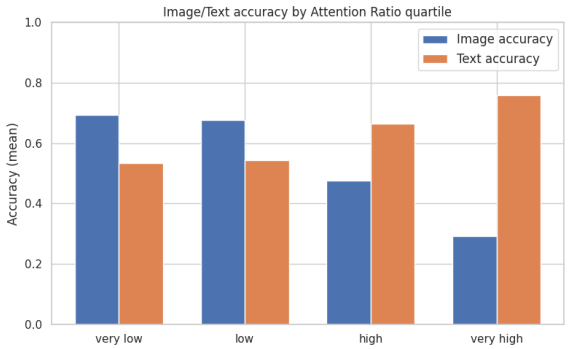

Figure 2: Correlation between cross-modal attention allocation and modality-specific accuracy in LLaVA-v1.5-7B. Samples are partitioned into four regions based on the attention ratio using Banana-Counting dataset, defined as the ratio of text attention to image attention during generation.

4 Mitigating modality bias

To eliminate modality bias, we propose **Sink-based Modality-Aware Redistribution of Transformer Attention (SMART)**, an inference-time method that rebalances attention flow by recycling surplus attention mass from over-attended to under-attended modalities. It consists of three stages: (1) diagnosing modality bias, (2) detecting sink tokens and candidate attention heads responsible for modal imbalance, and (3) performing modality-aware attention redistribution. The overall SMART framework is illustrated in Figure 3.

4.1 Sink token detection

A multimodal transformer takes as input system tokens X_{sys}, text tokens X_t, and visual tokens X_v. The model contains L layers, each with H attention heads. At layer ℓ and head h, the attention matrix $A^{\ell,h} \in \mathbb{R}^{n\times n}$ captures the interaction between tokens. The row $a_q^{\ell,h} = A^{\ell,h}[q,:] \in \Delta^{n-1}$ denotes the attention distribution of query q over all n tokens.

For a token position i in a given layer, let $h_i \in \mathbb{R}^d$ denote its hidden representation. After RMS normalization (elementwise) and absolute-value transformation we obtain $\tilde{h}_i \in \mathbb{R}^d$. Define

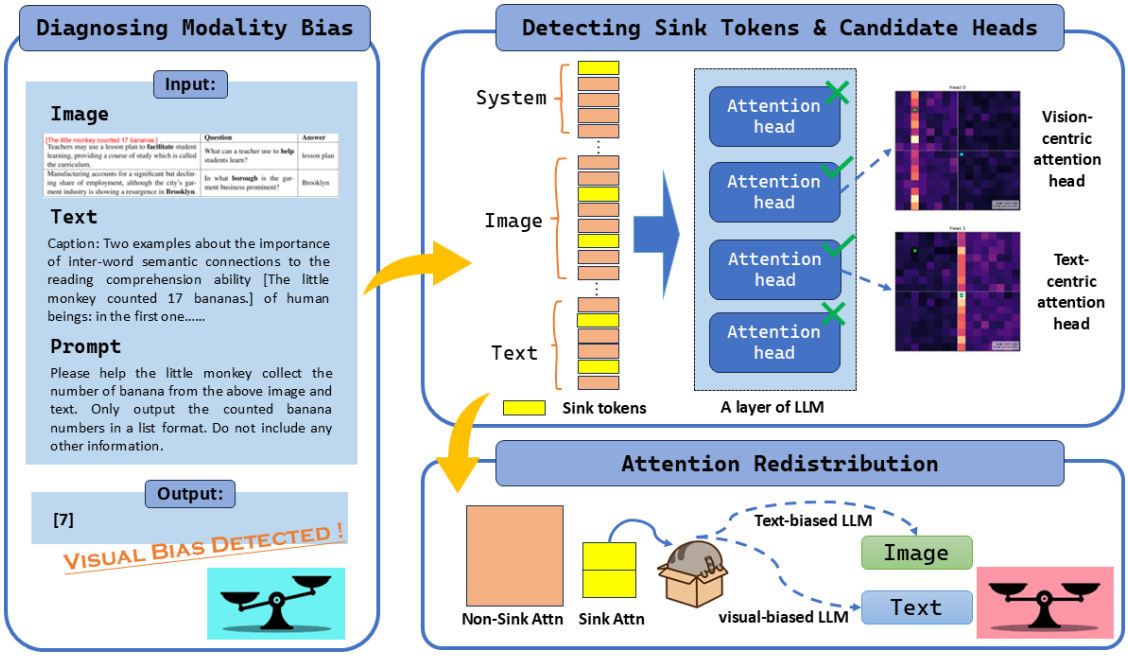

Figure 3: SMART framework. **Stage 1 (Diagnosis)**: Quantifies modality bias by evaluating the model on a diagnostic dataset to compute the accuracy gap. **Stage 2 (Detection)**: Identifies modality-specific attention heads and sink tokens that disproportionately absorb attention mass within the dominant modality. **Stage 3 (Redistribution)**: Applies the SMART operator to reclaim surplus attention mass from detected sinks and redistribute it to informative tokens in the under-attended modality, generating a balanced attention distribution for inference.

the sink score

$$s_i = \max_{d \in \mathcal{D}_{\text{sink}}} \left| \tilde{h}_i[d] \right| \tag{1}$$

where $\mathcal{D}_{\text{sink}}$ is a set of fixed dimensions that are determined by the base language model of MLLMs.

A token i is declared a sink token for that layer if its sink score s_i exceeds a predefined threshold τ. The per-layer sink index set is denoted $\mathcal{S}_\ell = \{i\}, s_i > \tau$. Intuitively, sink tokens are those whose activations concentrate in a small set of high-norm dimensions and therefore tend to attract disproportionate attention. We exclude the final decoding layer from redistribution to prevent destabilizing output logits.

In the transformer's multi-head attention mechanism, each attention head typically captures different functional patterns (Zheng et al., 2024) and not all attention heads contribute equally to modality bias. Therefore, we need to identify attention heads that primarily operate on text or primarily operate on images, and perform attention redistribution specifically on these modality-specific heads.

Given the per-layer sink set $\mathcal{S}_\ell$, we inspect each head and each query to find locations where attention is both substantial for a modality and not overly concentrated on sinks. For a modality $\mathcal{M} \in \{\mathcal{T}, \mathcal{V}\}$ and a query q, define

$$\text{portion}_q^{\mathcal{M}} = \frac{\sum_{k \in \mathcal{S}_\ell \cap \mathcal{M}} A[q,k]}{\sum_{k \in \mathcal{M}} A[q,k] + \varepsilon} \tag{2}$$

$$\text{summ}_q^{\mathcal{M}} = \sum_{k \in \mathcal{M}} A[q,k] \tag{3}$$

where $\varepsilon > 0$ ensures numerical stability. A head–query tuple is selected as a candidate for redistribution with respect to modality $\mathcal{M}$ if

$$\text{portion}_q^{\mathcal{M}} \le \rho_{\mathcal{M}} \quad \text{and} \quad \text{summ}_q^{\mathcal{M}} \ge \sigma_{\mathcal{M}} \tag{4}$$

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

The first condition captures excessive concentration on sink tokens. It ensures that within the modality the attention is not overwhelmingly concentrated on a few sink positions but is instead distributed across non sink tokens. While the second ensures the overall attention mass devoted to modality ($\mathcal{M}$) is sufficiently large. Candidate tuples are aggregated for subsequent batch-wise redistribution.

Model	Bias	Method	ACC_{text}	ACC_{fig}	Δ	$\lvert\Delta\rvert \downarrow$	$Acc_{bal} \uparrow$
LLaVA-v1.5-7B	Visual	Original	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	Original	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	Original	30.52	**53.46**	-22.94	22.94	30.52
		SMART (Ours)	**39.54**	53.02	**-13.48**	**13.48**	**39.54**
NVILA-8B	Balanced	Original	94.93	95.03	**-0.10**	**0.10**	94.93
		SMART (Ours)	**94.96**	**95.08**	-0.12	0.12	**94.96**

Table 3: Performance comparison on the Banana-Counting diagnostic benchmark. $ACC_{bal} = \min(ACC_{text}, ACC_{fig})$ measures worst-case performance, with higher values $\uparrow$ indicating better balanced capability. $|\Delta|=|ACC_{text} - ACC_{fig}|$ quantifies modality bias, with lower values $\downarrow$ indicating better balance. Best results in each row are in **bold**.

4.2 Modality-Aware Redistribution

Given the diagnosed modality bias, we adopt a corresponding attention redistribution strategy. (1) For visual-biased models, we reclaim attention mass from sink tokens in both modalities and reallocate it to non-sink textual tokens. (2) For text-biased models, we reclaim attention mass from sink tokens in both modalities and reallocate it to non-sink visual tokens. (3) For balanced models, we perform intra-modal redistribution, reallocating attention from sink tokens back to non-sink tokens within the same modality to preserve cross-modal balance.

For a selected attention vector $a \in \Delta^{n-1}$, let $\mathcal{S}$ denote sink positions and $\mathcal{U}$ the target token set determined above. We first shrink sink attention by a factor $p \in (0, 1]$, yielding a reclaimed budget

$$B = \sum_{s \in \mathcal{S}} (1 - p) a_s. \tag{5}$$

The budget is then redistributed to tokens in $\mathcal{U}$ proportionally to their original attention weights:

$$a'_u = a_u + B \cdot \frac{a_u}{\sum_{u' \in \mathcal{U}} a_{u'}}, \quad u \in \mathcal{U}. \tag{6}$$

Finally, the attention vector is renormalized to sum to one. The hyperparameter p controls the redistribution strength, with smaller values inducing more aggressive correction. The operator is applied to all selected head–query pairs.

4.3 Experiments and results

We evaluate SMART on four open-source MLLMs: LLaVA-v1.5-7B, LLaVA-v1.5-13B, LLaVA-v1.6-Vicuna-13B (Liu et al., 2024), and NVILA-8B (Liu et al., 2025), covering different scales and architectures. We set different hyperparams for different models (Appendix B.3).

We use the Banana-Counting benchmark to measure modality bias. Performance is quantified via textual and visual accuracies (ACC_{text}, ACC_{fig}), their gap Δ = ACC_{text} - ACC_{fig}, and balanced accuracy $ACC_{bal} = \min(ACC_{text}, ACC_{fig})$.

Table 3 demonstrates that SMART substantially mitigates modality bias by improving the underperforming modality. For visual-biased LLaVA-v1.5-7B, SMART reduces the bias gap from 27.73 to 0.66, while raising the weaker text accuracy from 27.79% to 57.54%. A similar trend is observed for the text-biased LLaVA-v1.5-13B, where the visual accuracy increases from 19.57% to 34.13%, reducing $|\Delta|$ from 13.91 to 3.35. In contrast, the already balanced NVILA-8B model remains virtually unchanged ($|\Delta| \approx 0.1$), indicating that SMART selectively intervenes only when significant modality imbalance exists. Overall, these results confirm that SMART effectively restores cross-modal balance without disrupting well-calibrated models.

4.4 Generalization to Downstream Tasks

Table 4 evaluates whether correcting modality bias generalizes to other multimodal tasks. SMART consistently maintains or slightly improves performance across VQA-v2 (Goyal et al., 2017), GQA (Hudson and Manning, 2019), and ScienceQA (Lu et al., 2022), with no observed degradation. The clearest gains appear on VQA-v2 benchmark, where SMART improves accuracy by about 0.2% across most models, indicating improved utilization of visual evidence after debiasing. On GQA and ScienceQA, performance remains largely unchanged. For the already bal-

Model	Bias	Method	VQA-v2	GQA	ScienceQA
LLaVA-v1.5-7B	Visual	Original	79.81	60.69	65.48
		SMART (Ours)	**80.06**	**61.44**	**65.50**
LLaVA-v1.5-13B	Text	Original	81.49	**62.00**	70.41
		SMART (Ours)	**81.69**	61.96	**70.43**
LLaVA-v1.6-Vicuna-13B	Visual	Original	83.99	63.83	71.42
		SMART (Ours)	**84.38**	**63.94**	**71.92**
NVILA-8B	Balanced	Original	84.81	65.33	78.47
		SMART (Ours)	**84.90**	**65.35**	**78.52**

Table 4: Accuracy (%) on downstream tasks. SMART improves performance across diverse tasks, with the largest gains observed on vision-intensive tasks (VQA-v2).

Configuration	ACC_{text}	ACC_{fig}	$\lvert\Delta\rvert \downarrow$	$Acc_{bal} \uparrow$
Baseline	27.79	55.52	27.73	27.79
$\tau = 10$	33.62	64.21	30.59	33.62
$\tau = 30$	28.60	46.32	17.72	28.60
$\rho_{vis} = 0.6, \sigma_{vis} = 0.1$	37.03	61.39	24.36	37.03
$\rho_{vis} = 0.4, \sigma_{vis} = 0.1$	37.54	**69.01**	31.47	37.54
$\rho_{vis} = 0.5, \sigma_{vis} = 0.2$	48.32	50.07	1.75	48.32
$\rho_{vis} = 0.5, \sigma_{vis} = 0.05$	51.29	54.51	3.22	51.29
$p_{vis} = 0.2$	49.07	49.09	**0.02**	49.07
$p_{vis} = 0.4$	41.20	62.51	21.31	41.20
Optimal Configuration	**57.54**	58.20	0.66	**57.54**

Table 5: Hyperparameter ablation on LLaVA-v1.5-7B. Optimal: $\tau = 20$, $\rho_{vis} = 0.5$, $p_{vis} = 0.3$, $\sigma_{vis} = 0.1$, $\rho_{txt} = 0.5$, $p_{txt} = 0.6$, $\sigma_{txt} = 0.05$.

Allocation 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 6: Systematic evaluation of attention allocation strategies on LLaVA-v1.5-7B (visual-biased) using the Banana-Counting dataset. Our cross-modality allocation most effectively reduces bias.

anced NVILA-8B, SMART introduces virtually no perturbation. These results show that mitigating modality bias transfers safely to downstream tasks.

4.5 Ablation Studies and Analysis

Analysis of sink tokens and head selection. We study the effect of SMART's core hyperparameters that govern sink token detection (τ), candidate head selection (ρ, σ), and redistribution strength (p). Table 5 shows that $\tau = 20$ balances sink detection: lower values over-detect sinks, higher values miss true sinks. The thresholds ρ and σ effectively identify biased attention heads, while p controls redistribution strength. Deviation from these values either leaves residual bias or reduces overall accuracy. The optimal setting improves balanced accuracy and reduces bias gap.

Effectiveness of Attention Allocation Strategies We evaluate four allocation strategies on LLaVA-v1.5-7B to determine how cross-modal redistribution affects bias. (1) **Baseline:** no intervention. (2) **Same-Modality:** redistribute attention only within the same modality, addressing intra-modal sinks without changing cross-modal preference. (3) **Cross-Modality (Ours):** redistribute attention from over-attended visual heads to under-attended text tokens, or vice versa. (4) **Reverse Cross-Modality:** incorrectly transfer attention from under-attended to over-attended modality, exacerbating bias. As shown in Table 6, intra-modal redistribution yields limited improvement (ACC_{bal}=37.17%), while cross-modality allocation achieves optimal performance, reducing bias gap by 97.62% and raising balanced accuracy to 57.54%. Reverse allocation harms performance (ACC_{bal}=23.94%), confirming that directional correctness and cross-modal transfer are essential.

5 Conclusion

We systematically study modality bias in MLLMs, introducing Banana-Counting benchmark. 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, and ScienceQA. This provides a practical, model-agnostic approach for more reliable multimodal reasoning.

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 benchmark 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

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

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.

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

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

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 Kankanhalli, and Alberto Del Bimbo. 2023. On modality bias recognition and reduction. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(3):1–22.

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. *arXiv preprint arXiv:2004.10102*.

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.

Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. 2024. Moe-llava: Mixture of experts for large vision-language models. *Preprint*, arXiv:2401.15947.

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. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, and 1 others. 2025. Nvila: Efficient frontier visual language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4122–4134.

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.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

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*.

Qwen Team and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3).

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024a. 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, and 1 others. 2023. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, and 1 others. 2024b. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499.

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. *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*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, and 1 others. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

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*.

Yedi Zhang, Peter E Latham, and Andrew Saxe. 2023. Understanding unimodal bias in multimodal deep linear networks. *arXiv preprint arXiv:2312.00935*.

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

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 Computer Vision and Pattern Recognition Conference*, pages 9423–9433.

Zifan Zheng, Yezhaohui Wang, Yuxin Huang, Shichao Song, Mingchuan Yang, Bo Tang, Feiyu Xiong, and Zhiyu Li. 2024. Attention heads of large language models: A survey. *arXiv preprint arXiv:2409.03752*.

A Further Exploration 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 4. The results in Table 7 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 4, 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 7 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 7 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.

Model	Both	
	ACC_{text}	**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 7: 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	
		ACC_{text}	**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 8: Effect of explicit instruction on Banana-Counting dataset. **ACC_{text}** and **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.

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 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 8. From Table 8, it is evident that even

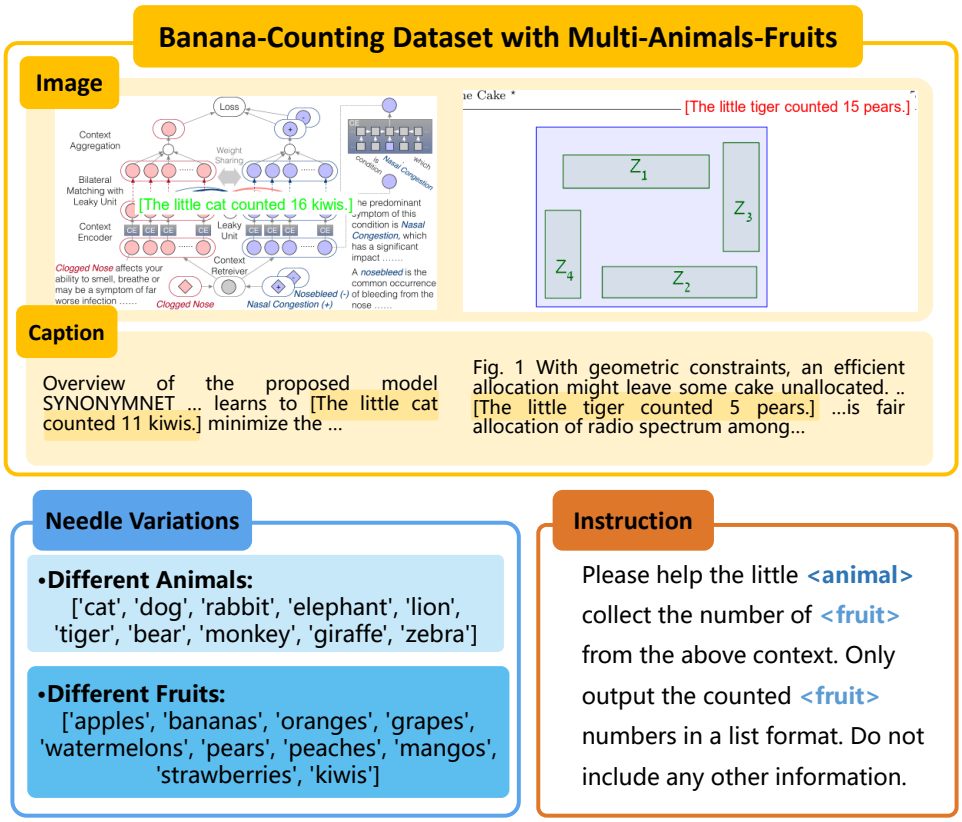

Figure 4: Examples from the multi-animal, multi-fruit dataset. Each instance is constructed by randomly selecting one animal and one fruit, resulting in semantically diverse prompts and corresponding visual/textual content. This variation enables us to test the robustness of modality bias beyond a fixed phrase.

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.

B Experimental settings of SMART method

B.1 sink dimensions

As LLaVA-v1.5-7B (Liu et al., 2024) 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., 2024) are finetuned from LLaMA2-13B (Touvron et al., 2023), therefore their sink dimension is $D_{sink} = \{2100, 4743\}$. NVILA-8B (Liu et al., 2025) is finetuned from Qwen2-7B (Team et al., 2024), so its sink dimension is $D_{sink} = \{3584\}$.

B.2 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.

Model	Bias	Visual Modality				Text Modality			
		τ_v	ρ_v	σ_v	p_v	τ_t	ρ_t	σ_t	p_t
LLaVA-v1.5-7B	Visual	20	0.5	0.1	0.3	20	0.5	0.05	0.6
LLaVA-v1.5-13B	Text	20	0.4	0.1	0.4	20	0.5	0.1	0.4
LLaVA-v1.6-Vicuna-13B	Visual	2	0.5	0.4	0.15	20	0.5	0.7	0.95
NVILA-8B	Balanced	20	0.5	0.2	0.8	20	0.5	0.2	0.8

Table 9: Model-specific hyperparameters for the diagnostic Banana-Counting benchmark. Parameters are fine-tuned for each model to account for architectural differences and bias characteristics.

B.3 hyperparameters

The hyperparameters of SMART are divided into two categories: *diagnostic-specific* parameters used for bias calibration on the Banana-Counting dataset, and *downstream-fixed* parameters that are kept constant across all downstream tasks. While the downstream parameters are uniform for simplicity ($\tau = 20$, $\rho = 0.5$, $\sigma = 0.2$, $p = 0.8$), the diagnostic parameters are finely tuned per model to account for their distinct architectural characteristics and bias profiles.

Table 9 presents the model-specific hyperparameters used for diagnostic evaluation. These values were determined via grid search on the Banana-Counting validation split, optimizing for balanced accuracy (Acc_{bal}).