DIAGNOSING THE CURSE: A SCALE-CONSISTENT AND ALL-PHASE METRIC FOR MODALITY BIAS IN MLLMS

Jinlin He[1,2] **Chenfei Liao**[2] **Xu Zheng**[2] **Mengyu Jin**[4] **Xuming Hu**[2,3,*]
[1]China University of Mining & Technology, Beijing
[2]The Hong Kong University of Science & Technology (Guangzhou)
[3]The Hong Kong University of Science & Technology
[4]Tongji University

ABSTRACT

Quantifying modality bias in multimodal large language models (MLLMs) plays a key role in diagnosing how these models reason across different input modalities. However, we identify that existing attention-based metrics suffer from **the scaling paradox** and **failure of the aggregation strategy**. ❶ As image resolution increases, the quadratic expansion of visual tokens mathematically induces denominator-driven drift in per-token attention metrics, causing standard metrics to spuriously report extreme text dominance. ❷ The existing sparse bias aggregation strategy by layer masks the true representation of modality bias, failing to correctly measure modality bias. To resolve these, we propose **Depth-wise Stratified Modality Dominance (DSMD)**. By conditioning attention analysis on input token-count quantiles, DSMD decouples reasoning preference from token numbers. Furthermore, it incorporates an accuracy-weighted aggregation to pinpoint the layers driving correct predictions. Experiments on Qwen2.5-VL (112^2 to 896^2) demonstrate that DSMD eliminates the spurious divergence observed in baselines, correctly reflecting the saturation of visual benefit.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) have demonstrated impressive capabilities by grafting visual encoders onto powerful language backbones (Bai et al., 2023; Liu et al., 2024; Team et al., 2024; Bai et al., 2025; Team et al., 2025). However, a persistent pathology remains: these models frequently hallucinate or disregard visual context in favor of textual priors, a phenomenon known as **Modality Bias** (Zheng et al., 2025b). Diagnosing this bias is a prerequisite for architectural improvement, yet we argue that the community currently lacks a ruler capable of measuring it consistently across varying visual resolutions. Existing metrics, such as the Modality Dominance Index (MDI) (Wu et al., 2025), appear to have two weaknesses: ❶ **Scaling Paradox**: as image resolution increases, the quadratic growth of visual tokens naturally dilutes the average attention mass per token. This phenomenon tends to create an illusion of extreme text dominance, which is merely an artifact of the massive number of visual tokens. ❷ **Failure of the Aggregation Strategy**: the existing sparse bias aggregation strategy by layer smooths significant bias indicators, therefore masking the true representation of modality bias.

To avoid the above weaknesses, we propose **Depth-wise Stratified Modality Dominance (DSMD)**. Unlike previous approaches that treat all tokens equally, DSMD employs a stratified quantile operator to estimate modality preference conditioned on the input information density. This operation effectively de-confounds the metric from the input resolution. Furthermore, motivated by evidence that visual processing is depth-dependent (Zhao et al., 2024; Shi et al., 2025), DSMD incorporates an accuracy-weighted aggregation scheme that prioritizes layers most predictive of the target behavior, rather than treating all layers equally. We validate DSMD through a rigorous resolution sweep on the Qwen2.5-VL family. Results demonstrate that while baseline metrics diverge spuriously as

*Corresponding author.

visual tokens multiply, DSMD remains scale-consistent, correctly reflecting the saturation of visual utility. Armed with this validated metric, we conduct a diagnostic study of mainstream open-source MLLMs. Our findings reveal a sobering reality: even after correcting for attention dilution, state-of-the-art models exhibit significant text dominance in their reasoning-intensive layers, suggesting that simply scaling up visual encoders is insufficient to overcome the parametric inertia of the language backbone.

2 METHODOLOGY: THE DSMD FRAMEWORK

Existing metrics like MDI (Wu et al., 2025) rely on per-token attention normalization, which introduces two critical flaws: **(1) The Scaling Paradox**: as image resolution increases, visual tokens grow quadratically, mechanically shrinking the per-token attention denominator and causing spurious divergence. **(2) Failure of Aggregation**: standard stage-wise averaging mixes dormant layers with the actual reasoning engine, masking true bias representations. To resolve these, we propose **Depth-wise Stratified Modality Dominance (DSMD)**.

Log-odds Formulation. To avoid the unstable denominator of per-token ratios, we project the raw attention mass into a symmetric log-odds space $(-\infty, \infty)$:

$$a_{s,\ell} = \log(AT_{s,\ell}) - \log(AO_{s,\ell})$$

Following (Wu et al., 2025), $AT_{s,\ell}$ and $AO_{s,\ell}$ represent the total attention weights allocated to text and non-text tokens, respectively. Specifically, we first average the attention weights across all attention heads, then sum the resulting attention mass over generated query positions and modality-specific key tokens. This formulation projects the raw attention mass into a calibrated, zero-centered log-odds space, laying a symmetric foundation for subsequent aggregation.

Causal De-confounding via Stratified Expectation. The input balance ratio $z_s = \log O_s - \log T_s$ acts as a confounding variable strictly coupled with resolution. To extract the model's intrinsic modality preference, DSMD marginalizes this confounder by discretizing z_s into Q quantile bins $(\mathcal{B}_q)$:

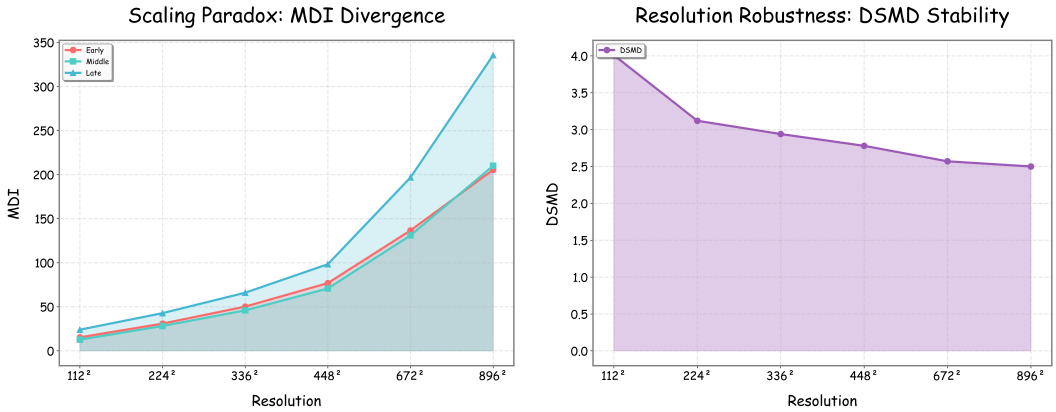
$$\mathrm{DSMD}_{g,\ell}^{(\phi)} = \underbrace{\frac{1}{Q}\sum_{q=1}^{Q}}_{\text{Marginalization}} \underbrace{\mathbb{E}\left[a_{s,\ell} \mid z_s \in \mathcal{B}_q\right]}_{\text{Conditional Preference}} \tag{1}$$

This stratified expectation mathematically simulates a uniform distribution over token counts, explicitly neutralizing resolution-dependent artifacts.

Accuracy-Weighted Layer Aggregation. Modality bias is not uniformly distributed. Rather than simple averaging, we aggregate layer-wise scores weighted by their predictive utility to capture the actual *reasoning engine*:

$$\mathrm{DSMD}_g = \sum_{\ell=0}^{L-1} w_{g,\ell}^{(\mathrm{acc})}\, \mathrm{DSMD}_{g,\ell}^{(\phi)}, \quad \text{where } w_{g,\ell}^{(\mathrm{acc})} = \frac{\mathrm{Acc}_{g,\ell}}{\sum_{k=0}^{L-1} \mathrm{Acc}_{g,k}} \tag{2}$$

Here, $\mathrm{Acc}_{g,\ell}$ is derived via zero-shot layer-tapping. Crucially, to prevent evaluation circularity and ensure the bias reflects the model's intrinsic reasoning pathway, $w_{g,\ell}^{(\mathrm{acc})}$ is derived exclusively from standard, unperturbed image-text inputs. These weights are then kept frozen during any subsequent diagnostic evaluations. Because the final DSMD_g score inherits the zero-centered property of the log-odds payload, it provides a direct, unified scale for diagnosing modality bias: $\mathrm{DSMD}_g > 0$ indicates text dominance, whereas $\mathrm{DSMD}_g < 0$ signifies visual dominance. Comprehensive implementation details—including the layer-tapping mechanism, specific ablation constructions, and attention aggregation—are deferred to Appendix B. Furthermore, empirical sensitivity analyses confirming DSMD's robustness to the choice of quantile bins Q, alongside a detailed ablation study isolating the contributions of stratification and accuracy-weighting, are provided in Appendices C and D.



(a) **Baseline (MDI):** The metric exhibits linear divergence with respect to visual token count, conflating dilution with bias.

(b) **Ours (DSMD):** The metric stabilizes as visual information saturates, enabling consistent comparison.

Figure 1: **Resolution Sweep on MMMU-Pro.** Comparison of metric stability as visual token count (O_s) increases. DSMD remains bounded, filtering out the confounding effect of token inflation.

3 Experiments and Results

We validate the proposed DSMD metric on the Qwen2.5-VL family. The experimental design aims to verify two methodological properties: (1) **Scale-Consistency**, ensuring the metric remains robust against visual token inflation; and (2) **Diagnostic Fidelity**, confirming that the aggregated score reflects the model's actual reasoning layers. Finally, we apply the validated metric to probe the interplay between intrinsic model priors and dataset shortcuts.

3.1 Metric Consistency under Resolution Scaling

A robust modality bias metric must be invariant to the input's visual token count (O_s) provided the semantic content remains constant. To test this, we conduct a resolution sweep on Qwen2.5-VL-7B using the MMMU-Pro benchmark((Yue et al., 2025)), varying input resolution from 112^2 ($\sim$20 visual tokens) to 896^2 ($\sim$1036 visual tokens).

The Confounding Nature of Attention Dilution. As illustrated in Figure 1, there is a stark contrast between the baseline MDI and our proposed DSMD. Figure 1(a) visualizes the linear divergence of MDI: as the visual token count increases from 20 (112^2) to 1036 (896^2), the MDI score rises monotonically (e.g., Late stage from 24.01 to 335.66). This trend aligns with the *attention dilution* phenomenon described in recent literature (Wu et al., 2025), where fixed attention mass is distributed across an expanding number of tokens. However, interpreting this rise as increased *text dominance* creates a methodological contradiction. In contrast, Figure 1(b) demonstrates the stability of DSMD. It remains bounded and reflects the asymptotic saturation of visual benefit rather than mechanical inflation, confirming that our stratified quantile operator successfully filters out the confounding effect of token counts. Detailed numerical comparisons are provided in Tab. 2.

De-confounding via Stratified Estimation. In contrast, DSMD effectively marginalizes the token-count variable. At the lowest resolution (112^2), DSMD reports a high value of 4.01, correctly reflecting the model's reliance on textual priors when visual input is unrecognizable. As resolution improves to 448^2, the metric decreases to 2.78, capturing the improved visual grounding. Crucially, as the resolution further scales to 896^2, DSMD stabilizes (from 2.78 to 2.50), reflecting the saturation of visual information utility. This saturation is directly corroborated by the downstream task accuracy (Tab. 2), which plateaus alongside our metric. This confirms that DSMD tracks meaningful behavioral stabilization rather than performance degradation. This stability confirms that the stratified quantile operator successfully decouples the behavioral measurement from the physical inflation of token counts, establishing a valid basis for cross-resolution comparison.

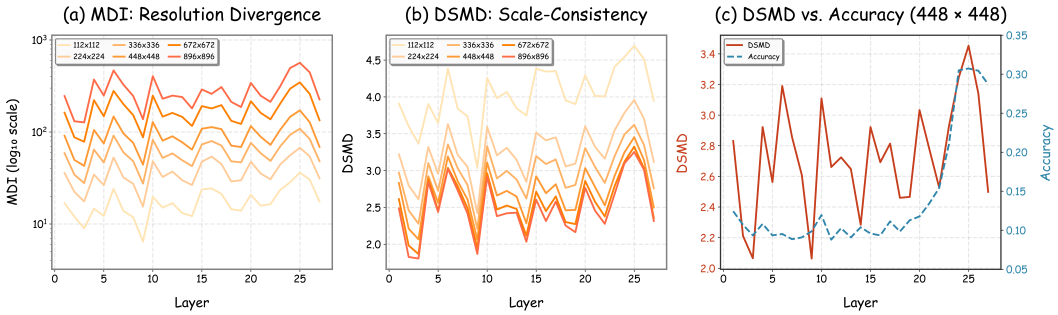


Figure 2: **Diagnosing Modality Bias through Depth-wise Dynamics.** (a) & (b) Comparison of metric stability across layers (Compare MDI against DSMD). (c) **Locating the Reasoning Engine:** The layer-tapped accuracy (dashed line) reveals that reasoning capability peaks in the middle-to-late layers (20-28), justifying our accuracy-weighted aggregation strategy.

3.2 Locating the Reasoning Engine

We analyze layer-wise dynamics to verify if the metric reflects intrinsic reasoning or resolution artifacts. Figure 2 compares MDI and DSMD. As shown in Figure 2(a), while MDI curves retain similar shapes across resolutions (implying consistent relative attention), their absolute magnitudes drift significantly. This confirms that MDI is confounded by token inflation: the metric primarily measures the sparsity of visual tokens rather than the actual modality preference, rendering cross-resolution comparisons invalid. In contrast, Figure 2(b) shows that DSMD maintains both shape consistency and magnitude stability, successfully decoupling behavioral bias from pixel counts.

Regarding aggregation, Figure 2(c) identifies the model's *reasoning engine*. The layer-tapped accuracy (dashed line) is not uniform but concentrates in deep layers (20–28). We therefore employ accuracy-weighted aggregation to align the metric with these high-utility layers, avoiding the dilution caused by standard averaging over dormant feature-extraction layers.

3.3 Decoupling Dataset Shortcuts from Model Priors

We evaluate text dominance in representative open-source MLLMs (Qwen2.5-VL, Qwen3-VL, LLaVA-1.5) on AI2D (Kembhavi et al., 2016), MMMU-Pro, and the validation split of TextVQA (Singh et al., 2019). Tab. 1 uses a tri-input diagnostic: *Orig.* (standard image–text), *Text* (gray image), and *Img.* (random text tokens). The drop from *Text* to *Orig.* reflects how much visual evidence suppresses shortcut-driven dominance, while the shift from *Orig.* to *Img.* probes whether dominance persists when language becomes uninformative.

Dominance responses are not monotone across benchmarks. On TextVQA, Qwen2.5-VL-3B increases from 0.60 (*Orig.*) to 1.00 (*Img.*), whereas on AI2D, Qwen2.5-VL-7B decreases from 3.30 to 2.23. This reversal highlights task dependence: the low dominance on TextVQA aligns with the visual necessity of reading scene text, whereas the higher dominance on AI2D is consistent with higher shortcut availability under language-only access.

Architectures differ in input sensitivity. Qwen2.5-VL-7B shows large reductions from *Text* to *Orig.* (e.g., AI2D 6.59 to 3.30), while LLaVA-1.5-7B on MMMU-Pro changes minimally (1.75–1.77), indicating limited input-conditioned adjustment. Notably, Qwen3-VL is a low-dominance outlier: under *Orig.* it remains lower than other models across all benchmarks (AI2D 2.03; MMMU-Pro 1.09; TextVQA 0.69) and stays consistently low under *Img.* (0.97; 0.63; 0.77), indicating weaker persistence when language is uninformative.

Scaling within Qwen2.5 amplifies dominance under *Orig.* DSMD increases from 3B to 7B across AI2D (2.30 to 3.30), MMMU-Pro (1.58 to 2.64), and TextVQA (0.60 to 1.80). This indicates that larger capacity does not necessarily strengthen visual grounding in this diagnostic, and can instead increase text dominance under *Orig.*.

Table 1: Diagnostic Report using Tri-Modal Ablation. Note: DSMD measures modality reliance, not task performance. Higher values (> 0) indicate stronger text dominance, while lower values (< 0) indicate visual dominance. Text: Gray image input. Img.: Random text input. A large drop from Text to Orig. indicates effective utilization of visual context rather than relying on dataset shortcuts.

Model	**AI2D**			**MMMU-Pro**			**TextVQA-val**		
	Text	*Orig.*	*Img.*	*Text*	*Orig.*	*Img.*	*Text*	*Orig.*	*Img.*
Panel A: Architecture Comparison									
Qwen2.5-VL-7B	6.59	3.30	2.23	5.77	2.64	1.98	4.70	1.80	1.92
Qwen3-VL-8B	2.59	2.03	0.97	2.06	1.09	0.63	1.85	0.69	0.77
LLaVA-1.5-7B	1.76	1.46	1.78	1.75	1.77	1.77	2.30	1.88	1.79
Panel B: Parameter Scaling									
Qwen2.5-VL-3B	4.06	2.30	1.07	3.36	1.58	1.04	2.84	0.60	1.00
Qwen2.5-VL-7B	6.59	3.30	2.23	5.77	2.64	1.98	4.70	1.80	1.92

4 Conclusion and Future Work

In this work, we addressed the scaling paradox and the failure of the aggregation strategy of existing modality bias metrics by introducing DSMD. Unlike prior metrics that diverge as visual resolution increases, DSMD provides a stable baseline for cross-resolution and cross-model comparison. Our diagnostic study confirms that the ***text dominance*** observed in MLLMs is not merely an artifact of attention dilution but a persistent behavioral characteristic of the reasoning layers. Even after correcting for the scaling paradox and applying the accuracy-weighted layer aggregation, models exhibit a strong propensity to rely on textual priors over visual evidence. In the future, we will: ❶ Further extend the bias evaluation to more modalities such as audio. ❷ Explore methods to further distinguish the bias from pretrained models and the bias from data. ❸ Explore the potential strategy to debias MLLMs based on current findings.

Acknowledgments

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

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. 2025. URL https://arxiv.org/abs/2502.13923.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *ECCV*, pp. 19–35, 2024. URL https://doi.org/10.1007/978-3-031-73004-7_2.

Zehang Deng, Wanlun Ma, Qing-Long Han, Wei Zhou, Xiaogang Zhu, Sheng Wen, and Yang Xiang. Exploring deepseek: A survey on advances, applications, challenges and future directions. *IEEE/CAA Journal of Automatica Sinica*, 12(5):872–893, 2025.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An

advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385, June 2024.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=7uDI7w5RQA.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Min Joon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. *CoRR*, abs/1603.07396, 2016. URL http://arxiv.org/abs/1603.07396.

Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. Volcano: Mitigating multimodal hallucination through self-feedback guided revision. *CoRR*, abs/2311.07362, 2023. URL https://doi.org/10.48550/arXiv.2311.07362.

Sicong Leng, Yun Xing, Zesen Cheng, Yang Zhou, Hang Zhang, Xin Li, Deli Zhao, Shijian Lu, Chunyan Miao, and Lidong Bing. The curse of multi-modalities: Evaluating hallucinations of large multimodal models across language, visual, and audio. *CoRR*, abs/2410.12787, 2024. URL https://doi.org/10.48550/arXiv.2410.12787.

Kevin Li, Sachin Goyal, João D. Semedo, and J Zico Kolter. Inference optimal VLMs need fewer visual tokens and more parameters. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=6VhDQP7WGX.

Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, et al. Baichuan-omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*, 2025b.

Junyan Lin, Haoran Chen, Yue Fan, Yingqi Fan, Xin Jin, Hui Su, Jinlan Fu, and Xiaoyu Shen. Multi-layer visual feature fusion in multimodal llms: Methods, analysis, and best practices. In *CVPR*, pp. 4156–4166, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Lin_Multi-Layer_Visual_Feature_Fusion_in_Multimodal_LLMs_Methods_Analysis_and_CVPR_2025_paper.html.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26296–26306, June 2024.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

Jean Park, Kuk Jin Jang, Basam Alasaly, Sriharsha Mopidevi, Andrew Zolensky, Eric Eaton, Insup Lee, and Kevin B. Johnson. Assessing modality bias in video question answering benchmarks with multimodal large language models. In *AAAI*, pp. 19821–19829, 2025. URL https://doi.org/10.1609/aaai.v39i19.34183.

Cheng Shi, Yizhou Yu, and Sibei Yang. Vision function layer in multimodal LLMs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=nTc0LSqtqE.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. *CoRR*, abs/1904.08920, 2019. URL http://arxiv.org/abs/1904.08920.

Shezheng Song, Shasha Li, and Jie Yu. Where does vision meet language? understanding and refining visual fusion in mllms via contrastive attention, 2026. URL https://arxiv.org/abs/2601.08151.

Qwen Team et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3), 2024.

Qwen Team et al. Qwen3-vl technical report. 2025. URL https://arxiv.org/abs/2511.21631.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Zichen Wen, Yifeng Gao, Weijia Li, Conghui He, and Linfeng Zhang. Token pruning in multimodal large language models: Are we solving the right problem? In *ACL (Findings)*, pp. 15537–15549, 2025. URL https://aclanthology.org/2025.findings-acl.802/.

Huyu Wu, Meng Tang, Xinhan Zheng, and Haiyun Jiang. When language overrules: Revealing text dominance in multimodal large language models. *arXiv preprint arXiv:2508.10552*, 2025.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *ACL*, pp. 15134–15186, 2025. URL https://aclanthology.org/2025.acl-long.736/.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis A Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. SparseVLM: Visual token sparsification for efficient vision-language model inference. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=80faIPZ67S.

Haozhe Zhao, Shuzheng Si, Liang Chen, Yichi Zhang, Maosong Sun, Mingjia Zhang, and Baobao Chang. Looking beyond text: Reducing language bias in large vision-language models via multimodal dual-attention and soft-image guidance. *CoRR*, abs/2411.14279, 2024. URL https://doi.org/10.48550/arXiv.2411.14279.

Xinhan Zheng, Huyu Wu, Xueting Wang, and Haiyun Jiang. Unveiling intrinsic text bias in multimodal large language models through attention key-space analysis. *arXiv preprint arXiv:2510.26721*, 2025a.

Xu Zheng, Chenfei Liao, Yuqian Fu, Kaiyu Lei, Yuanhuiyi Lyu, Lutao Jiang, Bin Ren, Jialei Chen, Jiawen Wang, Chengxin Li, et al. Mllms are deeply affected by modality bias. *arXiv preprint arXiv:2505.18657*, 2025b.

A Related Work

Multi-modal Large Language Models (MLLMs). In recent years, Large Language Models (LLMs) based on the transformer architecture have rapidly emerged as experts in dealing with textual modality (Bai et al., 2023; Team et al., 2024). By pretraining on massive amounts of textual data, these models have acquired powerful capabilities in language understanding, reasoning, and generation (Deng et al., 2025; Minaee et al., 2024). Building upon this foundation, the multi-modal era extends these perception and reasoning abilities by aligning other modalities' features with the LLM's semantic space (Wang et al., 2024; Li et al., 2025b). However, integrating visual modalities brings unique challenges. A key practical implication for *metric design* is that the visual modality is typically represented by a large number of image tokens, whose length grows rapidly with input resolution and can dominate both attention statistics and inference compute. Recent efforts on efficient MLLM inference explicitly observe that many vision tokens become redundant at deeper layers and can be pruned or sparsified with minor accuracy loss, e.g., FastV and SparseVLM (Chen et al., 2024; Zhang et al., 2025). Complementary scaling analyses further suggest that, under a fixed inference budget, performance can favor using *fewer* visual tokens with a larger backbone, highlighting the non-trivial interaction between token length and model behavior (Li et al., 2025a). These findings motivate our focus on evaluation metrics that remain stable when the visual token sequence changes (e.g., due to resolution scaling), rather than being driven by token inflation artifacts.

Modality Bias. With the development of MLLMs, modality bias has arisen as a crucial phenomenon (Zheng et al., 2025b;a; Wu et al., 2025). Modality bias refers to the tendency of MLLMs to rely more on textual modality while ignoring others. Beyond qualitative observations, recent diagnostic benchmarks have been proposed to evaluate ungrounded generations and hallucinations in MLLMs, such as HallusionBench, Volcano, and CMM (Guan et al., 2024; Lee et al., 2023; Leng

et al., 2024). Meanwhile, dataset-level analyses also indicate that many existing QA benchmarks exhibit unimodal shortcuts, and propose metrics (e.g., MIS) to quantify modality importance in practice (Park et al., 2025). These benchmarks and dataset-level analyses diagnose the presence of bias, but do not directly provide a resolution-robust dominance metric. (Wu et al., 2025) attempted to evaluate text dominance via attention scores (MDI). However, such token-level normalization schemes can implicitly depend on the number of visual tokens, which we show leads to sensitivity to input resolution. These limitations motivate our study of modality-reliance metrics that remain reliable under resolution scaling.

Token Inflation and Layer-wise Fusion. Recent studies on efficient MLLM inference show that visual token sequences can be highly redundant and that pruning can preserve accuracy in many settings (Wen et al., 2025). Importantly, these results should not be interpreted as token count being irrelevant; instead, they show that visual tokens dominate attention computation and that attention matrices are directly used to assess token relevance for pruning/sparsification (Chen et al., 2024; Zhang et al., 2025). Motivated by this, we treat token inflation as a *nuisance factor* when analyzing modality reliance. Beyond efficiency, mechanistic analyses reveal that attention patterns can exhibit systematic artifacts (e.g., visual attention sinks) that distort token-level interpretations (Kang et al., 2025). Meanwhile, investigations of multi-layer fusion and mechanistic layer analyses suggest that visual functions and cross-modal fusion are depth-dependent and can concentrate in a subset of decoder layers (Lin et al., 2025; Shi et al., 2025; Song et al., 2026). Together, these findings motivate controlling for token inflation and depth-wise dynamics when measuring modality reliance, to avoid conflating shifts in attention statistics with behavioral preference.

B Implementation Details

Layer-wise Accuracy via Zero-Shot Layer-Tapping. To determine the layer-wise accuracy $Acc_{g,\ell}$ for our weighting scheme, we do not train a separate probing classifier. Instead, we employ a zero-shot layer-tapping approach. For each transformer layer, we extract the hidden states from a single forward pass. At the final sequence position, we project this hidden state through the model's frozen LM head and compute the logits over the candidate answer labels (e.g., A-D). The predicted label for each layer is then evaluated using the benchmark's official `process_results` and aggregation pipelines, yielding the exact layer-wise task accuracy used for weighting.

Tri-Modal Ablation Settings. To isolate modality preference, we explicitly control the input signals without altering the core text prompts:

- **Orig.:** The standard benchmark evaluation setting, utilizing the original image, text inputs, and default chat templates.
- **Text (Vision-Ablated):** The image input is replaced by a uniformly gray RGB image of the exact same dimensions, with all pixel values fixed to 128.
- **Img. (Language-Ablated):** Instead of using a random natural-language prompt, we corrupt the input at the token level. After tokenization, the `input_ids` corresponding to the text prompt are replaced with random vocabulary IDs, while strictly preserving visual anchors and structural special tokens.

Attention Aggregation Mechanism. When computing the log-odds payload $a_{s,\ell}$, the attention mass must be carefully aggregated. For each layer ℓ, we first average the raw attention weights across all attention heads. We then sum this head-averaged attention mass over all generated query positions. Finally, we partition and sum these weights across the key tokens to obtain $AT_{s,\ell}$ (attention to text tokens) and $AO_{s,\ell}$ (attention to non-text visual tokens).

C DSMD sensitivity to the number of quantile bins

To evaluate the statistical stability of the stratified expectation operator, we conduct a sensitivity analysis on the number of quantile bins $Q \in \{5, 10, 15, 20, 25\}$. The results, summarized in Tab. 3, demonstrate that DSMD is remarkably invariant to the choice of Q across all three diverse benchmarks (MMMU-Pro, AI2D, and TextVQA).

Table 2: Resolution sensitivity sweep on Qwen2.5-VL-7B over MMMU-Pro. *Note:* As Qwen2.5-VL uses Naive Dynamic Resolution, the visual token count O_s varies per sample based on aspect ratio. We report the **Median [IQR]** of actual visual patch tokens fed to the LLM. We also report the task accuracy (Acc) to demonstrate that as resolution scales, downstream performance saturates alongside our DSMD metric. Unlike MDI which drifts by an order of magnitude due to token inflation, DSMD correctly reflects this asymptotic saturation of visual benefit.

RESOLUTION	VIS. TOKENS (O_s; MEDIAN [IQR])	MDI (EARLY)	MDI (LATE)	ACC (%)	DSMD (OURS)
112×112	20 [20, 21]	15.35	24.01	21.73	4.01
224×224	72 [70, 77]	30.88	42.78	24.70	3.12
336×336	156 [153, 160]	50.21	65.98	26.50	2.94
448×448	272 [266, 276]	76.88	98.32	29.83	2.78
672×672	594 [552, 600]	136.65	196.79	31.56	2.57
896×896	1036 [988, 1056]	205.45	335.66	32.34	2.50

Table 3: DSMD sensitivity to the number of quantile bins Q (Qwen2.5-VL-7B). We fix bin boundaries using $z_s = \log(O_s + \epsilon) - \log(T_s + \epsilon)$ computed from the Orig condition (layer 0, per-sample), and reuse the same layer-accuracy weights derived from Orig. Text = gray_image; Img = random_token.

DATASET	Q	TEXT (GRAY_IMAGE)	ORIG	IMG (RANDOM_TOKEN)
MMMU-PRO	5	5.7702	2.6425	1.9849
	10	5.7697	2.6421	1.9845
	15	5.7704	2.6420	1.9871
	20	5.7690	2.6420	1.9841
	25	5.7712	2.6423	1.9895
AI2D	5	6.5889	3.3036	2.2283
	10	6.5889	3.3037	2.2282
	15	6.5891	3.3037	2.2284
	20	6.5891	3.3038	2.2281
	25	6.5894	3.3040	2.2285
TEXTVQA-VAL	5	4.6998	1.7963	1.9163
	10	4.6992	1.7963	1.9163
	15	4.7012	1.7976	1.9170
	20	4.6991	1.7974	1.9192
	25	4.7040	1.7904	1.9226

Specifically, the fluctuation in DSMD values across different binning granularities remains within a negligible range (± 0.005), regardless of whether the model is evaluated under the standard (*Orig*), vision-ablated (*Text*), or language-ablated (*Img*) settings. This stability confirms that the stratified quantile operator successfully marginalizes the token-count confounding factor without introducing sensitivity to discretization noise. Such robustness ensures that the metric captures the intrinsic behavioral preference of the MLLM rather than artifacts of the statistical estimation process. Consequently, we fix $Q = 10$ for all primary experiments as it provides a reliable balance between estimation resolution and sample density per bin.

D ABLATION STUDY: THE ROLE OF STRATIFICATION AND AGGREGATION

In this section, we investigate whether the effectiveness of DSMD stems from the choice of the log-odds payload or from our proposed stratified estimation and aggregation schemes. We define a **Payload-Only (PO)** baseline as a non-stratified counterpart. It adopts the same log-odds payload $a_{s,\ell}$ but computes a naive expectation over the entire sample set:

$$\mathrm{PO} = \sum_{\ell=0}^{L-1} w_\ell \, \bar{a}_\ell, \quad \text{where } \bar{a}_\ell = \mathbb{E}_s[a_{s,\ell}]. \tag{3}$$

In contrast, the full **DSMD** metric incorporates the stratified marginalization over Q quantile bins of the input balance ratio $z_s = \log O_s - \log T_s$:

$$\mathrm{DSMD} = \sum_{\ell=0}^{L-1} w_\ell \cdot \left(\frac{1}{Q} \sum_{q=1}^{Q} \mathbb{E}[a_{s,\ell} \mid z_s \in \mathcal{B}_q] \right). \tag{4}$$

We compare these two estimators under two aggregation schemes: *accuracy-weighted* ($w^{(\mathrm{acc})}$) and *uniform* ($w^{(\mathrm{uni})}$). The empirical results are reported in Tab. 4.

Stratification Corrects Distributional Bias (Analysis of Δ_{strat}). The *Impact of Stratification* columns in Panel A highlight the need for de-confounding when token distributions shift across resolutions. At the extreme low resolution (112×112), the model operates in a near-blind regime where visual evidence is severely limited, so stronger text dominance is expected. However, the naive estimator (PO) reports 3.56, while DSMD yields 4.01. The resulting gap ($\Delta_{\text{strat}} = +\mathbf{0.4516}$) indicates that **PO is affected by distributional bias** under extreme token sparsity. Our diagnostics show that the aggregate is mechanically coupled to a highly skewed token distribution (with $> 99\%$ of samples collapsing into low-information buckets), which distorts the global estimate and drives PO downward relative to bucket-conditioned behavior. DSMD mitigates this effect by stratifying on the token-count proxy and aggregating within strata before cross-strata aggregation, effectively marginalizing out the skew. Crucially, as resolution increases and the token distribution stabilizes, Δ_{strat} rapidly shrinks (from $+0.4516$ at 112px to $+0.0083$ at 896px). This trend provides evidence that DSMD behaves as an adaptive correction: it intervenes strongly only when token-count confounding is severe, thereby improving scale-consistency across the resolution spectrum.

Accuracy-Weighting Amplifies Diagnostic Sensitivity (Analysis of Δ_{weight}). The *Impact of Weighting* columns in Panel B demonstrate why all-phase aggregation must be weighted by predictive utility. Under the *Text* ablation (where inputs are gray images), the model is forced to rely on dataset shortcuts. Here, accuracy-weighted DSMD consistently yields higher dominance scores than the uniform baseline (e.g., $\Delta_{\text{weight}} = +\mathbf{0.4469}$ on MMMU-Pro and $+\mathbf{0.3826}$ on AI2D). This positive gap implies that the model's reasoning-intensive layers, characterized by high predictive accuracy, exhibit more severe text bias than the average layer. Uniform aggregation dilutes this critical signal by averaging it with noisy, low-utility layers. By prioritizing high-accuracy layers, DSMD achieves greater diagnostic sensitivity to intrinsic model tendencies.

E Statistical Quantification of Metric Drift

To rigorously quantify the impact of resolution scaling on metric stability, we perform a linear regression analysis between the layer-wise dominance curves at lower resolutions (r_1) and higher resolutions (r_2). This includes **comparisons to the highest resolution limit ($r_2 = 896^2$) as well as intermediate pairs**. We fit the model $m(r_2) \approx \alpha \cdot m(r_1) + \beta$, where α represents the scaling factor (drift) and R^2 measures **similarity under an affine transform** (i.e., how well a scaled-and-shifted curve matches the target).

MDI: Mechanical Inflation. As shown in Tab. 5, MDI exhibits high affine similarity ($R^2 > 0.8$) alongside *explosive drift*. The scaling coefficient α rises from 1.82 to 14.12 as the token-count imbalance grows. Even when comparing moderate resolutions ($448 \rightarrow 896$), MDI still implies a scaling factor of $\alpha \approx 3.41$. This combination of high correlation and exploding magnitude indicates a near-affine rescaling phenomenon. It suggests that the metric is mathematically dominated by the inflating token-count denominator (O_s), which mechanically amplifies the curve while preserving its relative shape. This scaling artifact risks masking subtle behavioral shifts under the guise of stability.

DSMD: Behavioral Sensitivity. In contrast, Tab. 6 confirms the scale-consistency of DSMD. The scaling factor α remains tightly bounded ($0.80 < \alpha < 0.91$) across all resolution pairs, confirming that DSMD effectively mitigates the token-count confounder. We note that the R^2 for DSMD is lower when comparing extreme resolutions ($112 \rightarrow 896$, $R^2 = 0.66$) but recovers when comparing sufficient resolutions ($448 \rightarrow 896$, $R^2 = 0.82$). The moderate decrease in R^2 across regimes is

Table 4: Unified Ablation Study of DSMD on Qwen2.5-VL-7B. This Tab. decouples the two core contributions of our method: (1) **Stratification Effect** (Δ_{strat}): Comparing PO against DSMD under fixed accuracy-weighting reveals that stratification corrects the underestimation bias at low resolutions. (2) **Weighting Effect** (Δ_{weight}): Comparing Uniform against Accuracy-weighted aggregation reveals that our weighting scheme amplifies sensitivity to intrinsic bias in high-utility layers.

Setting	Condition	Impact of Stratification			Impact of Weighting	
		PO (Acc)	**DSMD (Acc)**	Δ_{STRAT}	DSMD (Uni)	Δ_{WEIGHT}
Panel A: Resolution Sweep on MMMU-Pro (Probing Scale-Consistency)						
@ 112×112	Orig	3.5676	**4.0192**	+0.4516	4.0112	+0.0080
@ 224×224	Orig	3.1231	**3.3297**	+0.2066	3.2496	+0.0801
@ 336×336	Orig	2.8039	**2.9693**	+0.1654	2.9618	+0.0075
@ 448×448	Orig	2.7794	**2.8062**	+0.0268	2.7212	+0.0850
@ 672×672	Orig	2.5739	**2.5738**	−0.0001	2.5607	+0.0131
@ 896×896	Orig	2.4909	**2.4992**	+0.0083	2.4821	+0.0171
Panel B: Cross-Dataset Interventions (Probing Diagnostic Sensitivity)						
MMMU-Pro	*Orig*	2.6420	**2.6421**	+0.0001	2.5431	+0.0990
	Text	5.7685	**5.7697**	+0.0012	5.3228	+0.4469
	Img	1.9844	**1.9845**	+0.0001	2.0774	−0.0929
AI2D	*Orig*	3.3035	**3.3037**	+0.0002	3.2548	+0.0489
	Text	6.5887	**6.5889**	+0.0002	6.2063	+0.3826
	Img	2.2283	**2.2282**	−0.0001	2.3338	−0.1056
TextVQA-val	*Orig*	1.7960	**1.7963**	+0.0003	1.7914	+0.0049
	Text	4.6953	**4.6992**	+0.0039	4.7268	−0.0276
	Img	1.9155	**1.9163**	+0.0008	1.9214	−0.0051

compatible with resolution-dependent deviations beyond affine rescaling. As inputs transition from unrecognizable (112^2) to detailed (896^2), the model reorganizes its reasoning pathways. Crucially, this variance aligns with our findings in Section 4.2, suggesting a qualitative shift from broad prior-driven attention (at low resolution) to concentrated evidence-driven reasoning (at high resolution), a dynamic that inherently alters the curve shape. While MDI's high R^2 reflects mechanical scaling, DSMD's variance preserves these genuine behavioral adaptations.

Table 5: Shape consistency test for **MDI**. The explosion of α persists even in non-extreme pairs ($224 \to 896$, $\alpha \approx 7.85$), confirming that MDI drift is a systemic issue driven by token inflation.

Compare ($r_1 \to r_2$)	**corr**	$\boldsymbol{\alpha}$ (Scaling)	$\boldsymbol{\beta}$	$\mathbf{R^2}$ (Affine Fit)
112→224	0.9652	1.8223	4.5473	0.9315
112→448	0.9035	4.3054	16.0625	0.8163
112→896	0.8456	**14.1218**	29.8054	0.7151
Robustness Check				
224→896	0.9102	**7.8462**	12.4501	0.8285
448→896	0.9455	**3.4140**	5.1022	0.8940

Table 6: Shape consistency test for **DSMD**. α remains stable near $0.8 \sim 0.9$ across all pairs. R^2 is lower for the extreme pair ($112 \to 896$) but increases for closer regimes ($448 \to 896$), **consistent with reduced regime mismatch at sufficient resolutions**.

Compare ($r_1 \to r_2$)	**corr**	$\boldsymbol{\alpha}$ (Scaling)	$\boldsymbol{\beta}$	$\mathbf{R^2}$ (Affine Fit)
112→224	0.9683	0.9087	-0.3953	0.9377
112→448	0.8763	0.8050	-0.5076	0.7679
112→896	0.8169	**0.8261**	-0.8316	0.6673
Robustness Check				
224→896	0.8650	**0.8013**	-0.1250	0.7482
448→896	0.9082	**0.8993**	-0.0845	0.8248