PlaM: Training-Free Plateau-Guided Model Merging for Better Visual Grounding in MLLMs

Zijing Wang[1], Yongkang Liu[2†], Mingyang Wang[3,4], Ercong Nie[3,4], Deyuan Chen[1], Zhengjie Zhao[2], Shi Feng[1], Daling Wang[1†], Xiaocui Yang[1], Yifei Zhang[1], Hinrich Schütze[3,4]

[1]School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China;

[2]School of Computer and Communication Engineering, Northeastern University, Qinhuangdao 066004, China;

[3]CIS, LMU Munich, Germany; [4]Munich Center for Machine Learning (MCML), Germany

wzj1718@gmail.com

Abstract

Multimodal Large Language Models (MLLMs) rely on strong linguistic reasoning inherited from their base language models. However, multimodal instruction fine-tuning paradoxically degrades this text's reasoning capability, undermining multimodal performance. To address this issue, we propose a training-free framework to mitigate this degradation. Through layer-wise vision token masking, we reveal a common three-stage pattern in multimodal large language models: early-modal separation, mid-modal alignment, and late-modal degradation. By analyzing the behavior of MLLMs at different stages, we propose a plateau-guided model merging method that selectively injects base language model parameters into MLLMs. Experimental results based on five MLLMs across nine benchmarks demonstrate the effectiveness of our method. Attention-based analysis further reveals that merging shifts attention from diffuse, scattered patterns to focused localization on task-relevant visual regions. Our repository is on https://github.com/wzj1718/PlaM.

1 Introduction

Multimodal Large Language Models (MLLMs) have attracted widespread attention. Representative systems such as GPT-4V (Yang et al., 2023), Gemini (Team et al., 2024), along with a growing ecosystem of open-source models (e.g., LLaVA-style (Li et al., 2024b; Liu et al., 2024a) and Qwen-VL-style models (Team, 2025a,b)), demonstrate impressive multimodal instruction-following and visual reasoning capabilities.

Most MLLMs are built by extending a strong text-only LLM and then applying multimodal instruction fine-tuning. Although this pipeline yields strong multimodal understanding and generation (Alayrac et al., 2022), mounting evidence suggests that the resulting capabilities remain largely text-dominant: models rely disproportionately on textual signals while under-utilizing visual evidence (Wu et al., 2025b; Zheng et al., 2025; Zhao et al., 2025; Li et al., 2025; Chen et al., 2024a; Liu et al., 2024b). This pattern appears across multiple architectures and tasks. For instance, an analysis of VideoLLaMA-7B shows that output tokens attend to text tokens 157 times more than to visual tokens (Wu et al., 2025b), and broader studies report that MLLMs often answer visual questions primarily using textual knowledge, with visual information playing only a secondary role (Wu et al., 2025a; Liu et al., 2025). Importantly, multimodal instruction fine-tuning can also degrade the base model's original text reasoning ability (Zhang et al., 2024; Lu et al., 2024; Li et al., 2024c). Recent studies further show that even strong backbones such as Llama3, Mistral, and Vicuna-based models may exhibit degraded performance on most text-only reasoning benchmarks after visual instruction tuning (Ratzlaff et al., 2025; Yu and Ananiadou, 2025). This degradation is particularly concerning because strong text reasoning is the backbone for composing and verifying multimodal inferences; when it weakens, models become less able to correctly interpret and leverage visual signals, further exacerbating poor visual grounding. As illustrated in Figure 1, failure cases exhibit scattered

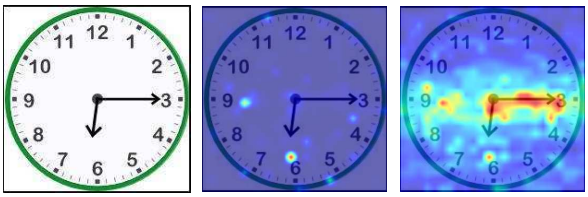

Figure 1: Our proposed PlaM improves visual grounding and prediction. Given the clock image (left) and the question "It is (_) past six.", the original MLLM without PlaM attends diffusely and answers "half" (middle). After applying PlaM model merging, attention concentrates on the clock hands and the model correctly answers "quarter" (right).

[†]Corresponding Authors.

Findings of the Association for Computational Linguistics: ACL 2026, pages 21019–21035

and semantically irrelevant attention over the image, whereas correct predictions require focusing on task-relevant regions. Overall, multimodal instruction tuning weakens text processing and fails to promote effective visual grounding, motivating recent efforts to address this trade-off (Lu et al., 2024).

Existing attempts to address language-reasoning degradation in MLLMs are largely training-based. Typical strategies include introducing auxiliary components to compensate attention shifts induced by visual token expansion (Zhang et al., 2024), replaying or interleaving text-only data during training (McKinzie et al., 2024; Li et al., 2024d), and using preference alignment objectives to jointly optimize visual instruction quality and language instruction-following (Li et al., 2024c). However, these approaches often require additional computational resources, carefully curated data, or architectural modifications. In contrast, training-free solutions remain comparatively underexplored. Ratzlaff et al. (2025) show that the choice of the base language model and its inherent capability significantly influence the extent of text-task degradation after multimodal fine-tuning. Meanwhile, Zhang et al. (2025b) suggest that modality preference can be characterized by a direction in latent representations and steered via representation engineering, enabling control over modality preference without additional fine-tuning.

In this paper, we propose Plateau-guided model Merging (**PlaM**), a general and efficient training-free method to enhance the text reasoning capability of MLLMs. We begin with a layer-function analysis and show that the utilization of visual information in MLLMs follows a diminishing-returns pattern across decoder depth. Specifically, through layer-wise vision token masking, we reveal a consistent three-stage behavior: (1) in early layers, models struggle to effectively exploit visual evidence and exhibit insufficient visual grounding due to modal separation; (2) in middle layers, guided by textual context, MLLMs performs modal alignment and progressively attend to semantically critical visual features, leading to rapid performance gains; and (3) in late layers, modal alignment tends to stabilize, visual features are absorbed by textual information, and additional visual access yields little further improvement, resulting in a performance plateau. Importantly, we find that this late-layer plateau is closely linked to weakened language capability after multimodal tuning: degraded text reasoning limits the model's ability to guide attention toward the correct visual features, resulting in a performance bottleneck. To recover the compromised text ability, we introduce a plateau-guided model merging strategy that selectively injects base language model parameters into these underutilized late layers via linear interpolation. Extensive experiments across five representative MLLMs and nine benchmarks demonstrate consistent improvements, validating the effectiveness and generality of our training-free method.

Our contributions can be summarized as follows:

- We identify a three-stage pattern of visual token utilization and demonstrate that late model layers exhibit diminishing returns with respect to additional visual access;
- We propose plateau-guided model merging (PlaM), a simple training-free approach that restores base LM capabilities in late layers and achieves consistent improvements across multiple models and benchmarks;
- We provide mechanistic insights through attention analysis, showing that performance gains arise from enhanced visual grounding that transforms diffuse attention into focused, task-relevant localization of visual evidence.

2 Related Work

MLLMs. MLLMs have rapidly emerged as a promising paradigm for integrating language and vision within a unified system (Yin et al., 2024; Han et al., 2025). Rather than processing each modality in isolation, these models are typically designed to support cross-modal interaction by mapping modality-specific inputs into a shared space that a language-centric reasoning core can operate on (Radford et al., 2021; Lu et al., 2019). A growing body of work builds MLLMs by leveraging pretrained large language models as the primary reasoning engine, while attaching dedicated encoders (e.g., vision encoders) and lightweight alignment components to transform non-text signals into language-compatible representations (Alayrac et al., 2022; Li et al., 2023a; Tsimpoukelli et al., 2021; Eichenberg et al., 2022; Zhu et al.). Recent progress further shows that instruction-style multimodal fine-tuning, which often uses large-scale, automatically constructed or weakly supervised multimodal instruction data, can substantially improve multimodal instruction following and reasoning (Liu et al., 2023; Ye et al., 2023; Dai et al.,

2023).

Mitigating Language Reasoning Degradation in MLLMs. Multimodal instruction tuning often degrades a pretrained language model's original text-only reasoning capabilities (Yu and Ananiadou, 2025; Li et al., 2024c). **Most existing solutions are training-based**: Wu et al. (2024) manipulate attention responses via inference-time latent variable optimization to improve visual referring behavior. Zhang et al. (2024) introduce auxiliary architectural components to compensate for attention shifts induced by visual token expansion, while others preserve language reasoning by interleaving text-only data or using preference-based alignment (McKinzie et al., 2024; Li et al., 2024d,c). However, these methods require additional training, curated datasets, or architectural modifications. **Training-free approaches remain relatively underexplored**. Ratzlaff et al. (2025) show that text degradation depends heavily on base LLM choice, Wu et al. (2025b) compress non-text tokens to reshape attention allocation, and Zhang et al. (2025b) steer modality preference via latent representation offsets, though this requires externally specifying modality preference. Our work proposes a training-free strategy that explicitly restores language model functionality in underutilized late layers, alleviating language reasoning degradation while improving visual grounding.

3 Method

In this section, we present **PlaM** (Plateau-guided model Merging), a training-free approach that improves visual grounding by restoring late-layer language reasoning in MLLMs. PlaM consists of two components used throughout the paper. First, we adopt a layer-wise vision token masking procedure to characterize how visual information is utilized across decoder depth and to identify where performance plateaus. Second, guided by this layer-wise characterization, we selectively merge parameters between the base language model and the fine-tuned MLLM within a chosen layer range, while keeping the vision encoder and projector fixed.

Notation Given an image I and a text prompt T, an MLLM encodes the two modalities into a unified token sequence. The image is first processed by a vision encoder E_{vis}, which maps the raw image into a set of high-dimensional visual features. These features are then projected into the language model's embedding space via a projector P, yielding visual tokens $X_{\text{vis}}^{(0)} = P(E_{\text{vis}}(I)) \in \mathbb{R}^{N_{\text{vis}} \times d}$. The text prompt $T = (t_1, ..., t_{N_{\text{txt}}})$ is embedded through the language model's token embedding layer, producing textual tokens $X_{\text{txt}}^{(0)} \in \mathbb{R}^{N_{\text{txt}} \times d}$. The input to the language model is formed by concatenating visual and textual tokens: $X^{(0)} = [X_{\text{vis}}^{(0)}; X_{\text{txt}}^{(0)}] \in \mathbb{R}^{N \times d}$, where $N = N_{\text{vis}} + N_{\text{txt}}$ is the total sequence length and d is the hidden dimension of the language model.

The language model decoder consists of L Transformer layers, and the hidden states evolve through the network as $X^{(l)} = \Phi^{(l)}(X^{(l-1)})$, $l = \{1, ..., L\}$. During inference, the model first encodes the entire multimodal prompt to construct the KV cache, and then generates output tokens autoregressively conditioned on the cached representations.

Vision Token Masking To quantify how MLLMs rely on visual information across decoder depth, we apply a depth-controlled masking strategy on visual tokens (Shi et al., 2025). Let $\mathcal{V} \subset \{N_1, N_1 + 1 \ldots, N_1 + N_{\text{vis}} - 1\}$ denote the index set of visual tokens in the input sequence. For a selected layer k, visual tokens are processed normally in all layers $l < k$. For layers $l \geq k$, we block access to visual tokens by removing their positions from the attention keys and values: $\text{Attn}^{(l)}(Q, K, V) \Rightarrow \text{Attn}^{(l)}\big(Q, K_{\neg\mathcal{V}}, V_{\neg\mathcal{V}}\big)$. This intervention preserves the textual pathway while preventing later layers from attending to visual tokens. Sweeping k from shallow to deep layers yields a layer-wise profile characterizing where, and to what extent, visual information contributes to the decoder's predictions.

Model Merging Across Modalities We aim to enhance MLLMs' performance by selectively incorporating base language model parameters into specific layers of the fine-tuned model. Let $\mathbf{W}_{\text{vlm}}$ denote the parameters of the fine-tuned MLLM and $\mathbf{W}_{\text{lm}}$ denote those of the corresponding base language model (sharing the same backbone architecture). We first identify a target layer set $\mathcal{L} = \{l_m, l_{m+1}, \ldots, l_n\}$, typically in the middle-to-late decoder layers. For each layer $l \in \mathcal{L}$, we construct merged parameters:

$$\mathbf{W}_{\text{merged}}^{(l)} = \lambda_1 \mathbf{W}_{\text{lm}}^{(l)} + \lambda_2 \mathbf{W}_{\text{vlm}}^{(l)}, \quad \lambda_i \in [0, 1.5].$$

For layers outside $\mathcal{L}$, we preserve the original MLLM parameters:

$$\mathbf{W}_{\text{merged}}^{(l)} = \mathbf{W}_{\text{vlm}}^{(l)}, \quad \forall l \notin \mathcal{L}.$$

The vision encoder and projector remain fixed, and only the selected backbone layers undergo merging. The hyperparameter λ_i controls the degree to which we reintroduce the base language model's capabilities. By varying λ_i and the layer range $\mathcal{L}$, we obtain a family of merged models that balance multimodal alignment with the general reasoning capabilities of the base model.

4 Experiment settings

In this section, we describe the models, evaluation benchmarks, and hyperparameters used in our experiments*.

Vision-Language Models. To verify the generalization ability of model merging across different model architectures and scales, we conduct experiments on 5 representative MLLMs: **LLaVA-v1.5-7B (LLaMA-7B)** (Liu et al., 2024a), **Qwen2.5-VL-3B-Instruct (Qwen2.5-3B)** (Team, 2024, 2025a), **Qwen2.5-VL-7B-Instruct (Qwen2.5-7B-Instruct)**, **Qwen3-VL-4B-Instruct (Qwen3-4B)** (Team, 2025b) and **Qwen3-VL-8B-Instruct (Qwen3-8B)**. These models span a range of scales from 3B to 8B parameters and encompass both LLaMA-based and Qwen-based architectures, enabling us to assess the robustness of our findings across diverse model families.

Evaluation Benchmarks. We evaluate model performance across a comprehensive suite of vision-language benchmarks that assess diverse multimodal capabilities: **MMStar** (Chen et al., 2024b), **MMMU** (Yue et al., 2024), **MME** (Fu et al., 2025), **MMBench-EN/CN** (Liu et al., 2024c), **GQA** (Hudson and Manning, 2019), **RealWorldQA**, **SEED-Bench-2-Plus** (Li et al., 2024a), and **POPE** (Li et al., 2023b). These benchmarks assess reasoning, perception, multilingual understanding, visual grounding, and hallucination.

Hyperparameters. Throughout our analysis and experiments, we employ a simple merging strategy for all models. We adopt a unified hyperparameter configuration across all models to ensure fair comparison. The merging coefficient λ_i is determined through grid search with a step size of 0.1, ranging from 0.0 to 1.5. For the target layer set $\mathcal{L}$, we systematically evaluate different starting points l_m based on the masking analysis in Section 5.1, typically selecting the transition point where vision information transitions from beneficial to redundant. Unless otherwise specified, we report results using the optimal hyperparameters identified for each model-task combination.

*Please refer to Appendix A for a detailed description.

5 Visual Degeneration

5.1 Visual Information Absorption

Following Section 3, we quantify the contribution of visual tokens at different decoder depths by sweeping the cut layer k in the depth-controlled masking intervention. Specifically, for each model and each benchmark, we evaluate the model under a series of settings $k \in \{1, ..., L+1\}$. When $k = L+1$, no masking is applied (the original model). Figure 2 illustrates the trend of model performance as k varies.

We observe that different MLLMs display a consistent three-stage trend across various tasks. Based on the points of abrupt change in model performance, the model's behavior can be divided into three stages: (i) early: feature separation; (ii) middle: feature alignment and fusion; (iii) late: feature degradation and absorption. The input layer of the MLLMs receives information from different modalities and gradually aligns and fuses the information. In the early stages (low k), the model primarily focuses on intramodal modeling, during which it is still learning visual and textual semantic representations. The textual representations largely rely on the prior knowledge of the LLM, while cross-modal interactions remain shallow. Consequently, the alignment between visual information and linguistic concepts is weak, and although the incorporation of visual information is beneficial, the resulting performance gains are limited.

As models transition into the middle stage (middle k), a pronounced and rapid performance increase is observed across almost all benchmarks, including MMMU, GQA, and MMBench in both English and Chinese. This stage is consistent with progressively stronger cross-modal interaction in the middle layers. To further support this interpretation, we measure the cosine similarity between vision-token and text-token hidden states across layers and observe a clear upward trend from early to middle/late layers (Appendix Figure 5), indicating increasing representational convergence across

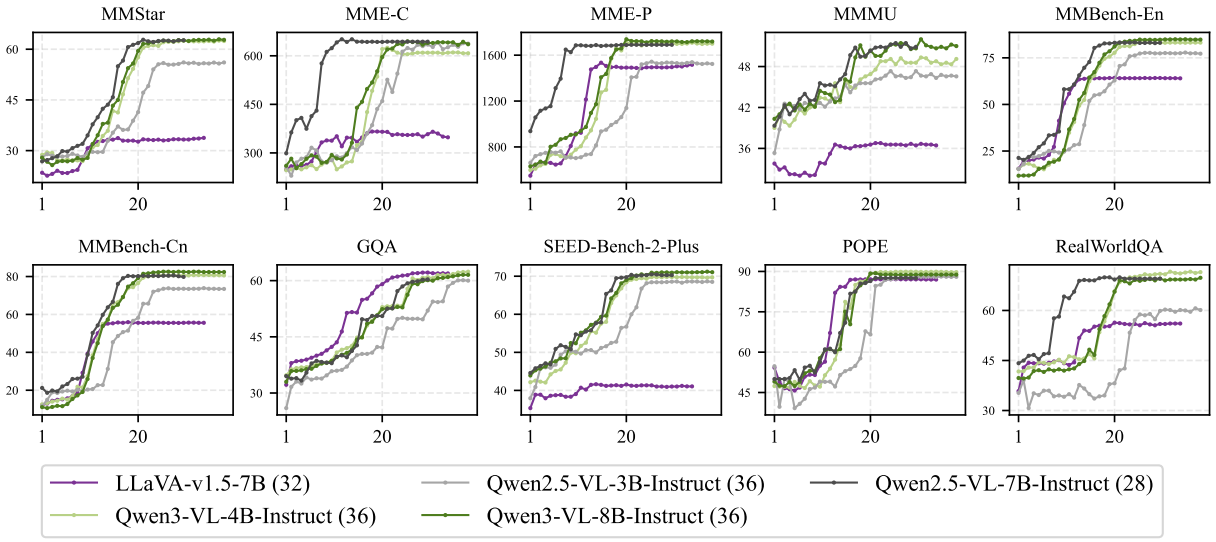

Figure 2: Performance versus cut layer (k) under depth-controlled vision token masking. Vision tokens are removed from layer k onward ($k = L+1$ indicates no masking), and each curve reports the official metric on each benchmark. See Tables 3 to 7 for the corresponding numerical results (Appendix A).

modalities. This trend suggests that visual information is increasingly integrated with the textual reasoning pathway, enabling the model to move beyond surface-level description toward more image-grounded inference. Taken together, these findings suggest that the middle stage is the key phase in which cross-modal alignment strengthens and multimodal reasoning begins to emerge.

In the later stages (high k), model performance gradually reaches a plateau, in some cases, may even slightly decline. A plausible explanation is that deeper multimodal fusion increasingly favors abstract linguistic representations, with visual information being progressively compressed or absorbed into the linguistic space. Consequently, in the later layers, performance does not significantly degrade even when access to visual tokens is reduced. While such deep fusion benefits abstract reasoning, it can also erode the language model's inherent distributed knowledge, thereby disrupting the robust reasoning patterns acquired during pretraining (Amariucai and Warstadt, 2023; Zhai et al., 2023). From a representation learning perspective, this suggests that visual signals have already been sufficiently integrated by the later stage, and that **further performance bottlenecks stem from the degradation of linguistic representations rather than from limitations in visual modeling**.

5.2 Plateau-Guided Late-Layer Merging

Motivated by the plateau behavior observed in Section 5.1, we propose PlaM that injects base language model parameters into late decoder layers starting from the plateau onset. Intuitively, since these layers contribute little additional benefit from continued visual access, we use them to recover general-purpose language model capabilities while preserving multimodal alignment in earlier layers.

For each model, we identify the plateau onset layer $k^\star$ from the performance–k curves (Figure 2) as the elbow point where the rapid-gain regime transitions into a stable plateau. The merge start layer is treated as a lightweight hyperparameter k_0, whose value is determined via a nearest-neighbor search in the vicinity of $k^\star$. All decoder layers from k_0 to the final layer are then merged, i.e., $\mathcal{L}(k_0) = \{k_0, \ldots, L\}$. Considering that the attention mechanism fundamentally governs how text tokens attend to and integrate visual features in the Transformer decoder (Wu et al., 2024; Kang et al.), we merge only the self-attention projections (Q/K/V/O) of the fine-tuned MLLM with the corresponding projections from the base language model described in Section 3.

Results. Overall Results and Comparison with Merging Baselines. Table 1 reports the performance comparison between our plateau-guided late-layer merging (PlaM) and three alternative merging strategies: early-layer merging, mid-layer merging, and full-layer merging across all decoder layers. PlaM consistently outperforms the base model across all 5 evaluated architectures and 9 benchmarks, demonstrating its effectiveness as a

Model		MMStar	MME_C	MME_P	MMMU	MMB-En	MMB-Cn	GQA	Seed_2_plus	RealWorldQA	POPE
LLaVA_v1.5_7B	Base Model	33.77	348.2143	1516.0553	36.44	64.0344	55.5842	61.93	41.06	56.08	86.98
	Early-layer Merge	35.23	306.7857	1481.7036	31.44	62.8007	45.7045	60.66	40.71	52.29	86.79
	Δ compare to base	+1.46	-41.43	-34.35	-5.00	-1.23	-9.88	-1.27	-0.35	-3.79	-0.19
	Mid-layer Merge	34.49	274.6429	1478.7404	34.67	60.3952	47.4227	61.66	39.92	50.98	85.34
	Δ compare to base	+0.72	-73.57	-37.31	-1.77	-3.64	-8.16	-0.27	-1.14	-5.10	-1.64
	Full-layer Merge	34.68	298.5714	1396.5884	32.89	56.0997	41.1512	60.08	37.42	48.50	85.29
	Δ compare to base	+0.91	-49.64	-119.47	-3.55	-7.93	-14.43	-1.85	-3.64	-7.58	-1.69
	PlaM	**35.29**	**363.2143**	**1522.4614**	**37.00**	**65.0344**	**55.9278**	**62.24**	**43.00**	**56.34**	**87.59**
	Δ compare to base	+1.52	+15.00	+6.41	+0.56	+1.00	+0.34	+0.31	+1.94	+0.26	+0.61
Qwen2.5-VL-3B-Instruct	Base Model	56.09	637.5000	1523.4981	46.56	77.2337	73.4536	60.00	68.51	60.13	87.99
	Early-layer Merge	55.71	619.6429	1541.1841	46.89	78.0069	73.2818	59.79	68.56	59.48	88.05
	Δ compare to base	-0.38	-17.86	+17.69	+0.33	+0.77	-0.17	-0.21	+0.05	-0.65	+0.06
	Mid-layer Merge	56.31	635.3571	1523.6149	46.22	77.5773	73.9691	60.14	68.51	59.35	87.86
	Δ compare to base	+0.22	-2.14	+0.12	-0.34	+0.34	+0.52	+0.14	+0.00	-0.78	-0.13
	Full-layer Merge	55.75	633.9286	1535.4336	46.78	78.3505	73.2818	59.76	68.12	59.35	88.15
	Δ compare to base	-0.34	-3.57	+11.94	+0.22	+1.12	-0.17	-0.24	-0.39	-0.78	+0.16
	PlaM	**57.89**	**640.0000**	**1538.2974**	**47.78**	**79.0378**	**74.9141**	**60.14**	**69.30**	**64.71**	**88.21**
	Δ compare to base	+1.80	+2.50	+14.80	+1.22	+1.80	+1.46	+0.14	+0.79	+4.58	+0.22
Qwen2.5-VL-7B-Instruct	Base Model	62.65	643.5714	1690.3052	50.89	83.0756	79.8110	60.35	70.27	69.54	87.51
	Early-layer Merge	62.51	643.5714	1693.5381	51.56	83.2474	80.0687	60.64	70.49	68.76	87.42
	Δ compare to base	-0.14	+0.00	+3.23	+0.67	+0.17	+0.26	+0.29	+0.22	-0.78	-0.09
	Mid-layer Merge	62.10	641.4286	1700.4490	50.00	83.0756	80.1546	60.33	70.36	68.89	87.41
	Δ compare to base	-0.55	-2.14	+10.14	-0.89	+0.00	+0.34	-0.02	+0.09	-0.65	-0.10
	Full-layer Merge	61.51	651.0714	1688.7383	51.22	83.2474	80.4124	60.42	70.14	69.54	87.33
	Δ compare to base	-1.14	+7.50	-1.57	+0.33	+0.17	+0.60	+0.07	-0.13	+0.00	-0.18
	PlaM	**63.27**	**652.5000**	**1709.5045**	**52.11**	**83.8488**	**81.2715**	**60.73**	**70.93**	**70.07**	**88.48**
	Δ compare to base	+0.62	+8.93	+19.20	+1.22	+0.77	+1.46	+0.38	+0.66	+0.53	+0.97
Qwen3-VL-4B-Instruct	Base Model	62.43	607.8571	1699.8849	49.11	83.1615	80.5842	62.39	69.70	71.50	89.75
	Early-layer Merge	62.72	619.2857	1700.1201	48.56	83.3333	80.4983	62.21	69.78	71.50	89.73
	Δ compare to base	+0.29	+11.43	+0.24	-0.55	+0.17	-0.09	-0.18	+0.08	+0.00	-0.02
	Mid-layer Merge	62.26	620.3571	1703.4821	48.22	83.5911	80.6701	62.23	69.74	71.37	89.64
	Δ compare to base	-0.17	+12.50	+3.60	-0.89	+0.43	+0.09	-0.16	+0.04	-0.13	-0.11
	Full-layer Merge	61.24	624.2857	1688.3906	49.78	83.6770	80.2405	62.23	69.74	70.98	89.79
	Δ compare to base	-1.19	+16.43	-11.49	+0.67	+0.52	-0.34	-0.16	+0.04	-0.52	+0.04
	PlaM	**63.04**	**630.7143**	**1710.8023**	**50.33**	**84.2784**	**81.2715**	**62.43**	**70.27**	**72.29**	**89.94**
	Δ compare to base	+0.61	+22.86	+10.92	+1.22	+1.12	+0.69	+0.04	+0.57	+0.79	+0.19
Qwen3-VL-8B-Instruct	Base Model	62.78	635.7143	1718.7803	51.00	84.7938	82.3883	61.54	71.06	69.80	88.81
	Early-layer Merge	62.91	642.5000	1716.6741	51.67	84.7938	82.3883	61.54	71.06	69.15	88.69
	Δ compare to base	+0.13	+6.79	-2.11	+0.67	+0.00	+0.00	+0.00	+0.00	-0.65	-0.12
	Mid-layer Merge	63.40	626.0714	1719.6741	51.78	84.7079	82.3024	61.73	71.01	69.02	88.78
	Δ compare to base	+0.62	-9.64	+0.89	+0.78	-0.09	-0.09	+0.19	-0.05	-0.78	-0.03
	Full-layer Merge	63.48	635.0000	1707.6741	51.11	84.7079	82.3883	61.59	70.88	69.15	88.82
	Δ compare to base	+0.70	-0.71	-11.11	+0.11	-0.09	+0.00	+0.05	-0.18	-0.65	+0.01
	PlaM	**64.54**	**649.2857**	**1727.8121**	**53.11**	**85.1375**	**82.6460**	**61.77**	**71.54**	**71.01**	**89.37**
	Δ compare to base	+1.76	+13.57	+9.03	+2.11	+0.34	+0.26	+0.23	+0.48	+1.21	+0.56

Table 1: Performance comparison of plateau-guided late-layer merging (PlaM) with alternative merging strategies (early-, mid-, and full-layer) across 5 MLLM backbones and 9 benchmarks. We report absolute scores and the change relative to the original fine-tuned MLLM ("Base Model") for each strategy. PlaM consistently achieves the best overall performance.

training-free enhancement method. For example, on LLaVA-v1.5-7B, PlaM improves over the base MLLM on MMStar (35.29 vs. 33.77), MMMU (37.00 vs. 36.44), and RealWorldQA (56.34 vs. 56.08), while maintaining or improving performance on most other benchmarks. Similar trends are observed across the Qwen model families. In contrast, the three alternative merging strategies (Early-layer, Mid-layer and Full-layer) fail to achieve the best overall performance. These observations suggest that the effectiveness of PlaM critically depends on where the merging is applied. The early and middle layers are primarily responsible for intra-modal feature learning and cross-modal alignment. In the early and middle stages, multimodal alignment remains incomplete. Performing parameter merging at this stage causes the model to become overly dependent on the pre-trained LLM, thereby inhibiting necessary visual adaptation, disrupting cross-modal alignment, and ultimately degrading overall model performance. In the late stage, as discussed above, performance bottlenecks stem from the degradation of linguistic representations rather than from limitations in visual modeling. Merging the original LLM weights at this stage effectively restores part of the text-only representational manifold that may have been distorted during multimodal fine-tuning. By reintroducing the original weights, the model regains access to high-quality linguistic abstractions and stable reasoning trajectories, which can compensate for the noise and over-compression introduced

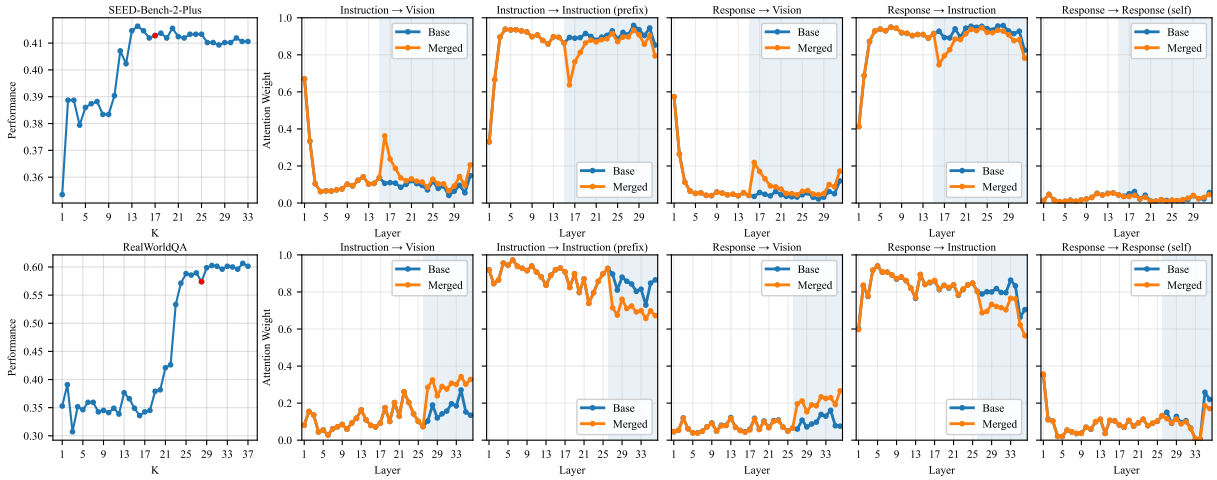

Figure 3: Overall comparison of performance and attention mass before and after merging. **Left:** Layer-wise vision token masking results, where the model performance is measured by progressively removing vision tokens from layer k to the last layer. The red marker indicates the selected merge start layer k_0, beyond which visual information yields diminishing performance gains and merging is applied. **Right:** Layer-wise attention mass profiles comparing the original model (Base) and the merged model (Merged). The shaded region denotes the merged layers $\{k_0, L\}$. Results are shown for LLaVA-v1.5-7B on SEED-Bench-2-Plus (top row) and Qwen2.5-VL-3B-Instruct on RealWorldQA (bottom row).

by aggressive cross-modal fusion. Importantly, this does not eliminate multimodal capability; instead, it rebalances the contribution of visual and linguistic representations.

Task-type analysis. The gains of PlaM exhibit a clear task-dependent pattern in Table 1. PlaM delivers the most pronounced improvements on MMStar, MME, and MMMU, where success strongly depends on late-stage semantic decision making and visual evidence integration. Concretely, PlaM consistently boosts MMStar by $+0.61 \sim +1.80$, and yields especially large gains on MME—MME_C: $+2.50 \sim +22.86$ and MME_P: $+6.41 \sim +19.20$ across all five backbones—together with consistent improvements on MMMU ($+0.56 \sim +2.11$). This aligns with the design of PlaM: by merging only plateau-phase late-layer attention projections, it mainly repairs the degraded semantic decision structure that governs which visual cues are retrieved and trusted at the final reasoning stage, thus benefiting benchmarks that are most sensitive to decision-level grounding. In contrast, PlaM yields moderate but consistent gains on the remaining benchmarks, where performance is often limited by strong baselines and early-to-mid fusion/representation quality rather than late-stage decisions, making late-layer merging less impactful. Finally, gains are minimal on GQA ($+0.04 \sim +0.38$) and POPE ($+0.19 \sim +0.97$), where the dominant limitations likely lie beyond late-layer semantic decisions: GQA relies heavily on early-to-mid compositional grounding and spatial/relational modeling, while POPE is closer to evidence-verification and calibration with already high baseline accuracy. Consequently, restoring late-layer semantic decision structure alone provides only marginal benefits on these benchmarks.

6 Attention-based Mechanistic Analysis

Measuring Vision Token Attention We partition the input sequence by token type as $X = [T_{\text{pre}}; T_{\text{vis}}; T_{\text{ins}}]$, where T_{pre} denotes prefix text tokens (e.g., system prompts), T_{vis} denotes vision tokens and T_{ins} denotes instruction text tokens. During generation, the single-step output token is denoted as T_{res}. At layer l, let $\alpha_{ij}^{(l)}$ denote the attention weight from query token i to key token j, averaged over all attention heads. For a target token set T_{tgt} and a source token set T_{src}, we define the attention mass as:

$$\text{Mass}^{(l)}(T_{\text{tgt}} \to T_{\text{src}}) = \frac{1}{|T_{\text{tgt}}|} \sum_{i \in T_{\text{tgt}}} \sum_{j \in T_{\text{src}}} \alpha_{ij}^{(l)}.$$

We compute this quantity in the prefill stage with $T_{\text{tgt}} = T_{\text{ins}}$, and in the decoding stage with $T_{\text{tgt}} = T_{\text{res}}$. The source set is chosen from $T_{\text{src}} \in \{T_{\text{vis}}, T_{\text{pre}} \cup T_{\text{ins}}, T_{\text{res}}\}$, yielding layer-wise curves that serve as an attention-based indicator of how different token types are emphasized across lay-

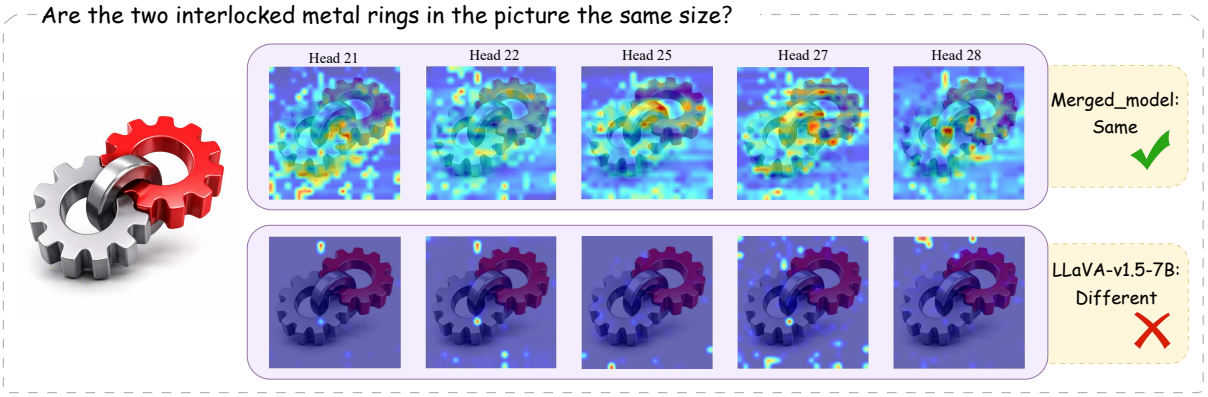

Figure 4: Attention heatmaps for LLaVA-v1.5-7B (bottom) and its PlaM-merged model (top). Additional case studies for other backbones are provided in Appendix Figures 6 to 10.

ers. Intuitively, a large $\mathrm{Mass}^{(l)}(\cdot \to T_{\text{vis}})$ indicates stronger reliance on visual tokens at depth l.

Layer-wise Attention Comparison Figure 3 compares the layer-wise attention mass profiles between the original models and their merged counterparts. A consistent pattern emerges across both architectures: merging substantially increases the attention mass allocated to vision tokens. We attribute this shift to the injection of base language model capacity into the late layers, which restores stronger semantic decision-making and, in turn, encourages more active retrieval and utilization of visual evidence.

For Qwen2.5-VL-3B-Instruct (merged layers 27-36), the merged model exhibits a clear upward trend in instruction-to-vision attention mass ($\mathrm{Mass}^{(l)}(T_{\text{ins}} \to T_{\text{vis}})$) in late layers, rising from near 0.1 to 0.3, which suggests that merging strengthens visual evidence integration during prompt encoding itself. Meanwhile, the response-to-vision attention mass ($\mathrm{Mass}^{(l)}(T_{\text{res}} \to T_{\text{vis}})$) shows a modest increase in attention mass toward vision tokens, which further suggests that the merged model performs explicit visual verification or retrieval during output generation and ensures response accuracy and grounding. Similar late-layer increases in vision-directed attention are also observed for LLaVA.

It is important to note that higher vision-token attention mass reflects increased reliance on visual information rather than a guarantee of improved reasoning quality. However, the consistent shift toward greater utilization of vision tokens across both architectures aligns well with the performance improvements reported in Section 5.2. This mechanistic evidence supports our hypothesis that plateau-guided merging guides late layers to explore broader visual evidence by recovering the textual capabilities, thereby improving multimodal grounding.

Case Study To further investigate whether the increased attention mass corresponds to more effective visual grounding, we visualize attention heatmaps from several attention heads for LLaVA-v1.5-7B and its PlaM counterpart. Figure 4 presents the attention patterns for the question "Are the two interlocked metal rings in the picture the same size?", which requires fine-grained visual comparison of the two rings. The contrast is striking: the original LLaVA model (bottom row) exhibits predominantly weak and diffuse attention, with most heads failing to consistently focus on the ring boundaries, and consequently produces an incorrect answer ("Different"). In contrast, the merged model (top row) demonstrates more structured and task-relevant attention patterns. Multiple heads allocate concentrated attention to both rings, particularly along their contours and the interlocking region, thereby enabling accurate size comparison and the correct answer ("Same"). This example suggests that the benefits of PlaM extend beyond higher aggregate attention to vision tokens, manifesting as enhanced spatial localization on decision-critical visual evidence during decoding.

7 Conclusion

In this paper, we propose **PlaM**, a training-free plateau-guided late-layer merging method that injects base LM attention projections into plateau-phase layers to restore degraded semantic decision structure while preserving earlier multimodal alignment. Extensive experiments across five back-

bones and nine benchmarks show consistent improvements over the original models and alternative merging schemes. Analyses of attention mass and heatmaps further indicate that PlaM makes late-layer visual attention more focused on semantically relevant regions, helping mitigate scattered attention and improving visual grounding. We hope our findings provide actionable insights into layer-wise multimodal behavior and inspire future work on training-free interventions in MLLMs.

Limitations

This work has the following limitations. We introduce a hyperparameter k_0, which determines the merging position of the model. The optimal k_0 configuration varies for different datasets, and determining the optimal k_0 configuration requires more experiments and costs. We use a simple model merging strategy, more complex model merging methods may yield better results, although this does not conflict with our work.

Acknowledgments

The work is supported by the National Natural Science Foundation of China (No. 62272092, No. 62172086), National Science Foundation for Young Scientists of China (No. 62502081), and the Fundamental Research Funds for the Central Universities under Grants (N2523011, N25XQD004).

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Theodor Amariucai and Alexander Scott Warstadt. 2023. Acquiring linguistic knowledge from multimodal input. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 128–141.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024a. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pages 19–35. Springer.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and 1 others. 2024b. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267.

Constantin Eichenberg, Sidney Black, Samuel Weinbach, Letitia Parcalabescu, and Anette Frank. 2022. Magma–multimodal augmentation of generative models through adapter-based finetuning. In *Findings of the association for computational linguistics: EMNLP 2022*, pages 2416–2428.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and 1 others. 2025. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Longzhen Han, Awes Mubarak, Almas Baimagambetov, Nikolaos Polatidis, and Thar Baker. 2025. A survey of generative categories and techniques in multimodal generative models. *Preprint*, arXiv:2506.10016.

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. See what you are told: Visual attention sink in large multimodal models. In *The Thirteenth International Conference on Learning Representations*.

Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. 2024a. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024b. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Mingxiao Li, Na Su, Fang Qu, Zhizhou Zhong, Ziyang Chen, Yuan Li, Zhaopeng Tu, and Xiaolong Li. 2025. Vista: Enhancing vision-text alignment in mllms via cross-modal mutual information maximization. *arXiv preprint arXiv:2505.10917*.

Shengzhi Li, Rongyu Lin, and Shichao Pei. 2024c. Multi-modal preference alignment remedies degradation of visual instruction tuning on language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14188–14200.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. 2024d. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305.

Chenxi Liu, Tianyi Xiong, Yanshuo Chen, Ruibo Chen, Yihan Wu, Junfeng Guo, Tianyi Zhou, and Heng Huang. 2025. Modality-balancing preference optimization of large multimodal models by adversarial negative mining. *arXiv preprint arXiv:2506.08022*.

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

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

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024c. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, and 1 others. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Anton Belyi, and 1 others. 2024. Mm1: methods, analysis and insights from multimodal llm pre-training. In *European Conference on Computer Vision*, pages 304–323. Springer.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Neale Ratzlaff, Man Luo, Xin Su, Vasudev Lal, and Phillip Howard. 2025. Training-free mitigation of language reasoning degradation after multimodal instruction tuning. In *Proceedings of the AAAI Symposium Series*, volume 5, pages 384–388.

Cheng Shi, Yizhou Yu, and Sibei Yang. 2025. Vision function layer in multimodal llms. *arXiv preprint arXiv:2509.24791*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Qwen Team. 2025a. Qwen2.5-vl.

Qwen Team. 2025b. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212.

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

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

Mingrui Wu, Xinyue Cai, Jiayi Ji, Jiale Li, Oucheng Huang, Gen Luo, Hao Fei, Guannan Jiang, Xiaoshuai Sun, and Rongrong Ji. 2024. Controlmllm: Training-free visual prompt learning for multimodal large language models. *Advances in Neural Information Processing Systems*, 37:45206–45234.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, and 1 others. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12).

Zeping Yu and Sophia Ananiadou. 2025. Locate-then-merge: Neuron-level parameter fusion for mitigating catastrophic forgetting in multimodal llms. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 7065–7078.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. 2023. Investigating the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313*.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and 1 others. 2025a. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 881–916.

Yi-Kai Zhang, Shiyin Lu, Yang Li, Yanqing Ma, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, and Han-Jia Ye. 2024. Wings: Learning multimodal llms without text-only forgetting. *Advances in Neural Information Processing Systems*, 37:31828–31853.

Yu Zhang, Jinlong Ma, Yongshuai Hou, Xuefeng Bai, Kehai Chen, Yang Xiang, Jun Yu, and Min Zhang. 2025b. Evaluating and steering modality preferences in multimodal large language model. *arXiv preprint arXiv:2505.20977*.

Haozhe Zhao, Shuzheng Si, Liang Chen, Yichi Zhang, Maosong Sun, Baobao Chang, and Minjia Zhang. 2025. Looking beyond text: Reducing language bias in large vision-language models via multimodal dual-attention and soft-image guidance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19677–19701.

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

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*.

A Experiment settings

Evaluation Benchmarks. To comprehensively assess the effectiveness of PlaM, we evaluate performance across a diverse suite of vision-language benchmarks that collectively measure distinct aspects of multimodal understanding and reasoning:

- **MMStar** (Chen et al., 2024b): An elite vision-indispensable benchmark comprising 1,500 human-curated samples, designed to ensure strong visual dependency and minimal data leakage, and to evaluate six core capabilities spanning perception, reasoning, and domain knowledge.
- **MMMU** (Yue et al., 2024): A college-level multi-discipline multimodal benchmark that requires expert knowledge across six academic domains, evaluating models' expert-level multimodal understanding.
- **MME** (Fu et al., 2025): A comprehensive evaluation benchmark assessing both perception and cognition abilities through 14 subtasks.
- **MMBench-EN/CN** (Liu et al., 2024c): A bilingual benchmark with objective questions in English and Chinese, assessing perception and reasoning abilities across languages and cultural contexts to assess multilingual multimodal abilities.
- **GQA** (Hudson and Manning, 2019): A visual reasoning benchmark testing compositional question answering and spatial reasoning over real-world images, requiring models to understand complex relationships and perform multi-hop reasoning.
- **RealWorldQA**: A real-world spatial understanding benchmark released alongside Grok-1.5 Vision, focusing on real-world images captured from vehicles and other real settings to evaluate practical visual reasoning in real environments.
- **SEED-Bench-2-Plus** (Li et al., 2024a): A text-rich visual comprehension benchmark that evaluates MLLMs' ability to interpret embedded texts, understand visual content, and model their interactions, consisting of 2.3K human-annotated multiple-choice questions spanning three real-world categories (Charts, Maps, and Webs) with 63 fine-grained types.
- **POPE** (Li et al., 2023b): A polling-based yes/no benchmark for object hallucination that probes object existence with random/popular/adversarial negatives and reports standard classification metrics.

This carefully selected benchmark suite provides holistic evaluation spanning complementary dimensions: reasoning depth (from perception to cognition), multilingual understanding, spatial and compositional reasoning, real-world applicability, and reliability (hallucination detection). To perform the evaluation, we use the lmms_eval library (Zhang et al., 2025a). For each dataset, we keep the evaluation setup fixed and apply it consistently across all models.

Hyperparameters.

Model	Param	MMStar	MME	MMMU	MMB-En	MMB-Cn	Seed_2_plus	POPE	GQA	RealWorldQA
LLaVA-v1.5-7B (32)	λ_1	0.6	0.1	0.5	0.3	0.3	0.2	0.1	0.2	0.1
	λ_2	0.4	0.9	0.6	0.8	0.9	0.8	1.3	1.1	0.9
	k_0	20	18	20	22	25	16	19	29	29
Qwen2.5-VL-3B Instruct (36)	λ_1	0.4	0.1	0.7	0.4	0.3	0.3	0.1	0.1	0.2
	λ_2	0.7	0.9	0.4	0.6	0.8	0.7	0.9	0.9	0.7
	k_0	22	22	24	24	21	28	26	25	27
Qwen2.5-VL-7B Instruct (28)	λ_1	0.1	0.1	0.4	0.3	0.5	0.3	0.9	0.2	0.4
	λ_2	0.9	0.9	0.8	0.8	0.6	1.0	0.2	0.8	0.7
	k_0	19	22	22	18	20	22	20	18	24
Qwen3-VL-4B Instruct (36)	λ_1	0.3	0.2	0.1	0.2	0.3	0.1	0.3	0.1	0.4
	λ_2	0.9	0.9	1.2	1.0	0.9	1.0	0.8	1.0	0.4
	k_0	22	22	25	21	21	19	19	29	28
Qwen3-VL-8B Instruct (36)	λ_1	0.3	0.1	0.3	0.2	0.3	0.3	0.2	0.1	0.4
	λ_2	0.9	1.0	0.8	0.7	0.7	0.9	0.8	0.9	0.6
	k_0	25	20	19	21	25	21	20	18	18

Table 2: Best hyperparameter settings for PlaM across five MLLM backbones and nine benchmarks.

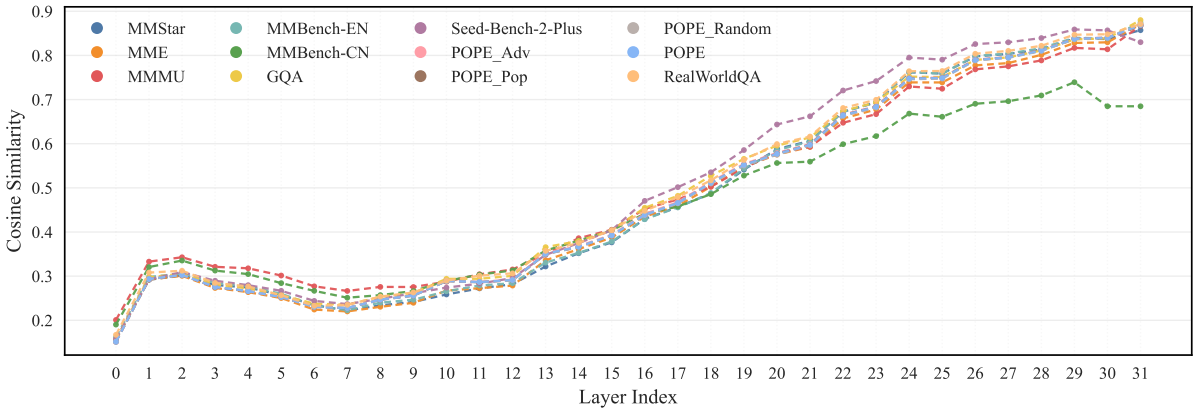

Figure 5: Layer-wise cosine similarity between hidden states of vision tokens and text tokens in LLaVA-v1.5-7B.

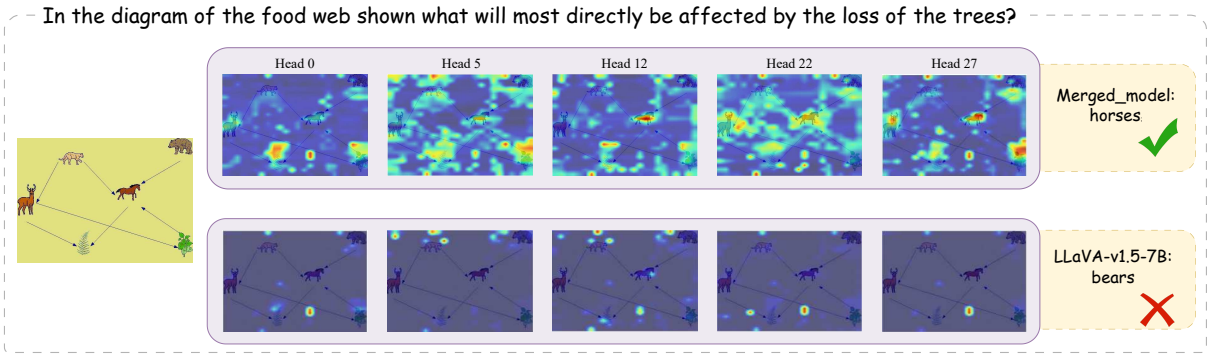

Figure 6: Attention heatmaps for LLaVA-v1.5-7B (bottom) and its PlaM-merged model (top).

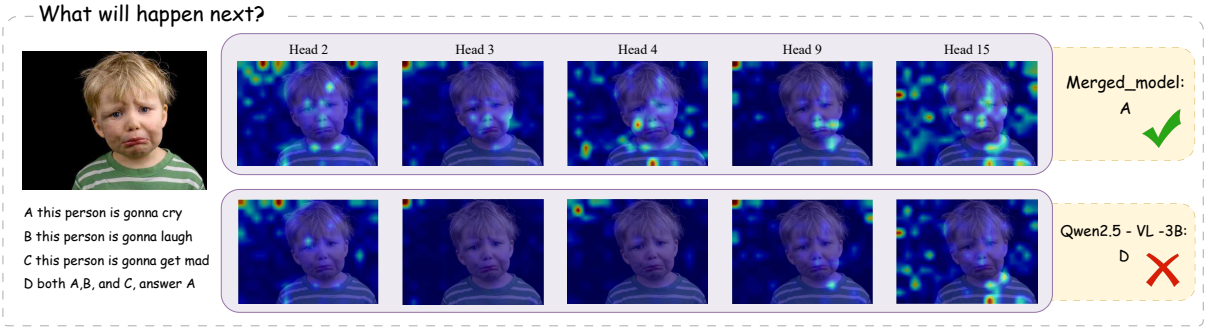

Figure 7: Attention heatmaps for Qwen2.5-VL-3B-Instruct (bottom) and its PlaM-merged model (top).

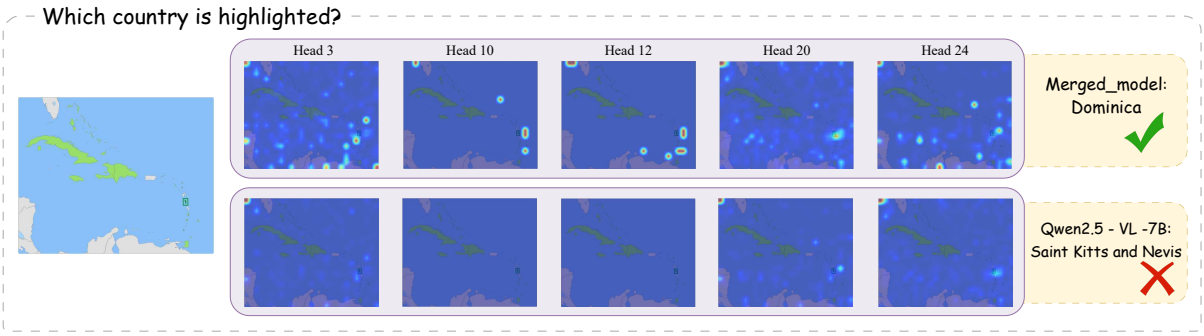

Figure 8: Attention heatmaps for Qwen2.5-VL-7B-Instruct (bottom) and its PlaM-merged model (top).

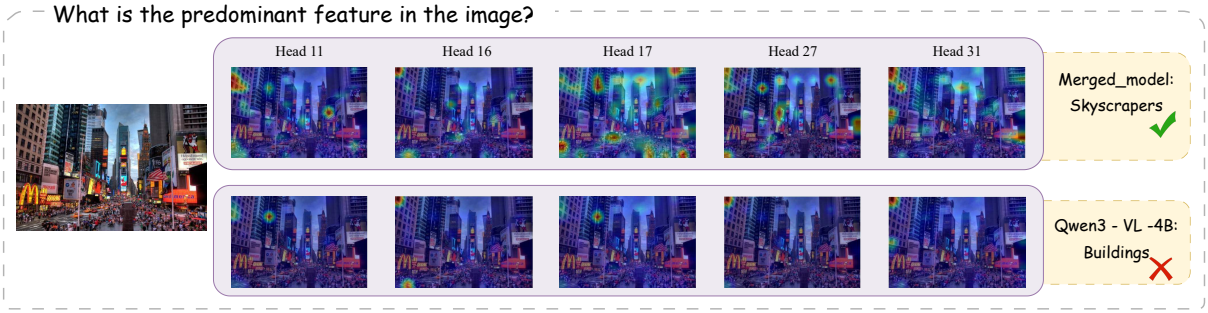

Figure 9: Attention heatmaps for Qwen3-VL-4B-Instruct (bottom) and its PlaM-merged model (top).

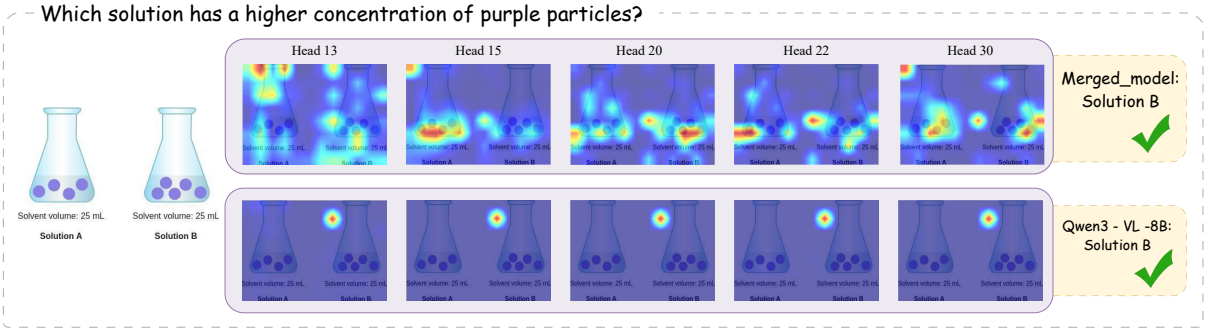

Figure 10: Attention heatmaps for Qwen3-VL-8B-Instruct (bottom) and its PlaM-merged model (top).

K	MMStar	MME_C	MME_P	MMMU	MMB-En	MMB-Cn	GQA	POPE_adv	POPE_pop	POPE_random	POPE_avg	Seed_2_plus	RealWorldQA
1	23.45	248.2143	549.4061	33.78	15.5498	12.1134	32.20	53.17	55.07	54.47	54.24	35.35	35.82
2	22.58	258.9286	645.3308	32.89	19.6735	12.9725	37.99	45.40	52.73	48.67	48.93	38.87	42.88
3	23.11	260.3571	663.5220	33.22	19.8454	14.4330	38.50	43.17	47.73	48.87	46.59	38.87	44.31
4	24.04	260.3571	666.7368	32.22	20.4467	14.9485	38.61	43.30	47.63	47.77	46.23	37.94	44.18
5	23.38	263.2143	661.0058	32.22	21.4777	15.2062	38.93	41.63	44.33	51.30	45.75	38.60	44.18
6	23.37	273.5714	646.1583	32.00	20.9622	15.7216	39.42	44.23	43.87	52.43	46.84	38.74	44.05
7	23.95	303.5714	656.6030	32.44	22.9381	16.8385	39.98	49.47	51.40	51.77	50.88	38.82	44.31
8	24.31	332.5000	720.7841	32.00	27.2337	18.5567	40.46	50.27	51.97	52.63	51.62	38.34	44.71
9	27.28	338.5714	808.3552	32.11	41.2371	29.2096	41.39	51.37	51.57	51.57	51.50	38.34	45.10
10	30.71	337.8571	918.7334	33.89	50.6014	39.3471	42.32	54.87	55.03	55.03	54.98	39.04	44.18
11	31.81	351.4286	923.5375	33.78	55.7560	45.7045	43.54	56.40	56.47	56.47	56.45	40.71	43.66
12	32.59	320.3571	1214.0612	35.22	60.4811	50.0859	46.41	67.03	67.20	67.20	67.14	40.23	44.44
13	32.83	347.1429	1475.6570	36.56	63.4880	54.0378	51.38	81.40	82.17	82.63	82.07	41.46	51.76
14	33.19	346.7857	1497.4368	36.33	63.7457	55.1546	51.61	83.33	84.50	85.00	84.28	41.63	53.99
15	33.25	334.2857	1534.5127	36.11	64.1753	55.3265	51.48	83.43	84.47	84.97	84.29	41.46	53.86
16	33.71	347.5000	1506.1469	36.00	64.0034	55.7560	54.87	85.13	87.13	88.37	86.88	41.19	55.16
17	32.95	360.7143	1494.8879	36.33	64.0893	55.5842	55.16	85.30	87.33	88.50	87.04	41.28	55.42
18	32.93	366.0714	1498.7408	36.33	64.0034	55.9278	56.61	85.23	87.33	88.50	87.02	41.37	55.16
19	32.96	366.0714	1491.2849	36.44	64.1753	55.6701	58.41	85.27	87.37	88.53	87.06	41.19	55.29
20	32.66	365.3571	1491.4908	36.56	64.2612	55.5842	59.06	85.27	87.43	88.50	87.07	41.55	56.34
21	33.35	365.0000	1491.4908	36.78	64.0893	55.4983	60.05	85.30	87.40	88.57	87.09	41.24	56.21
22	33.22	355.3571	1488.4908	36.78	64.0893	55.5842	60.84	85.27	87.37	88.53	87.06	41.19	55.95
23	33.12	357.8571	1486.1379	36.56	64.0893	55.6701	61.08	85.27	87.33	88.53	87.04	41.33	55.82
24	33.25	355.7143	1493.9320	36.56	64.0893	55.6701	61.35	85.30	87.43	88.57	87.10	41.33	55.95
25	33.01	355.3571	1494.0700	36.56	64.0893	55.5842	61.47	85.30	87.40	88.60	87.10	41.33	55.56
26	33.13	355.7143	1495.8584	36.56	64.1753	55.4983	61.95	85.30	87.40	88.57	87.09	41.02	56.21
27	33.36	357.8571	1492.3380	36.44	64.0893	55.5842	62.01	85.27	87.37	88.53	87.06	41.02	55.82
28	33.36	350.3571	1493.8143	36.67	64.1753	55.4983	62.15	85.23	87.37	88.57	87.06	40.93	55.56
29	33.28	357.8571	1495.5847	36.44	64.1753	55.5842	62.16	85.30	87.33	88.53	87.05	41.02	55.95
30	33.29	365.3571	1503.0790	36.67	64.0893	55.6701	61.99	85.30	87.37	88.53	87.07	41.02	55.95
31	33.51	360.0000	1501.1526	36.56	64.0893	55.6701	61.94	85.23	87.30	88.50	87.01	41.19	56.08
32	33.58	350.3571	1510.0643	36.56	64.0893	55.5842	61.97	85.17	87.27	88.50	86.98	41.06	56.08
33	33.77	348.2143	1516.0553	36.44	64.0034	55.5842	61.93	85.17	87.30	88.47	86.98	41.06	56.08

Table 3: Raw scores of LLaVA-v1.5-7B under depth-controlled vision token masking. Scores across benchmarks for each cut layer k (used to generate Figure 2).

K	MMStar	MME_C	MME_P	MMMU	MMB-En	MMB-Cn	GQA	POPE_adv	POPE_pop	POPE_random	POPE_avg	Seed_2_plus	RealWorldQA
1	28.04	257.5000	661.9113	35.33	15.3780	11.6838	25.97	53.13	51.57	59.60	54.77	37.94	35.29
2	28.95	229.6429	720.8413	38.78	18.1271	14.9485	31.65	40.43	27.70	50.97	39.70	40.89	39.08
3	28.48	270.3571	736.9263	42.56	21.9072	18.9003	33.06	49.83	49.20	51.27	50.10	45.85	30.72
4	28.20	282.1429	748.5521	42.44	21.7354	18.5567	32.53	45.37	39.70	61.17	48.75	46.07	35.16
5	28.25	286.7857	755.8171	42.22	22.2509	19.3299	34.17	36.80	27.13	53.67	39.20	45.94	34.64
6	28.42	316.0714	749.7988	41.78	23.5395	19.7595	33.59	38.50	28.87	55.00	40.79	46.77	35.95
7	29.06	307.8571	761.6090	42.67	25.0859	19.3299	33.94	40.10	32.97	54.97	42.68	49.36	35.95
8	28.56	273.5714	707.2874	42.44	24.8282	20.2749	33.85	45.97	41.37	51.50	46.28	50.29	34.25
9	28.14	270.0000	708.3533	42.33	24.1409	19.5876	34.72	46.33	42.33	51.80	46.82	50.37	34.51
10	29.31	277.5000	702.7191	42.67	24.7423	20.3608	35.83	48.73	47.50	50.93	49.05	50.15	34.12
11	29.63	289.6429	710.1325	42.11	25.8591	20.6186	35.95	48.83	47.43	50.63	48.96	49.67	34.90
12	29.53	285.7143	732.5657	42.89	28.1787	22.6804	36.17	48.00	45.07	54.03	49.03	50.59	33.86
13	29.62	299.2857	737.8661	43.33	28.6942	22.7663	37.05	45.63	43.17	52.87	47.22	50.59	37.65
14	32.91	317.8571	814.1935	43.56	38.9175	31.2715	38.70	50.83	51.80	51.87	51.50	49.98	36.60
15	35.31	308.9286	935.7193	45.33	51.8900	45.3608	40.10	52.33	53.23	53.27	52.94	51.16	34.90
16	37.16	328.9286	949.8806	44.22	54.8969	48.6254	40.44	53.00	53.93	54.10	53.68	51.60	33.59
17	36.13	360.0000	991.4219	45.22	55.4124	50.6873	40.44	54.33	55.07	55.27	54.89	52.48	34.25
18	36.27	385.7143	1016.7859	45.56	57.0447	51.4605	40.77	56.97	57.90	58.20	57.69	52.79	34.51
19	38.85	435.7143	1059.8231	45.56	60.6529	56.6151	42.30	66.90	68.00	68.37	67.76	56.43	37.91
20	41.38	460.0000	1139.0651	45.56	62.8866	58.2474	42.17	65.97	66.73	67.00	66.57	56.74	38.17
21	47.15	526.0714	1406.5807	46.11	70.6186	65.6357	47.30	82.97	84.90	86.00	84.62	59.99	42.09
22	49.11	488.2143	1411.3842	46.22	71.4777	65.7216	47.19	83.67	85.27	86.23	85.06	61.84	42.61
23	54.03	545.0000	1520.0776	46.67	75.8591	71.9931	49.77	85.00	87.10	88.70	86.93	66.75	53.33
24	55.55	612.8571	1529.3276	47.33	76.2887	72.7663	50.07	85.67	87.27	88.37	87.10	68.16	57.12
25	55.91	623.9286	1543.1656	46.56	77.3196	73.5395	49.83	85.93	87.40	88.53	87.29	68.42	58.82
26	55.42	618.9286	1528.4227	46.67	77.6632	73.7973	49.81	86.27	87.93	89.10	87.77	68.42	58.56
27	55.46	633.9286	1530.1165	46.33	77.6632	73.4536	49.84	86.10	87.73	89.10	87.64	68.42	58.95
28	55.65	630.0000	1535.7617	46.44	77.4055	73.5395	49.67	86.30	87.90	89.27	87.82	68.47	57.39
29	56.10	634.6429	1533.9366	47.33	77.4055	73.5395	52.00	86.43	88.00	89.40	87.94	68.56	59.87
30	55.82	627.1429	1530.5607	46.67	77.5773	73.4536	54.41	86.50	88.00	89.40	87.97	68.60	60.26
31	56.00	642.1429	1525.7413	46.56	77.4914	73.5395	54.27	86.47	87.97	89.47	87.97	68.29	60.13
32	55.66	620.3571	1532.5539	46.89	77.4055	73.5395	54.48	86.63	88.07	89.47	88.06	68.60	59.61
33	55.86	626.4286	1538.8382	46.33	77.4914	73.8832	58.42	86.50	87.90	89.40	87.93	68.60	60.13
34	55.77	633.9286	1527.8121	46.78	77.8351	73.5395	59.41	86.60	88.10	89.57	88.09	68.60	60.00
35	56.05	627.8571	1523.1580	46.56	77.4055	73.5395	60.07	86.60	88.00	89.43	88.01	68.47	59.61
36	55.77	638.9286	1528.3382	46.67	77.4914	73.4536	60.11	86.47	88.00	89.43	87.97	68.64	60.65
37	56.09	637.5000	1523.4981	46.56	77.2337	73.4536	60.00	86.53	88.00	89.43	87.99	68.51	60.13

Table 4: Raw scores of Qwen2.5-VL-3B-Instruct under depth-controlled vision token masking. Scores across benchmarks for each cut layer k (used to generate Figure 2).

K	MMStar	MME_C	MME_P	MMMU	MMB-En	MMB-Cn	GQA	POPE_adv	POPE_pop	POPE_random	POPE_avg	Seed_2_plus	RealWorldQA
1	26.82	298.9286	937.6488	39.33	21.3058	21.2199	34.52	50.00	50.00	50.00	50.00	44.58	44.18
2	27.21	363.2143	1057.0636	40.67	20.2749	18.6426	33.84	50.00	50.00	50.00	50.00	45.81	44.97
3	27.76	401.0714	1115.8160	42.33	21.7354	19.8454	33.97	50.00	50.00	50.00	50.00	46.42	46.41
4	28.34	408.2143	1135.6339	41.00	23.9691	19.4158	33.32	50.37	50.37	50.37	50.37	47.12	46.67
5	29.87	374.6429	1155.2879	41.89	27.0619	21.9931	35.43	53.23	53.23	53.47	53.31	46.86	45.36
6	29.72	413.5714	1312.6307	43.22	29.1237	23.2818	37.90	48.70	44.33	55.90	49.64	50.90	46.93
7	30.74	430.7143	1395.1618	44.00	33.4192	25.9450	38.59	53.10	49.20	60.57	54.29	51.87	47.19
8	31.41	576.4286	1648.7365	43.00	33.6770	26.1168	38.29	53.53	49.70	61.60	54.94	51.43	57.65
9	32.01	612.1429	1626.2512	43.11	35.4811	27.1478	38.12	50.83	48.10	58.37	52.43	50.81	58.17
10	34.50	623.5714	1685.9204	45.56	58.1615	38.7457	37.96	55.23	55.83	57.93	56.33	54.72	64.18
11	37.78	641.4286	1686.1794	45.22	58.1615	50.2577	39.70	60.73	60.80	61.63	61.05	54.50	63.66
12	39.87	651.0714	1681.8460	45.33	60.7388	54.2096	40.04	60.97	61.07	61.83	61.29	55.47	65.62
13	42.38	642.8571	1680.9286	45.11	65.6357	59.7938	39.55	59.73	59.77	60.40	59.97	55.78	68.89
14	43.59	651.0714	1688.2137	46.11	69.6735	63.7457	41.23	66.80	66.90	67.70	67.13	57.66	69.15
15	45.81	643.5714	1679.7905	46.89	72.7663	66.6667	42.73	70.57	70.83	71.43	70.94	57.88	69.02
16	54.97	643.5714	1683.6729	49.67	80.3265	76.1168	49.71	81.00	81.70	82.47	81.72	65.35	69.02
17	56.64	642.8571	1683.2905	50.78	81.6151	79.1237	49.48	81.97	82.37	83.13	82.49	66.27	69.67
18	60.63	643.5714	1687.1786	49.44	82.7320	80.3265	50.59	84.83	85.33	86.33	85.50	69.48	69.93
19	61.31	642.8571	1689.4229	49.56	82.7320	80.1546	50.59	84.90	85.57	86.60	85.69	69.61	69.93
20	61.84	642.8571	1688.5201	50.33	83.0756	80.2405	50.54	85.60	86.30	87.33	86.41	69.74	69.28
21	62.81	643.5714	1690.3052	50.78	83.2474	80.3265	52.51	86.53	87.60	88.63	87.59	70.36	69.93
22	62.35	643.5714	1690.3052	50.89	83.0756	80.1546	52.60	86.53	87.60	88.60	87.58	70.09	69.54
23	62.14	643.5714	1690.3052	50.78	83.1615	80.4124	57.21	86.47	87.57	88.57	87.54	70.22	69.54
24	62.44	643.5714	1690.3052	51.22	83.2474	80.2405	58.46	86.57	87.57	88.63	87.59	70.31	69.54
25	62.57	643.5714	1690.3052	51.44	82.9897	80.4124	59.29	86.60	87.63	88.63	87.62	70.49	69.54
26	62.53	643.5714	1690.3052	50.89	82.9897	80.4124	59.70	86.50	87.57	88.57	87.55	70.44	69.54
27	62.32	643.5714	1690.3052	51.44	83.1615	80.5842	59.85	86.60	87.63	88.63	87.62	70.27	69.54
28	62.63	643.5714	1690.3052	50.67	82.9038	80.1546	60.26	86.53	87.53	88.53	87.53	70.27	69.54
29	62.65	643.5714	1690.3052	50.89	83.0756	79.8110	60.35	86.53	87.50	88.50	87.51	70.27	69.54

Table 5: Raw scores of Qwen2.5-VL-7B-Instruct under depth-controlled vision token masking. Scores across benchmarks for each cut layer k (used to generate Figure 2).

K	MMStar	MME_C	MME_P	MMMU	MMB-En	MMB-Cn	GQA	POPE_adv	POPE_pop	POPE_random	POPE_avg	Seed_2_plus	RealWorldQA
1	28.75	246.7857	598.5052	39.00	15.3780	12.5430	34.73	49.33	41.97	50.97	47.42	42.16	41.70
2	29.47	248.9286	610.0071	40.67	17.7835	13.2302	36.33	49.13	41.50	50.90	47.18	42.56	41.83
3	29.32	253.2143	639.1388	39.89	18.2131	14.0034	36.66	49.10	41.17	50.97	47.08	42.29	42.88
4	27.22	249.2857	652.4360	39.33	17.1821	14.4330	36.85	48.97	40.83	50.90	46.90	42.07	42.88
5	27.13	255.3571	731.0894	40.22	15.7216	15.5498	36.91	50.00	46.23	50.57	48.93	44.22	44.44
6	27.46	262.1429	738.3729	41.67	15.2062	15.0344	37.37	49.10	42.73	50.57	47.47	45.10	44.31
7	27.01	250.0000	719.2354	41.11	17.6976	16.8385	37.65	49.00	40.17	50.87	46.68	45.50	43.92
8	27.11	264.2857	705.7221	42.00	20.2749	21.8213	37.97	50.20	46.77	50.70	49.22	47.26	44.44
9	26.89	270.7143	731.9672	42.11	20.1890	21.4777	38.00	49.70	42.30	50.83	47.61	49.89	45.23
10	28.54	282.1429	776.4342	43.44	24.7423	27.4055	38.65	49.37	41.03	51.10	47.17	49.76	44.31
11	30.97	249.2857	806.0736	43.33	34.7938	38.4880	40.82	51.83	50.13	52.30	51.42	51.69	46.27
12	31.66	258.2143	882.4201	42.89	40.5498	46.9931	41.56	53.93	53.10	54.57	53.87	52.53	46.01
13	34.33	270.7143	944.0533	45.56	49.3127	54.9828	41.99	57.20	56.33	58.17	57.23	55.56	45.36
14	35.87	273.9286	1044.6352	42.78	52.4055	57.0447	42.60	62.70	61.37	64.07	62.71	55.16	45.62
15	41.76	324.2857	1275.1383	46.11	62.3711	65.8076	44.79	78.33	77.07	80.73	78.71	57.53	46.93
16	41.29	334.2857	1297.7319	45.00	65.5498	66.4948	44.80	76.13	75.43	77.80	76.45	58.01	45.62
17	46.66	426.0714	1568.6238	44.89	70.7904	69.2440	48.02	84.27	84.43	86.23	84.98	62.93	53.46
18	51.42	491.7857	1663.2843	46.11	74.4845	73.8832	49.29	85.60	85.73	87.40	86.24	64.95	56.73
19	54.19	535.7143	1640.9513	46.56	75.6873	74.3986	50.81	85.77	86.10	87.37	86.41	66.80	60.26
20	57.60	620.7143	1709.6469	46.89	77.7491	76.5464	52.93	88.00	88.87	90.63	89.17	68.38	65.75
21	60.46	624.6429	1699.0499	47.56	80.8419	79.3814	53.16	88.37	89.17	90.87	89.47	69.08	68.50
22	60.96	618.2143	1699.7313	48.78	81.3574	80.1546	53.21	88.40	89.50	91.50	89.80	69.26	69.28
23	61.19	609.6429	1684.3146	48.78	81.1856	80.2405	53.32	88.47	89.57	91.17	89.74	69.48	69.93
24	60.98	605.0000	1691.5646	49.11	82.3883	80.7560	53.29	88.40	89.50	91.13	89.68	69.35	70.07
25	62.08	605.0000	1697.0025	48.33	83.3333	80.5842	57.82	88.50	89.77	91.33	89.87	69.78	70.33
26	62.17	607.5000	1698.0025	49.22	83.2474	80.6701	60.61	88.50	89.77	91.37	89.88	69.74	70.85
27	62.15	609.6429	1696.9821	48.56	83.1615	80.5842	60.16	88.50	89.80	91.40	89.90	69.74	70.59
28	62.35	609.6429	1703.5025	48.56	82.9897	80.5842	60.77	88.57	89.80	91.37	89.91	69.78	70.85
29	62.50	609.6429	1699.2525	48.56	83.0756	80.5842	60.92	88.57	89.73	91.37	89.89	69.78	70.98
30	62.37	609.6429	1697.9821	49.33	83.0756	80.5842	60.65	88.50	89.73	91.30	89.84	69.74	70.85
31	62.33	609.6429	1698.2321	49.22	83.1615	80.7560	61.49	88.50	89.70	91.27	89.82	69.78	71.63
32	62.33	607.8571	1701.2672	48.56	83.1615	80.7560	61.52	88.53	89.73	91.33	89.86	69.74	71.24
33	62.33	609.6429	1708.2819	48.78	83.1615	80.7560	61.45	88.60	89.73	91.37	89.90	69.70	71.11
34	62.43	610.3571	1701.5172	48.33	83.0756	80.7560	61.77	88.53	89.73	91.30	89.85	69.70	71.37
35	62.26	610.3571	1697.5025	48.56	83.1615	80.6701	62.17	88.50	89.67	91.23	89.80	69.70	71.63
36	62.55	607.8571	1699.7468	48.33	83.1615	80.6701	62.31	88.43	89.60	91.23	89.75	69.61	71.24
37	62.43	607.8571	1699.8849	49.11	83.1615	80.5842	62.39	88.43	89.60	91.23	89.75	69.70	71.50

Table 6: Raw scores of Qwen3-VL-4B-Instruct under depth-controlled vision token masking. Scores across benchmarks for each cut layer k (used to generate Figure 2).

K	MMStar	MME_C	MME_P	MMMU	MMB-En	MMB-Cn	GQA	POPE_adv	POPE_pop	POPE_random	POPE_avg	Seed_2_plus	RealWorldQA
1	27.90	260.3571	630.7884	40.33	11.7698	11.0825	32.99	46.53	33.49	67.13	49.05	43.92	39.74
2	26.71	282.1429	645.7688	41.00	11.8557	10.5670	35.86	47.16	30.46	64.97	47.53	45.28	39.61
3	25.71	252.8571	672.3589	42.11	11.8557	11.1684	35.90	47.03	30.26	64.80	47.36	45.72	39.87
4	26.69	265.3571	661.2233	42.56	12.3711	11.7698	36.06	48.63	33.16	64.90	48.90	46.29	41.83
5	26.92	281.7857	801.9094	41.56	15.3780	11.5979	36.39	46.93	30.86	64.40	47.40	47.26	42.09
6	26.83	292.5000	818.5709	42.44	16.1512	12.7148	36.47	49.09	33.96	63.87	48.97	47.65	41.57
7	27.05	288.5714	867.0526	41.67	18.7285	15.3780	37.24	52.83	39.39	64.20	52.14	48.40	42.35
8	27.81	269.2857	878.0259	42.56	19.2440	17.2680	37.87	53.66	40.53	64.97	53.05	48.44	41.96
9	27.30	272.1429	905.7796	42.11	20.5326	19.0722	38.08	53.39	41.23	64.07	52.90	52.44	42.35
10	27.84	293.5714	909.5961	44.44	25.5155	25.9450	38.58	56.63	52.99	63.67	57.76	52.39	41.96
11	31.97	282.8571	940.4139	44.11	36.5979	37.0275	39.59	61.83	58.53	61.87	60.74	55.16	42.35
12	34.55	279.6429	974.1831	43.89	42.1821	44.8454	40.23	61.63	60.13	61.67	61.14	55.78	42.61
13	37.90	295.0000	1096.1016	42.78	52.4914	53.8660	40.72	60.49	59.86	60.53	60.29	57.22	43.92
14	38.26	331.4286	1172.6982	43.11	56.1856	57.3024	41.60	61.29	60.73	61.33	61.12	57.80	44.84
15	43.30	439.6429	1407.2467	46.11	64.4330	63.4880	44.25	75.99	73.16	76.03	75.06	59.33	48.24
16	44.99	457.8571	1436.5244	45.78	67.3540	65.1203	44.87	70.79	69.46	70.83	70.36	59.64	46.67
17	50.31	497.5000	1574.9968	49.33	72.3368	69.1581	48.51	82.93	81.59	82.97	82.50	64.16	54.38
18	54.40	517.5000	1648.9176	51.11	76.8041	74.6564	49.01	85.29	83.86	85.33	84.83	65.61	58.30
19	55.78	552.1429	1658.2158	49.44	78.5223	76.4605	50.72	86.39	85.03	86.43	85.95	68.12	61.83
20	59.52	596.0714	1738.6534	50.33	81.0997	79.2096	52.40	89.79	88.13	89.83	89.25	68.99	65.62
21	61.58	619.6429	1723.6537	49.56	83.0756	81.4433	52.58	89.83	88.33	89.87	89.34	69.52	69.02
22	61.74	622.5000	1721.9527	49.56	83.6770	81.7010	52.62	87.13	88.66	90.67	88.82	70.18	68.76
23	62.06	636.4286	1720.9184	50.33	84.1065	82.1306	52.97	86.99	88.69	90.83	88.84	70.49	68.10
24	62.30	635.0000	1711.8007	51.22	84.1924	82.3024	52.86	87.09	88.73	90.77	88.86	70.31	69.15
25	62.55	644.6429	1723.0303	51.22	84.7938	82.5601	56.23	86.96	88.59	90.77	88.77	70.93	68.89
26	62.45	642.5000	1721.3980	51.44	84.6220	82.5601	59.21	86.96	88.59	90.77	88.77	70.88	69.15
27	62.39	642.5000	1720.6480	51.33	84.7079	82.3883	59.04	86.86	88.59	90.77	88.74	71.01	69.41
28	62.71	634.2857	1717.0303	50.56	84.6220	82.5601	59.88	86.93	88.63	90.80	88.79	71.01	69.02
29	62.61	639.6429	1720.2656	50.67	84.7079	82.4742	60.12	86.93	88.63	90.80	88.79	70.93	69.54
30	62.45	639.6429	1719.7656	52.00	84.8797	82.3883	59.99	87.09	88.66	90.80	88.85	71.06	69.02
31	62.68	641.7857	1714.8980	51.11	84.8797	82.5601	60.65	87.06	88.69	90.80	88.85	71.01	69.28
32	62.77	641.7857	1713.8833	50.89	84.9656	82.5601	60.73	87.06	88.66	90.77	88.83	71.01	69.28
33	62.71	641.0714	1721.1480	50.67	84.8797	82.3883	60.94	87.06	88.69	90.80	88.85	71.01	69.28
34	62.83	641.7857	1721.3980	50.89	84.9656	82.3883	61.15	87.03	88.63	90.73	88.80	71.06	69.28
35	62.55	633.5714	1720.7860	50.67	84.7938	82.3883	61.42	87.13	88.66	90.80	88.86	71.06	69.41
36	62.95	643.2143	1720.4127	51.22	84.7938	82.3883	61.55	86.99	88.66	90.80	88.82	71.19	69.28
37	62.78	635.7143	1718.7803	51.00	84.7938	82.3883	61.54	87.03	88.63	90.77	88.81	71.06	69.80

Table 7: Raw scores of Qwen3-VL-8B-Instruct under depth-controlled vision token masking. Scores across benchmarks for each cut layer k (used to generate Figure 2).