MaLoRA: Mitigating Text-Centric Bias in Multimodal Large Language Models through Key-Space Alignment

Anonymous authors
Paper under double-blind review

Abstract

Multimodal large language models (MLLMs) often exhibit text-centric bias under joint image–text inputs, over-relying on textual signals while under-utilizing visual evidence. We analyze the decoder self-attention and observe persistent cross-modal misalignment in the attention key space, where visual and textual keys form separated distributions consistent with preferential attention to text tokens. To mitigate this bias, we introduce MaLoRA, a text-centric bias mitigation framework that directly intervenes in the self-attention key space to re-balance visual and textual evidence. MaLoRA combines modality-aware gated key adaptation, multi-kernel maximum mean discrepancy (MMD) alignment between multimodal key distributions, and Gram-reference regularization to preserve within-modality key geometry during alignment. We compare MaLoRA with training-free decoding methods and trained adaptation baselines across multiple MLLM backbones. Evaluations cover general multimodal benchmarks and robustness tests targeting text-dominant failures, including image–text conflict and irrelevant-context settings, as well as hallucination benchmarks such as POPE, AMBER, and HallusionBench. The results show that MaLoRA reduces key-space divergence and improves robustness to text-dominant failures, with its clearest gains on hallucination-sensitive benchmarks and competitive performance in conflict-sensitive settings.

1 Introduction

Multimodal large language models (MLLMs) have achieved strong performance on vision–language tasks by integrating visual perception with the reasoning ability of large language models. Ideally, given image–text inputs, MLLMs should attend to the modality that provides task-relevant evidence. However, recent studies reveal a persistent **text-centric bias**: MLLMs often over-rely on textual information while under-utilizing visual cues, even when visual evidence is crucial. Such bias weakens the visual grounding and leads to hallucinated responses. Existing studies mitigate this problem through data construction, alignment objectives, architectural designs, and inference-time correction Chen et al. (2024a); Sun et al. (2024b); Huang et al. (2024); Leng et al. (2024); Yin et al. (2024). Bias-oriented methods such as HALVA Sarkar et al. (2025) and LACING Zhao et al. (2025b) further reduce language-prior dominance and enhance visual grounding. Prior work has documented a persistent modality gap between visual and textual representations in multimodal contrastive models Liang et al. (2022). In contrast, we investigate whether an analogous misalignment exists specifically in the decoder self-attention key space.

To investigate this mechanism, we analyze the key space of the decoder self-attention. We focus on the attention **key space** and examine multiple MLLM backbones, datasets, and transformer layers. Our analysis reveals a consistent pattern: visual keys and textual keys form separate distributions. Specifically, the cross-modal divergence is larger than within-modal variation, indicating that such separation is not a model or dataset-specific artifact, but a geometric property of MLLMs. Since the decoder is typically initialized from a pretrained language model, its query representations inherently preserve language-oriented geometric structure. As a result, decoder queries are more compatible with textual keys, causing text tokens to receive higher attention scores under dot-product attention.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

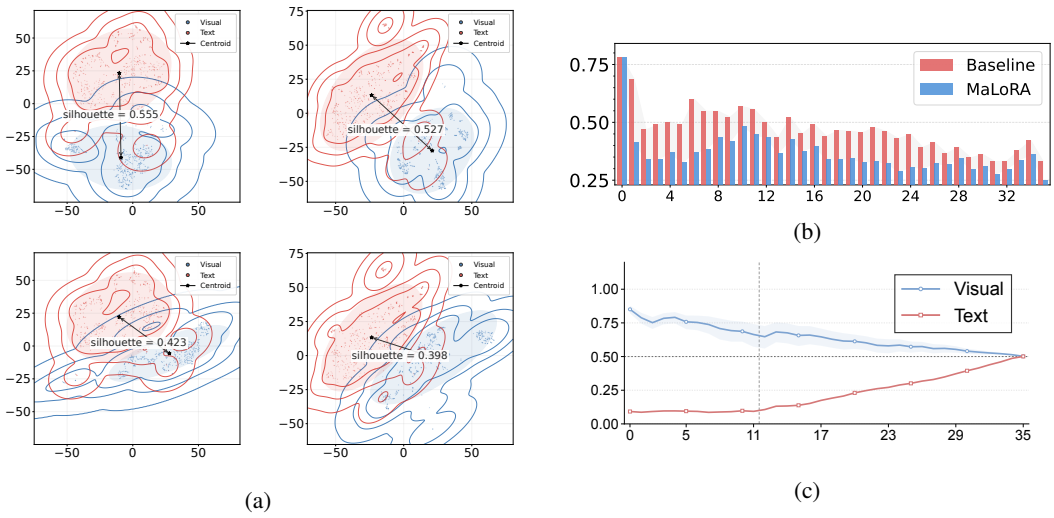


Figure 1: **(a)** t-SNE visualizations of hidden representations from LLaVA-1.5-7B. The left and right columns correspond to MMMU and MMBench-EN, respectively, while the top and bottom rows correspond to the Base model and MaLoRA. **(b)** Representation-gap analysis of Qwen3-VL-8B on MMBench-EN. **(c)** Layer-wise gate values for visual and text tokens in Qwen3-VL-8B, showing their evolution with decoder depth.

Thus, we argue that the text-centric bias arises because the keys are misaligned in the latent space that determines attention allocation.

This observation reframes bias mitigation as a *visual–textual key-space alignment problem*. However, effective alignment faces two challenges. First, visual and textual tokens originate from distinct semantic spaces, necessitating divergent adaptation directions. Constraining them to a shared transformation mechanism risks introducing gradient interference and conflicting updates. Besides, simple distribution alignment may collapse or distort the intrinsic structure of each modality, damaging the representations needed for perception and reasoning. Therefore, key-space alignment should reduce cross-modal discrepancy while preserving modality-specific geometry.

To address these challenges, we propose **MaLoRA**, a modality-aware adaptation framework for mitigating text-centric bias through key-space alignment. MaLoRA introduces three components: **Gated Modality Layer**, which assigns modality-specific adaptation paths to visual and textual tokens; **MK-MMD**, which explicitly reduces the distributional gap between visual and textual keys; and **Gram regularization**, which preserves the internal relational structure of key representations during alignment. Together, these components allow MaLoRA to reshape the attention key space in a controlled manner.

Extensive experiments show that MaLoRA consistently reduces the visual–textual key-space gap, increases Textual-to-Visual attention, and improves performance on diverse multimodal tasks. These results support our hypothesis that text-centric bias is closely related to key-space misalignment and demonstrate that addressing attention-internal geometry is an effective way to improve visual utilization in MLLMs.

Our contributions are summarized as follows:

- Through modality-gap profiling and layer-wise analysis, we identify a persistent distributional gap between visual and textual keys across models, datasets, and layers.
- We explain text-centric bias from the perspective of decoder self-attention, linking biased attention allocation to visual–textual key-space misalignment.
- MaLoRA is introduced as a modality-aware adaptation method that combines a gated modality layer, MK-MMD alignment, and Gram regularization to reduce misalignment of the key-space while preserving the structure of the original modality.

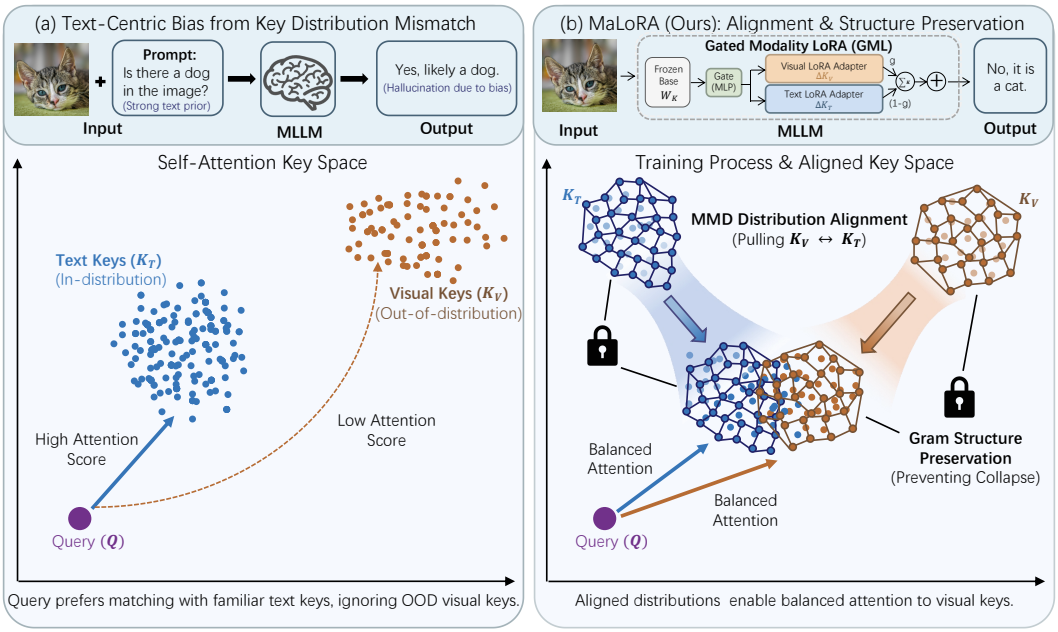


Figure 2: The MaLoRA framework addressing *text-centric bias* in MLLMs.

- Comprehensive experiments validate the effectiveness of MaLoRA in both attention behavior and downstream accuracy. MaLoRA achieves the best performance on all reported benchmarks across LLaVA-1.5-7B, Qwen2.5-VL-7B, and Qwen3-VL-8B.

2 KEY-SPACE MISALIGNMENT

Under image–text inputs, MLLMs often over-rely on textual information during generation, exhibiting *text-centric bias*. ***We hypothesize that this phenomenon is not solely due to data composition or alignment objectives, but stems from*** **key-space distribution misalignment** ***in decoder self-attention.*** We recall the standard self-attention formulation:

$$\mathrm{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}, \tag{1}$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ denote the query, key, and value matrices, respectively, and d is the per-head dimension. At the token level, the attention score between a query $\mathbf{q}_i$ and a key $\mathbf{k}_j$ is determined by their scaled dot product $\mathbf{q}_i^\top \mathbf{k}_j/\sqrt{d}$. In contrast, visual tokens are projected from a vision encoder and concatenated with text tokens, inheriting a cross-modal gap before entering the decoder. At each layer, both visual and textual keys are produced from their hidden states through a shared key projection $\mathbf{W}_k$. Despite sharing the projection, $\mathbf{K}_{\text{vis}}$ and $\mathbf{K}_{\text{text}}$ may still exhibit substantial distribution differences in the key space. As a result, $\mathbf{q}$ may be better aligned with $\mathbf{k}^{\text{text}}$, yielding higher similarity scores and attention allocation bias toward text tokens.

To test this hypothesis, we perform *key-space probing*, extracting keys and analyzing their geometry and divergence, on several decoder layers of representative MLLMs. Given joint image-text inputs, we extract the sets of visual and textual key vectors$\mathbf{k}^{\text{vis}}, \mathbf{k}^{\text{text}}$ from target layers and mask padding tokens. First, we visualize the geometry of the two key sets on selected layers using t-SNE. Second, we quantify cross-modal discrepancy using distributional measures (MMD) Gretton et al. (2012), comparing $P(\mathbf{K}^{\text{vis}})$ and $P(\mathbf{K}^{\text{text}})$ and contrasting them with within-modality baselines computed from split subsets. As shown in Table 1, the divergence between visual and textual keys is consistently larger than the within-modality baselines across different MLLM backbones and datasets, supporting our hypothesis of persistent key-space misalignment.

Qualitatively, the base model exhibits a clear separation between visual and textual key representations in the t-SNE space across both benchmarks, as shown in the left panel of Figure 1. After ap-

Table 1: Key-space divergence at decoder layer 1 for LLaVA-1.5-7B, Qwen2.5-VL-7B, and Qwen3-VL-8B on MMBench-EN and MMMU (10-option). MMD and JS show Image vs. Text consistently larger than Image vs. Image and Text vs. Text. This gap aligns with a persistent separation between visual and textual keys in the projected embedding space.

(a) MMD

	LLaVA-1.5-7B		Qwen2.5-VL-7B		Qwen3-VL-8B	
Comparison	MMB-EN	MMMU	MMB-EN	MMMU	MMB-EN	MMMU
Image vs. Text	**0.94**	**0.95**	**0.57**	**0.71**	**0.76**	**0.76**
Image vs. Image	0.01	0.01	0.01	0.01	0.01	0.09
Text vs. Text	0.02	0.02	0.01	0.07	0.04	0.03

(b) JS

	LLaVA-1.5-7B		Qwen2.5-VL-7B		Qwen3-VL-8B	
Comparison	MMB-EN	MMMU	MMB-EN	MMMU	MMB-EN	MMMU
Image vs. Text	**0.82**	**0.84**	**0.86**	**0.86**	**0.72**	**0.66**
Image vs. Image	0.04	0.03	0.04	0.04	0.04	0.09
Text vs. Text	0.14	0.10	0.14	0.14	0.15	0.15

plying MaLoRA, the two modalities appear more closely interleaved, suggesting a reduction in their cross-modal discrepancy. The upper-right panel reports the layer-wise modality gap for Qwen3-VL-8B on MMBench-EN and shows that MaLoRA consistently reduces the gap across decoder layers. The lower-right panel presents the corresponding layer-wise gate values, which become less polarized in deeper layers, indicating a gradual transition from modality-specific routing to more flexible cross-modal fusion. Together, these observations suggest that MaLoRA reduces cross-modal key-space discrepancy while enabling more flexible multimodal interaction in deeper layers.

This key space separation affects attention computation. From Eq. (1), when $\mathbf{q}$ is better aligned with the textual key distribution, $\mathbf{q}^\top \mathbf{k}^{\text{text}}$ tends to be larger, while $\mathbf{q}^\top \mathbf{k}^{\text{vis}}$ tends to be smaller and is further down-weighted after softmax. Consequently, even when visual tokens contain question-critical information, the model is more likely to allocate attention to text tokens or previously generated context, resulting in insufficient use of image evidence when generating the answer. These results suggest that visual–textual key misalignment is an important factor in *text-centric bias*.

3 METHOD

To mitigate *text-centric bias* in MLLMs under joint image–text inputs, we propose **MaLoRA**, illustrated in Figure 2. Its Gated Modality Layer (GML) augments the decoder key projection k_{proj} with modality-specific LoRA branches and a lightweight gate. During training, MK-MMD aligns visual and textual key distributions, while Gram regularization preserves within-modality geometry. Together, they promote a balanced key space while retaining standard autoregressive decoding at inference.

Following the self-attention formulation in Eq. (1), for the decoder at layer l, the input hidden states are $\mathbf{H}^{(l)}$. We assume that $\mathbf{K}^{(l)}_{\text{vis}}$ and $\mathbf{K}^{(l)}_{\text{text}}$ exhibit persistent misalignment in the key space, which is associated with decoder queries preferentially matching text tokens.

3.1 GATED MODALITY LAYER (GML)

To enable token-wise modality conditioning, we augment the key projection layer k_{proj} with a gated modality-aware adaptation layer, termed **Gated Modality Layer (GML)**. GML is implemented by two modality-specific LoRA branches, which separately model visual and textual key adaptations. Given the input hidden states $\mathbf{H}^{(l)}$, the resulting key representation is computed as:

$$\mathbf{K}^{(l)} = \mathbf{H}^{(l)}\mathbf{W}^{(l)}_k + \alpha \left[g\Delta\mathbf{K}^{(l)}_{\text{vis}} + (1-g)\Delta\mathbf{K}^{(l)}_{\text{text}} \right], \tag{2}$$

where $\Delta\mathbf{K}^{(l)}_{\text{vis}}$ and $\Delta\mathbf{K}^{(l)}_{\text{text}}$ denote the visual-specific and text-specific LoRA adaptations, respectively. The gate $g \in [0,1]$ controls the token-wise contribution of the two modality branches and is predicted by a lightweight gating network:

$$g = \sigma\left(\text{MLP}(\mathbf{H}^{(l)})\right), \tag{3}$$

where σ denotes the sigmoid activation.

To ensure modality-specific routing, we supervise the gating network with a token-level binary cross-entropy loss $\mathcal{L}_{\text{gate}}$ using modality labels obtained from the input packing:

$$\mathcal{L}^{(l)}_{\text{gate}} = -\left[y\log(g) + (1-y)\log(1-g) \right], \tag{4}$$

where y is the modality label of the current token, with $y = 1$ for visual tokens and $y = 0$ for textual tokens. During training, we adopt a linear annealing schedule that gradually transitions the gating mechanism from early label-guided routing to fully model-predicted routing.

To verify whether the routing behavior of GML across different layers aligns with our expectation, we further analyze the gate outputs after MaLoRA training on VQAv2 based on Qwen3-VL-8B. Specifically, we randomly select 20 cases and sample the gate output distributions of different layers during inference, as visualized in (Figure 1, right). The results show that gate outputs in shallow layers are typically closer to 0 or 1, exhibiting a more polarized modality selection pattern. This indicates that, at lower layers, the model mainly performs routing based on relatively explicit visual/textual token boundaries. As the network becomes deeper, information between textual and visual tokens is gradually fused through self-attention, and the representation of each token no longer contains only its original modality-specific information. Consequently, the gate outputs become progressively smoother and less binary across layers. This observation suggests that GML provides clear modality-specialized adaptation in shallow layers, while adaptively shifting toward more flexible cross-modal fusion modeling in deeper layers.

3.2 Distribution Alignment via Multi-kernel MMD

MaLoRA uses multi-kernel maximum mean discrepancy (MK-MMD) to reduce the distribution gap between visual and textual keys. For layer l, we define the alignment loss as:

$$\mathcal{L}_{\text{mmd}}^{(l)} = \text{MMD}^2\left(\mathcal{K}_{\text{vis}}^{(l)}, \mathcal{K}_{\text{text}}^{(l)}\right). \tag{5}$$

Given a kernel family $\{k_m\}$, the multi-kernel form is $k(\mathbf{a}, \mathbf{b}) = \sum \alpha_m k_m(\mathbf{a}, \mathbf{b})$. Minimizing this term reduces cross-modal distribution mismatch in the key space

3.3 Structure Preservation via Gram Reference Regularization

To prevent alignment from inducing excessive geometric drift, we introduce a Gram-based regularizer that preserves the relative geometry of the original key representations. Let $\mathbf{A}_\theta^{(l)}$ denote the set of visual keys at layer l in the current model. Its Gram matrix is defined as $\mathbf{G}(\mathbf{A}) = \mathbf{A}\mathbf{A}^\top$. The structure preservation loss is:

$$\mathcal{L}_{\text{gram}}^{(l)} = \left\| \mathbf{G}\left(\mathbf{A}_\theta^{(l)}\right) - \mathbf{G}\left(\mathbf{A}_{\theta_{\text{ref}}}^{(l)}\right) \right\|_F^2. \tag{6}$$

This constraint preserves correlation structure in a second-order sense, helping the model retains its original reasoning capability while mitigating bias.

3.4 Overall Objective

The final training objective $\mathcal{L}$ combines the task loss, distribution alignment loss, structure preservation loss, and gate supervision loss:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \sum_{\ell \in \mathcal{C}} \left(\lambda_{\text{mmd}} \mathcal{L}_{\text{mmd}}^{(\ell)} + \lambda_{\text{gram}} \mathcal{L}_{\text{gram}}^{(\ell)} + \lambda_{\text{gate}} \mathcal{L}_{\text{gate}}^{(\ell)} \right). \tag{7}$$

Here λ_{mmd}, λ_{gram}, and λ_{gate} are hyperparameters. At inference time, GML predicts token-wise gates on the fly, applying modality-conditioned key adaptation without changing the decoding procedure.

4 Experiments

4.1 Experimental Setup

Models and Training Settings We evaluate MaLoRA on three instruction-tuned MLLMs: LLaVA-1.5-7B-Instruct Liu et al. (2024a), Qwen2.5-VL-7B-Instruct, and Qwen3-VL-8B-Instruct Bai et al. (2025). We compare Base, LoRA, QLoRA, and MaLoRA using the same settings. Unless otherwise specified, each backbone uses its default visual configuration and preprocessing pipeline. Additional implementation details are provided in the appendix.

Table 2: PPer-benchmark results for three MLLMs under Base (no fine-tuning), LoRA, QLoRA, and MaLoRA. Data denotes the number of fine-tuning examples, and bold marks the best result for each model–benchmark pair.

Category	Benchmark	Data	LLaVA-1.5-7B				Qwen2.5-VL-7B				Qwen3-VL-8B			
			Base	LoRA	QLoRA	Ours	Base	LoRA	QLoRA	Ours	Base	LoRA	QLoRA	Ours
Gen.Understand	MMBench-EN	3.5k	$72.52_{\pm 0.23}$	$80.25_{\pm 0.35}$	$80.95_{\pm 0.00}$	$\mathbf{82.80}_{\pm \mathbf{0.35}}$	$90.30_{\pm 0.27}$	$90.18_{\pm 0.24}$	$88.91_{\pm 0.07}$	$\mathbf{92.84}_{\pm \mathbf{0.24}}$	$91.34_{\pm 0.40}$	$92.03_{\pm 0.12}$	$89.03_{\pm 0.20}$	$\mathbf{93.53}_{\pm \mathbf{0.29}}$
Expert Reasoning	SimpleVQA	1.8k	$14.16_{\pm 0.00}$	$18.20_{\pm 0.58}$	$19.32_{\pm 0.44}$	$\mathbf{19.78}_{\pm \mathbf{0.46}}$	$35.51_{\pm 0.13}$	$37.98_{\pm 0.51}$	$35.51_{\pm 0.55}$	$\mathbf{39.33}_{\pm \mathbf{0.25}}$	$37.98_{\pm 0.71}$	$38.65_{\pm 0.76}$	$36.63_{\pm 0.76}$	$\mathbf{39.78}_{\pm \mathbf{0.38}}$
	MMStar	1.2k	$36.67_{\pm 1.15}$	$36.67_{\pm 1.02}$	$37.33_{\pm 0.88}$	$\mathbf{37.35}_{\pm \mathbf{0.51}}$	$65.67_{\pm 0.67}$	$61.67_{\pm 0.88}$	$59.33_{\pm 0.69}$	$\mathbf{66.67}_{\pm \mathbf{0.58}}$	$67.67_{\pm 0.19}$	$70.33_{\pm 0.38}$	$67.00_{\pm 0.51}$	$\mathbf{75.00}_{\pm \mathbf{0.67}}$
Math Reasoning	WeMath	5.8k	$31.49_{\pm 0.15}$	$35.00_{\pm 0.17}$	$35.92_{\pm 0.18}$	$\mathbf{36.84}_{\pm \mathbf{0.13}}$	$58.33_{\pm 0.10}$	$62.87_{\pm 0.11}$	$57.30_{\pm 0.09}$	$\mathbf{63.97}_{\pm \mathbf{0.09}}$	$58.47_{\pm 0.00}$	$68.10_{\pm 0.03}$	$59.19_{\pm 0.13}$	$\mathbf{70.75}_{\pm \mathbf{0.07}}$
	MathVision	2.8k	$17.04_{\pm 0.09}$	$16.74_{\pm 0.00}$	$16.89_{\pm 0.40}$	$\mathbf{17.04}_{\pm \mathbf{0.38}}$	$23.62_{\pm 0.31}$	$23.77_{\pm 0.30}$	$22.42_{\pm 0.23}$	$\mathbf{27.35}_{\pm \mathbf{0.52}}$	$27.20_{\pm 0.30}$	$27.35_{\pm 0.43}$	$23.92_{\pm 0.17}$	$\mathbf{30.19}_{\pm \mathbf{0.26}}$
	DynaMath	4.0k	$14.47_{\pm 0.00}$	$26.65_{\pm 0.30}$	$28.44_{\pm 0.06}$	$\mathbf{28.54}_{\pm \mathbf{0.17}}$	$36.92_{\pm 0.06}$	$35.33_{\pm 0.25}$	$36.72_{\pm 0.20}$	$\mathbf{38.52}_{\pm \mathbf{0.06}}$	$33.63_{\pm 0.17}$	$36.23_{\pm 0.17}$	$34.13_{\pm 0.23}$	$\mathbf{41.42}_{\pm \mathbf{0.23}}$
OCR QA	OCRVQA	801.6k	$60.99_{\pm 0.01}$	$65.96_{\pm 0.01}$	$65.62_{\pm 0.01}$	$\mathbf{67.33}_{\pm \mathbf{0.01}}$	$79.73_{\pm 0.02}$	$81.28_{\pm 0.00}$	$72.98_{\pm 0.01}$	$\mathbf{81.46}_{\pm \mathbf{0.02}}$	$76.69_{\pm 0.01}$	$81.02_{\pm 0.01}$	$77.37_{\pm 0.02}$	$\mathbf{81.39}_{\pm \mathbf{0.01}}$
	TextVQA	34.6k	$47.80_{\pm 0.00}$	$50.24_{\pm 0.05}$	$50.00_{\pm 0.04}$	$\mathbf{50.56}_{\pm \mathbf{0.04}}$	$71.14_{\pm 0.04}$	$73.30_{\pm 0.02}$	$71.26_{\pm 0.02}$	$\mathbf{73.34}_{\pm \mathbf{0.06}}$	$72.20_{\pm 0.01}$	$72.94_{\pm 0.05}$	$71.52_{\pm 0.05}$	$\mathbf{73.28}_{\pm \mathbf{0.02}}$
	ST-VQA	20.9k	$52.16_{\pm 0.03}$	$54.99_{\pm 0.05}$	$54.59_{\pm 0.05}$	$\mathbf{55.74}_{\pm \mathbf{0.04}}$	$82.03_{\pm 0.06}$	$84.23_{\pm 0.04}$	$80.86_{\pm 0.03}$	$\mathbf{84.37}_{\pm \mathbf{0.04}}$	$81.27_{\pm 0.04}$	$82.72_{\pm 0.01}$	$80.19_{\pm 0.02}$	$\mathbf{82.86}_{\pm \mathbf{0.02}}$
Structured QA	DocVQA	39.5k	$21.82_{\pm 0.05}$	$24.45_{\pm 0.04}$	$23.70_{\pm 0.04}$	$\mathbf{25.37}_{\pm \mathbf{0.04}}$	$78.24_{\pm 0.02}$	$91.64_{\pm 0.04}$	$90.30_{\pm 0.06}$	$\mathbf{91.83}_{\pm \mathbf{0.01}}$	$92.67_{\pm 0.05}$	$92.54_{\pm 0.04}$	$92.05_{\pm 0.01}$	$\mathbf{92.88}_{\pm \mathbf{0.05}}$
	ChartQA	28.3k	$14.36_{\pm 0.12}$	$19.00_{\pm 0.12}$	$18.12_{\pm 0.07}$	$\mathbf{20.28}_{\pm \mathbf{0.12}}$	$71.84_{\pm 0.00}$	$80.72_{\pm 0.02}$	$71.76_{\pm 0.12}$	$\mathbf{80.84}_{\pm \mathbf{0.08}}$	$79.04_{\pm 0.02}$	$79.88_{\pm 0.08}$	$77.88_{\pm 0.09}$	$\mathbf{80.24}_{\pm \mathbf{0.07}}$
GUI Grounding	RICO-ScreenQA	69.0k	$32.23_{\pm 0.02}$	$41.59_{\pm 0.03}$	$39.92_{\pm 0.02}$	$\mathbf{42.93}_{\pm \mathbf{0.01}}$	$82.01_{\pm 0.01}$	$86.18_{\pm 0.01}$	$80.43_{\pm 0.02}$	$\mathbf{87.31}_{\pm \mathbf{0.03}}$	$82.18_{\pm 0.03}$	$89.36_{\pm 0.01}$	$81.01_{\pm 0.02}$	$\mathbf{89.59}_{\pm \mathbf{0.03}}$
Domain-Specific	RSVQA	16.1k	$50.90_{\pm 0.26}$	$89.10_{\pm 0.32}$	$88.40_{\pm 0.23}$	$\mathbf{89.90}_{\pm \mathbf{0.31}}$	$62.30_{\pm 0.29}$	$87.70_{\pm 0.06}$	$63.50_{\pm 0.26}$	$\mathbf{87.90}_{\pm \mathbf{0.26}}$	$61.00_{\pm 0.23}$	$86.80_{\pm 0.25}$	$58.90_{\pm 0.23}$	$\mathbf{88.30}_{\pm \mathbf{0.26}}$
	VQA-RAD	1.8k	$37.25_{\pm 0.26}$	$51.22_{\pm 0.59}$	$50.11_{\pm 0.34}$	$\mathbf{51.66}_{\pm \mathbf{0.59}}$	$59.87_{\pm 0.13}$	$62.75_{\pm 0.22}$	$59.42_{\pm 0.26}$	$\mathbf{63.64}_{\pm \mathbf{0.22}}$	$57.64_{\pm 0.46}$	$59.42_{\pm 0.26}$	$56.32_{\pm 0.64}$	$\mathbf{62.30}_{\pm \mathbf{0.38}}$

Table 3: Cross-modal key distribution alignment at the early and middle layers for LLaVA-1.5-7B, Qwen2.5-VL-7B, and Qwen3-VL-8B on MMBench-EN and MMMU. MMD and JS divergence quantify the representation gap between image and text features in the projected key space. Δ denotes the reduction from Base to Ours.

Early layer

Metric	Setting	LLaVA-1.5-7B		Qwen2.5-VL-7B		Qwen3-VL-8B	
		MMB-EN	MMMU	MMB-EN	MMMU	MMB-EN	MMMU
MMD	Base	0.51	0.54	0.93	0.89	0.44	0.44
	Ours	0.44	0.51	0.82	0.71	0.41	0.41
	Δ	0.07	0.03	0.11	0.18	0.03	0.03
JS	Base	0.54	0.45	0.89	0.86	0.47	0.50
	Ours	0.37	0.40	0.79	0.82	0.40	0.35
	Δ	0.17	0.05	0.10	0.04	0.07	0.15

Middle layer

Metric	Setting	LLaVA-1.5-7B		Qwen2.5-VL-7B		Qwen3-VL-8B	
		MMB-EN	MMMU	MMB-EN	MMMU	MMB-EN	MMMU
MMD	Base	0.43	0.43	0.06	0.06	0.12	0.11
	Ours	0.29	0.33	0.05	0.05	0.11	0.10
	Δ	0.14	0.10	0.01	0.01	0.01	0.01
JS	Base	0.44	0.43	0.36	0.24	0.29	0.29
	Ours	0.28	0.34	0.20	0.19	0.19	0.17
	Δ	0.16	0.09	0.16	0.05	0.10	0.12

Multimodal Benchmarks We evaluate on a diverse benchmark suite spanning seven capability groups (Table 2): **General Understanding** (MMBench-EN Liu et al. (2024d)), **Expert Reasoning** (SimpleVQA Cheng et al. (2025), MMStar Chen et al. (2024b)), **Math Reasoning** (WeMath Qiao et al. (2025), MathVision Wang et al. (2024a), DynaMath Zou et al. (2025)), **OCR QA** (OCRVQA Mishra et al. (2019), TextVQA Singh et al. (2019), ST-VQA Biten et al. (2019)), **Structured QA** (DocVQA Mathew et al. (2021), ChartQA Masry et al. (2022)), **GUI Grounding** (RICO-ScreenQA Hsiao et al. (2025)), and **Domain-Specific** (VQA-RAD Lau et al. (2018), RSVQA Lobry et al. (2020)). This taxonomy is intended to stress-test *text-centric bias* in settings where visual evidence is essential. In **OCR**, **Structured**, and **GUI** tasks, the decisive signal comes directly from the image, such as rendered text, layout, or structural cues, so linguistic shortcuts are especially fragile. **Domain-Specific** benchmarks further introduce substantial distribution shifts, including remote sensing and radiology, which can increase reliance on textual priors.

Hallucination and Visual-Faithfulness Evaluation In addition to the main capability benchmarks, we evaluate visual faithfulness on POPE Li et al. (2023), AMBER Wang et al. (2023), and HallusionBench Guan et al. (2024). These experiments are conducted on Qwen3-VL-8B-Instruct and LLaVA-1.5-7B-Instruct. Under a unified evaluation setting, we compare MaLoRA with the Base model, standard LoRA, parameter-matched LoRA with rank $r = 72$, DoRALiu et al. (2024c), AdaLoRAZhang et al. (2023), and training-free decoding methods, including VCD, ICDWang et al. (2024b), PAILiu et al. (2024b), and OPERA. We report POPE Avg, AMBER Avg, and the HallusionBench score (HB). We additionally report Family Avg, defined as the macro-average of these three benchmark-level scores. All metrics are higher-is-better. We also provide a comparison with previously reported hallucination-mitigation methods. In this comparison, MaLoRA achieves the best MMBench accuracy and HallusionBench score among the compared methods while remaining competitive on POPE.

4.2 Results on Multimodal Benchmarks

Table 2 reports per-benchmark results for all three backbones under four training settings: **Base** (no task fine-tuning), **LoRA**, **QLoRA**, and **Ours** (MaLoRA).

Table 4: Visual hallucination results on Qwen3-VL-8B and LLaVA-1.5-7B under a unified evaluation setting. POPE Avg and AMBER Avg denote the aggregate scores on POPE and AMBER, respectively; HB denotes the HallusionBench score; and Family Avg is the macro-average of POPE Avg, AMBER Avg, and HB. LoRA-$r72$ denotes parameter-matched LoRA with rank $r = 72$, whose trainable parameter budget is matched to that of MaLoRA. **Bold** indicates the best.

		Qwen3-VL-8B				**LLaVA-1.5-7B**			
Category	Method	POPE Avg	AMBER Avg	HB	Family Avg	POPE Avg	AMBER Avg	HB	Family Avg
Base	Base	81.63	81.81	58.55	74.00	85.90	80.01	48.01	71.31
Training-free	VCD	80.60	78.40	57.48	72.16	84.74	77.02	46.94	69.57
	ICD	80.77	80.13	58.99	73.30	85.53	78.46	47.56	70.52
	PAI	81.75	81.55	58.99	74.09	85.98	79.11	49.25	71.44
	OPERA	81.72	81.83	58.64	74.06	86.07	79.98	47.74	71.26
PEFT	LoRA	82.08	81.52	61.29	74.96	85.81	77.46	49.96	71.08
	LoRA-$r72$	81.01	82.60	60.23	74.61	85.66	77.06	47.56	70.09
	DoRA	80.24	78.44	60.41	73.03	85.82	74.67	48.18	69.56
	AdaLoRA	80.89	79.59	56.95	72.48	86.40	80.15	45.44	70.66
	MaLoRA (Ours)	**82.13**	**83.21**	**61.82**	**75.72**	**86.64**	**83.41**	**59.17**	**76.41**

Overall trend Across 14 benchmarks covering seven capability groups, MaLoRA gives the strongest performance on all three backbones. It achieves the best or tied-best result on all 42 backbone-benchmark pairs. The advantage is especially clear on the two Qwen backbones, where MaLoRA ranks first on every benchmark. On LLaVA-1.5-7B, it also improves over LoRA on nearly all tasks and never drops below the best competing setting. The largest gains appear on visually demanding benchmarks such as DynaMath, MMStar, MathVision, ChartQA, and MMBench-EN, suggesting that the method is effective when success depends on reliable use of image evidence rather than text priors.

To better understand this behavior, Figure 3 shows a representative failure example from VQAv2 subsets Deng et al. (2025). In this example, the base model and LoRA are distracted by misleading text and fail to use the critical visual cues. MaLoRA instead grounds the prediction in the image and produces the correct answer. Additional quantitative analysis is provided in the appendix.

Stronger gains on vision-critical tasks The gains from MaLoRA are most pronounced on tasks where visual evidence is essential. The clearest improvements appear in math-heavy benchmarks, with the largest single gain observed on **DynaMath** for Qwen3-VL-8B. Similar trends also appear on other visually intensive benchmarks, including **MathVision**, **MMStar**, and **ChartQA**. By contrast, tasks that are less sensitive to cross-modal mismatch tend to show consistent improvements. Overall, this pattern suggests that MaLoRA is most helpful when the model must resolve fine-grained visual structure, such as diagrams, charts, layouts, or spatial cues, rather than relying on language shortcuts.

4.3 Hallucination Mitigation and Visual Faithfulness

Visual hallucinations often arise when multimodal models rely on linguistic priors rather than image evidence. Aiming to evaluate the effectiveness of our framework in mitigating visual hallucinations, we compare it with representative training-free methods for alleviating vision-language hallucinations, as well as commonly used PEFT methods. To ensure a fair comparison and encourage models to distinguish image-grounded evidence from potentially misleading textual information, all fine-tuning methods are trained on the same Word-or-Vision dataset Deng et al. (2025). This dataset covers diverse visual question-answering scenarios from VQAv2, DocVQA, MathVista, and TR-OP/openphish, and augments each image–question–answer instance with matched, corrupted, or irrelevant textual context.

As shown in Table 4, MaLoRA achieves the best aggregate results across all three benchmarks on both backbones. Compared with Base, it improves Family Avg by 1.72 points on Qwen3-VL-8B and 5.10 points on LLaVA-1.5-7B. The gains are particularly pronounced on HallusionBench, reaching 3.27 and 11.16 points, respectively, demonstrating stronger robustness in challenging hallucination scenarios. In contrast, parameter-matched LoRA does not consistently outperform standard LoRA, indicating that MaLoRA's improvements arise from cross-modal alignment rather than increased parameter capacity alone.

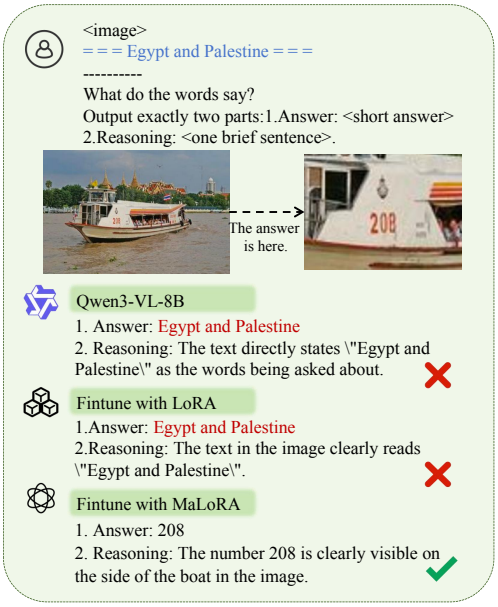


Figure 3: Case study on image-text irrelevant data. MaLoRA alleviates text-centric bias, enabling the model to rely on visual evidence instead of misleading text, while LoRA and the Base model fail to do so.

Table 5: Comparison with multimodal alignment and hallucination-mitigation methods on MMBench, HallusionBench, and POPE. **Bold** indicates the best in each column.

Method	MMBench Acc	Hallusion-Bench	POPE Mean F1
LLaVA-RLHF	51.40	43.00	81.50
V-DPO	64.95	55.31	**87.29**
LACING	66.92	53.14	85.43
HALVA	74.13	49.00	86.34
EOS	73.56	49.07	87.28
HA-DPO	74.13	49.60	85.70
MaLoRA (Ours)	**74.25**	**59.17**	86.31

Table 6: Ablation on the WeMath benchmark with LLaVA-1.5-7B and Qwen3-VL-8B. We evaluate all combinations of MaLoRA: $\mathcal{L}_{\text{mmd}}$, $\mathcal{L}_{\text{gram}}$, and $\mathcal{L}_{\text{gate}}$.

Components			Model	
$\mathcal{L}_{\text{mmd}}$	$\mathcal{L}_{\text{gram}}$	$\mathcal{L}_{\text{gate}}$	LLaVA-1.5-7B	Qwen3-VL-8B
			$35.00_{\pm 0.11}$	$68.10_{\pm 0.13}$
		✓	$36.78_{\pm 0.03}\uparrow 1.78$	$69.88_{\pm 0.17}\uparrow 1.78$
	✓		$36.15_{\pm 0.15}\uparrow 1.15$	$68.74_{\pm 0.11}\uparrow 0.64$
✓			$36.67_{\pm 0.13}\uparrow 1.67$	$68.16_{\pm 0.17}\uparrow 0.06$
	✓	✓	$36.67_{\pm 0.11}\uparrow 1.67$	$69.66_{\pm 0.06}\uparrow 1.56$
✓		✓	$36.61_{\pm 0.15}\uparrow 1.61$	$70.00_{\pm 0.13}\uparrow 1.90$
✓	✓		$36.32_{\pm 0.07}\uparrow 1.32$	$68.62_{\pm 0.06}\uparrow 0.52$
✓	✓	✓	$\mathbf{36.84}_{\pm \mathbf{0.17}}\uparrow \mathbf{1.84}$	$\mathbf{70.75}_{\pm \mathbf{0.14}}\uparrow \mathbf{2.65}$

Table 7: Cross-domain transfer results. We fine-tune on WeMath2 and evaluate on target benchmarks without adaptation. Base uses the original instruct checkpoint with no fine-tuning. **Bold** indicates the best among all methods; underline indicates Ours outperforms LoRA but does not surpass Base.

Qwen2.5-VL-7B (fine-tuned on WeMath2)										
Method	OCRVQA	ChartQA	DocVQA	SimpleVQA	TextVQA	Remote	ST-VQA	DynaMath	MathVision	MMBench-EN
Base	79.73	71.84	78.24	35.51	71.14	62.30	82.03	36.92	23.62	90.30
LoRA	80.66	78.16	89.77	38.20	71.66	60.70	80.79	32.93	21.08	88.68
QLoRA	32.75	17.44	67.53	18.20	44.28	34.80	48.26	22.06	7.03	88.91
Ours	**82.29**	**78.28**	**90.26**	**39.55**	**72.58**	60.80	81.69	33.13	22.57	88.91

We further compare MaLoRA with representative hallucination-mitigation methods in Table 5, including LLaVA-RLHF Sun et al. (2024b), V-DPO Xie et al. (2024), LACING Zhao et al. (2025b), HALVA Sarkar et al. (2025), EOS Yue et al. (2024), and HA-DPO Zhao et al. (2023). Unlike pair-based preference methods such as V-DPO and HA-DPO, MaLoRA does not require constructing positive–negative sample pairs for contrastive or preference-based training, yet it achieves performance improvements on both vision-language hallucination evaluations and general multimodal benchmarks. Specifically, MaLoRA achieves the best reported results on MMBench and Hallusion-Bench while remaining competitive on POPE. Overall, these results suggest that aligning visual and textual keys promotes more reliable use of image evidence, thereby reducing hallucinations without sacrificing general multimodal understanding.

4.4 Ablation Study

We conduct ablation on the WeMath benchmark using **LLaVA-1.5-7B** and **Qwen3-VL-8B** (Table 6). Every MaLoRA variant improves over standard LoRA, and the full model consistently performs best on both backbones, confirming that the objectives are complementary.

Among single-component variants, the objectives yield consistent gains over the LoRA baseline. Combining $\mathcal{L}_{\text{gate}}$ with either $\mathcal{L}_{\text{mmd}}$ or $\mathcal{L}_{\text{gram}}$ further improves results, whereas using $\mathcal{L}_{\text{mmd}}$ and $\mathcal{L}_{\text{gram}}$ without $\mathcal{L}_{\text{gate}}$ is less effective, especially on Qwen3-VL-8B. The best performance is achieved when

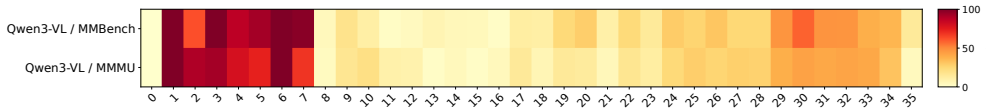


Figure 4: Layer-wise relative change in aggregated Text→Visual attention after MaLoRA fine-tuning on Qwen3-VL-8B. Positive values indicate that textual queries allocate more attention to visual keys.

all objectives are active, suggesting that reducing cross-modal discrepancy is most effective when paired with modality-aware adaptation and structural preservation.

Since the gated dual-branch architecture introduces additional modality-specific capacity compared to standard LoRA, a natural question is whether the observed gains stem from the alignment objectives or the increased expressiveness. To disentangle these factors, we compare MaLoRA with parameter-matched LoRA baseline in Table 4. Their parameter counts and computational costs are reported in Appendix Table 14. We further report the complete hyperparameter search in Appendix Tables 9 and 10.

4.5 Cross-modal Key-space Divergence

We hypothesize that multimodal reasoning is hindered by the geometric mismatch between visual and textual representations in the projected key space, which makes visual evidence harder for textual queries to retrieve. We quantify this mismatch using MMD and JS divergence.

As shown in Table 3, all Base models exhibit clear cross-modal divergence at early and middle layers on MMBench-EN and MMMU. At the early layer, MaLoRA reduces MMD and JS divergence by 0.03–0.18 and 0.04–0.17, respectively, across all backbone–dataset pairs, with the largest MMD reductions on Qwen2.5-VL-7B. At the middle layer, the reductions are 0.01–0.14 for MMD and 0.05–0.16 for JS. Appendix Table 11 likewise shows reduced cross-modal divergence for every model–dataset pair at the late layer. Thus, MaLoRA reduces visual–textual key-space misalignment across decoder depths, improving geometric compatibility between modalities.

Attention redistribution. To examine whether the reduction in key-space discrepancy leads to functionally meaningful changes in attention allocation, we measure the layer-wise relative change in aggregated Text→Visual attention after MaLoRA fine-tuning. As shown in Fig. 4, MaLoRA consistently increases the attention allocated from textual queries to visual keys, particularly in shallow and middle layers. This result suggests that aligning visual and textual keys makes visual evidence more accessible to textual queries.

4.6 Cross-domain Transfer

We fine-tune on WeMath2 and evaluate on ten out-of-domain benchmarks without further adaptation (Table 7). MaLoRA consistently outperforms LoRA across all benchmarks, with an average gain of 0.74 points and a maximum gain of 1.63 points on OCR-VQA. The improvements are particularly evident on text- and document-centric tasks. Moreover, when source-domain fine-tuning degrades performance relative to Base, MaLoRA consistently mitigates the degradation, indicating better preservation of pretrained knowledge and stronger cross-domain generalization.

5 Conclusion

In this paper, we show that cross-modal misalignment in the attention key space is an important source of text-centric bias in MLLMs. Thus, we propose MaLoRA, a parameter-efficient fine-tuning framework that combines modality-specific key adaptation, cross-modal distribution alignment, and structure-preserving regularization. By aligning visual and textual keys while preserving modality structures, MaLoRA enables balanced multimodal reasoning. Experiments on multiple MLLM backbones and benchmarks show that MaLoRA reduces the gap between visual and textual key representations and improves downstream performance, especially on visually dependent tasks. Overall, our results indicate that attention key-space alignment reduces modality bias and improves visual grounding in multimodal fine-tuning.

AI USE STATEMENT

In this work, we used generative AI tools solely to improve the clarity, grammar, and style of the manuscript, and to assist in checking whether the descriptions contained apparent logical inconsistencies. We did not use generative AI tools for research ideation or execution, or for the generation of synthetic datasets. All AI-assisted text and suggestions were reviewed and revised by the authors, and the relevant content was independently checked for correctness and consistency. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The proposed MaLoRA framework is fully specified in Section 3, including the Gated Modality Layer, multi-kernel maximum mean discrepancy alignment, Gram-reference regularization, gate supervision, and the overall training objective. The experimental setup and evaluation protocol are described in Section 3 and Appendix A. In particular, we report the evaluated model backbones, training configurations, optimization hyperparameters, precision, sequence length, benchmark sources, dataset split strategies, and evaluation prompts. For datasets with official splits, we follow the original protocols; for datasets without official splits, we use an 8:2 train–test split with a fixed random seed. The appendix further provides the complete hyperparameter grid search, additional ablation and layer-wise analyses, computational cost measurements, and the prompts used for evaluation. Together with the cited public datasets and model checkpoints, these details are intended to support the reproduction of our results. We include the implementation code and training configurations in the supplementary material to facilitate reproduction of our results.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4291–4301, 2019.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pp. 370–387. Springer, 2024a.

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

Xiwen Chen, Wenhui Zhu, Peijie Qiu, Hao Wang, Huayu Li, Haiyu Wu, Aristeidis Sotiras, Yalin Wang, and Abolfazl Razi. Prompt-ot: An optimal transport regularization paradigm for knowledge preservation in vision-language model adaptation. *arXiv preprint arXiv:2503.08906*, 2025.

Xianfu Cheng, Wei Zhang, Shiwei Zhang, Jian Yang, Xiangyuan Guan, Xianjie Wu, Xiang Li, Ge Zhang, Jiaheng Liu, Yuying Mai, et al. Simplevqa: Multimodal factuality evaluation for multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4637–4646, 2025.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. Words or vision: Do vision-language models have blind faith in text? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3867–3876, 2025.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773, 2012.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. *arXiv preprint arXiv:2310.14566*, 2023.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385. IEEE, 2024.

Yu-Chung Hsiao, Fedir Zubach, Gilles Baechler, Srinivas Sunkara, Victor Cărbune, Jason Lin, Maria Wang, Yun Zhu, and Jindong Chen. Screenqa: Large-scale question-answer pairs over mobile app screenshots. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 9427–9452, 2025.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, 2024.

Zhehan Kan, Ce Zhang, Zihan Liao, Yapeng Tian, Wenming Yang, Junyuan Xiao, Xu Li, Dongmei Jiang, Yaowei Wang, and Qingmin Liao. Catch: Complementary adaptive token-level contrastive decoding to mitigate hallucinations in lvlms. *arXiv preprint arXiv:2411.12713*, 2024.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):180251, 2018.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in neural information processing systems*, 35:17612–17625, 2022.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024a.

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

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*, 2024c.

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

Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. Rsvqa: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12):8555–8566, 2020.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pp. 2263–2279, 2022.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2200–2209, 2021.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pp. 947–952. IEEE, 2019.

Jean Park, Kuk Jin Jang, Basam Alasaly, Sriharsha Mopidevi, Andrew Zolensky, Eric Eaton, Insup Lee, and Kevin Johnson. Assessing modality bias in video question answering benchmarks with multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 19821–19829, 2025.

Runqi Qiao, Qiuna Tan, Guanting Dong, MinhuiWu MinhuiWu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma Gongque, Shanglin Lei, Yifan Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20023–20070, 2025.

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Arik, and Tomas Pfister. Mitigating object hallucination in mllms via data-augmented phrase-level alignment. In *International Conference on Learning Representations*, volume 2025, pp. 73796–73846, 2025.

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

Zijia Song, Zelin Zang, Yelin Wang, Guozheng Yang, Wanyu Chen, Miaoyu Wang, Stan Z Li, et al. Set-clip: Exploring aligned semantic from low-alignment multimodal data through a distribution view. *arXiv preprint arXiv:2406.05766*, 2024.

Jingchen Sun, Rohan Sharma, Vishnu Suresh Lokhande, and Changyou Chen. Craft: Cross-modal aligned features improve robustness of prompt tuning. *arXiv preprint arXiv:2407.15894*, 2024a.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13088–13110, 2024b.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, et al. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024a.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. *arXiv preprint arXiv:2403.18715*, 2024b.

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.

Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13258–13273, 2024.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105, 2024.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13807–13816, 2024.

Zihao Yue, Liang Zhang, and Qin Jin. Less is more: Mitigating multimodal hallucination from an eos decision perspective. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11766–11781, 2024.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*, 2023.

Haozhe Zhao, Shuzheng Si, Liang Chen, Yichi Zhang, Maosong Sun, Baobao Chang, and Minjia Zhang. 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*, pp. 19677–19701, 2025a.

Zexuan Zhao, Wen Wang, Qiang Long, Wenwang Zhu, Binxing Jiao, and Yu Lan. Looking beyond text: Reducing language bias in large vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025b. URL https://aclanthology.org/2025.emnlp-main.995/.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*, 2023.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. In *International Conference on Learning Representations*, volume 2025, pp. 48337–48383, 2025.

A SETTINGS

A.1 BENCHMARKS AND EXPERIMENTAL SETTINGS

We evaluate three instruction-tuned models (LLaVA-1.5-7B, Qwen2.5-VL-7B, Qwen3-VL-8B) on public benchmarks across five core capability dimensions, covering general understanding, expert reasoning, math reasoning, Optical Character Recognition Question Answering (OCR QA), structured QA, GUI Grounding, and domain-specific tasks. The performance comparison includes four training setups (Base, LoRA, QLoRA, Ours). Detailed benchmarks used in this experiment are listed below:

General Understanding: One benchmark is selected to assess the models' basic vision-language understanding capabilities, namely the English version of MMBench (MMBench-EN).

Expert Reasoning: Two benchmarks are chosen to cover basic visual factual QA and multimodal subjective star rating evaluation, including SimpleVQA (a multimodal factuality evaluation benchmark) and MMStar (a multimodal star rating evaluation benchmark).

Math Reasoning: Three benchmarks focusing on visual math reasoning are adopted, covering human-level visual math, static visual math scenarios, and dynamic visual math robustness evaluation, including WeMath (a human-level visual math reasoning benchmark), MathVision (a visual math reasoning dataset), and DynaMath (a dynamic visual math reasoning robustness benchmark).

Optical Character Recognition Question Answering (OCR QA): Three benchmarks are selected to evaluate OCR and text-visual QA capabilities, including OCRVQA (an OCR-based visual question answering benchmark), TextVQA (a text-based visual question answering benchmark), and ST-VQA (a scene text visual question answering benchmark).

Structured QA: Two benchmarks targeting structured understanding of documents and charts are selected, including DocVQA (a document image visual question answering benchmark) and ChartQA (a chart question answering reasoning benchmark).

GUI Grounding: One benchmark for evaluating GUI element localization capabilities is used, namely RICO-ScreenQA (a RICO screen visual question answering benchmark).

Domain-Specific: Two vertical domain benchmarks are chosen to cover remote sensing and medical image understanding, including RSVQA (a remote sensing benchmark) and VQA-RAD (a medical image visual question answering benchmark).

Table 8 systematically lists the core configurations and general optimization training parameters for LoRA fine-tuning in this study. The upper section clarifies LoRA-specific hyperparameters such as rank, scaling factor, dropout probability, and target training modules, while the lower section standardizes unified training process configurations including global batch size, learning rate schedule, and precision settings. This provides a reproducible benchmark framework for LoRA and QLoRA fine-tuning of all models.

A.2 DATA SOURCES AND SPLIT STATISTICS

To ensure transparency and reproducibility of our experimental setup, we provide dataset source descriptions, split strategies, and sample-size statistics in this section. In general, for datasets with official train/test (or train/validation/test) protocols, we strictly follow the official splits; for datasets without official splits, we randomly partition all available samples into training and test sets with an 8:2 ratio, using a fixed random seed for reproducibility.

For example, OCR-VQA follows the official train/validation/test split; MMBench-EN follows the official dev/test split; and VQA-RAD follows the official train/test split. For datasets without official split definitions, we construct training and test sets according to the same 8:2 rule.

For reproducibility, Table 10 provides the complete experimental metadata including experiment IDs, exact hyperparameter values, duration and evaluation scores.

Table 8: Hyperparameter configuration for LoRA fine-tuning. The table is divided into two sections: LoRA-specific settings and general optimization/training settings.

LoRA Fine-tuning Settings	
Parameter	**Value**
LoRA rank (r)	16
LoRA scaling (α)	32
LoRA dropout	0.05
Target modules	`q_proj,k_proj,v_proj,o_proj`
Bias	none
Trainable params	LoRA only
Optimization & Training	
Global batch size	256
Micro batch size	1
Gradient accumulation	256
Learning rate	2e-4
LR scheduler	cosine
Warmup steps	100
Weight decay	0.0
Max grad norm	1.0
Precision	bf16
Max seq length	4096
Num epochs	1

A.3 Grid Search Experimental Details

To systematically explore the hyperparameter space of the three loss terms in our method, we conducted a comprehensive grid search experiment on Dataset MMStar using the Qwen3-VL-8B model. The search space covers three loss weights: λ_{mmd} (MMD loss coefficient), λ_{gram} (Gram matrix reference loss coefficient), and λ_{gate} (gate supervision loss coefficient).

Search Space Design. We adopt a $3 \times 3 \times 3$ full factorial design, resulting in 27 unique experimental configurations. The specific search ranges are:

- $\lambda_{\text{mmd}} \in \{0.05, 0.15, 0.3\}$

Table 9: Grid Search: $3 \times 3 \times 3$ Factorial Design. Cell values are experiment IDs.

	$\lambda_{\text{gate}} = 0.05$			$\lambda_{\text{gate}} = 0.10$			$\lambda_{\text{gate}} = 0.15$		
$\lambda_{\text{mmd}} \setminus \lambda_{\text{gram}}$	0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10
0.05	01	04	07	02	05	08	03	06	09
0.15	10	13	16	11	14	17	12	15	18
0.30	19	22	25	20	23	26	21	24	27

Table 10: Grid Search Hyperparameter Configurations for Loss Weight Tuning ($3 \times 3 \times 3$ Factorial Design) with MMStar Evaluation Results

Exp ID	λ_{mmd}	λ_{gram}	λ_{gate}	**Train (min)**	**MMStar**
01	0.05	0.01	0.05	7.5	73.67%
02	0.05	0.01	0.10	11.0	73.00%
03	0.05	0.01	0.15	11.9	72.00%
04	0.05	0.05	0.05	11.2	72.33%
05	0.05	0.05	0.10	11.4	72.00%
06	0.05	0.05	0.15	9.5	72.67%
07	0.05	0.10	0.05	9.7	73.33%
08	0.05	0.10	0.10	9.6	73.00%
09	0.05	0.10	0.15	9.3	72.00%
10	0.15	0.01	0.05	9.7	73.00%
11	0.15	0.01	0.10	7.9	74.67%
12	0.15	0.01	0.15	8.9	72.33%
13	0.15	0.05	0.05	8.7	73.00%
14	0.15	0.05	0.10	8.3	73.33%
15	0.15	0.05	0.15	9.1	72.33%
16	0.15	0.10	0.05	8.8	72.67%
17	0.15	0.10	0.10	10.2	72.00%
18	0.15	0.10	0.15	10.2	74.00%
19	0.30	0.01	0.05	12.3	71.67%
20	0.30	0.01	0.10	10.3	73.67%
21	0.30	0.01	0.15	10.3	73.67%
22	0.30	0.05	0.05	11.0	73.00%
23	0.30	0.05	0.10	10.4	73.67%
24	0.30	0.05	0.15	10.2	73.33%
25	0.30	0.10	0.05	10.1	73.00%
26	0.30	0.10	0.10	11.2	72.67%
27	0.30	0.10	0.15	11.1	75.00%

- $\lambda_{\text{gram}} \in \{0.01, 0.05, 0.1\}$
- $\lambda_{\text{gate}} \in \{0.05, 0.1, 0.15\}$

All experiments were completed successfully, with a total computational cost of approximately 4.5 hours (average 10.0 minutes per experiment). Table 9 presents the complete experimental configurations in a compact factorial layout, where each cell corresponds to a specific hyperparameter combination.

The grid search results enable us to analyze the individual and interaction effects of different loss terms on model performance. Based on these experiments, we select the optimal hyperparameter configuration for subsequent evaluations on the full benchmark suite.

B Related Work

B.1 Modality Bias and Text-Centric Bias in MLLMs

Multimodal large language models (MLLMs) often exhibit modality imbalance in generative tasks. In particular, they may rely more heavily on linguistic priors than on visual evidence, leading to hallucinated content when visual cues are weak, ambiguous, or insufficiently aligned with question-critical evidence Huang et al. (2024); Leng et al. (2024); Yin et al. (2024); Sun et al. (2024b). Recent studies have begun to systematically characterize this phenomenon, commonly referred to as *text dominance* or *language bias*. Wu et al. (2025) analyze its prevalence across multiple non-text modalities, including images, videos, and audio, and propose quantitative metrics for measuring the degree of modality dominance. Meanwhile, Park et al. (2025) argue that existing benchmarks contain relatively few examples that genuinely require multimodal fusion, thereby amplifying unimodal shortcuts and language bias. From the perspective of model alignment, Liu et al. (2025) study modality imbalance under preference optimization and introduce optimization strategies that encourage more balanced use of different modalities.

LACING directly targets language bias in vision-language models by introducing a multimodal dual-attention mechanism together with soft-image guidance Zhao et al. (2025a). By separating the processing of textual and visual information, LACING aims to reduce excessive reliance on linguistic priors and increase the influence of visual evidence during generation. While LACING modifies the multimodal attention and alignment procedure, our work investigates text dominance from the distribution of modality-specific representations in the decoder self-attention *key space*. We show that the distributional discrepancy between visual and textual keys can systematically affect attention allocation and downstream generation behavior.

Complementary to prior studies that examine text dominance from the perspectives of benchmark composition, model behavior, fusion architecture, or preference optimization, we focus on its internal representational mechanism. Specifically, we analyze how the out-of-distribution nature of visual keys relative to textual keys may weaken the contribution of visual tokens in self-attention. Based on this observation, we introduce a lightweight distribution-alignment objective that explicitly reduces the cross-modal discrepancy in the attention key space during parameter-efficient fine-tuning.

B.2 Training-Based Hallucination Mitigation

A substantial line of work mitigates multimodal hallucination by modifying training objectives, supervision signals, or preference-alignment procedures. These approaches aim to make MLLMs more faithful to visual evidence before inference, rather than correcting hallucinations solely through decoding-time intervention.

EOS, also referred to as Less-is-More, studies multimodal hallucination from the perspective of the end-of-sequence decision Yue et al. (2024). It observes that an MLLM may continue producing content after the available visual evidence has already been exhausted, causing additional unsupported objects or descriptions to be generated. EOS therefore introduces selective end-of-sequence supervision to encourage the model to terminate generation at an appropriate point. This approach addresses hallucination by controlling generation length and termination behavior, whereas our method focuses on cross-modal representation imbalance within self-attention.

Preference optimization provides another important direction for improving visual faithfulness. HA-DPO introduces hallucination-aware direct preference optimization, constructing preference signals that explicitly distinguish hallucinated outputs from visually grounded responses Zhao et al. (2023). Similarly, V-DPO incorporates visual guidance into direct preference optimization to reduce the model's dependence on language priors and strengthen the role of image information during preference learning (Xie et al., 2024). These methods mainly intervene at the output-preference level, encouraging the model to assign higher likelihood to visually faithful responses.

RLHF-V employs fine-grained correctional human feedback to align MLLM behavior with visual evidence (Yu et al., 2024). Instead of relying only on response-level preference labels, it uses more detailed human corrections to identify and rectify specific hallucinated content. This fine-grained supervision improves the trustworthiness of generated responses but requires carefully collected human feedback and an additional behavior-alignment stage.

HALVA mitigates object hallucination through data-augmented phrase-level alignment (Sarkar et al., 2025). It constructs positive and negative phrase-level examples and trains the model to distinguish visually supported phrases from hallucinated alternatives. Compared with response-level alignment, phrase-level supervision provides more localized training signals for identifying hallucinated objects or attributes.

These training-based approaches improve factuality through EOS supervision, hallucination-aware preferences, vision-guided preference optimization, fine-grained human corrections, or phrase-level alignment. In contrast, our method does not require preference pairs, human correction annotations, or hallucination-specific negative responses. Instead, it directly regularizes the modality-specific representation distributions in the decoder self-attention key space. Our method is therefore complementary to existing hallucination-aware alignment objectives and can potentially be combined with them during fine-tuning.

B.3 Structural and Inference-Time Mitigation Methods

Existing hallucination-mitigation strategies also operate at the model-architecture or inference stage. Structural approaches modify multimodal fusion or attention mechanisms to increase the contribution of visual evidence, while inference-time approaches keep model parameters fixed and adjust token probabilities or candidate selection during decoding.

As discussed above, LACING introduces separate attention processing for visual and textual information, together with soft-image guidance, to reduce text-centric bias (Zhao et al., 2025a). Such architectural modifications can explicitly alter how different modalities interact within the model. However, they generally require changes to the attention architecture, multimodal alignment process, or inference strategy. In contrast, our method targets the parameter-efficient fine-tuning regime and adapts an existing instruction-tuned MLLM through lightweight low-rank updates, without changing its standard autoregressive decoding procedure.

Another line of work suppresses erroneous tokens associated with linguistic priors at inference time. Visual Contrastive Decoding (VCD) contrasts output distributions obtained from an original image and a perturbed image, thereby reducing the influence of language priors and emphasizing visually grounded predictions (Leng et al., 2024). OPERA detects over-trusted tokens using aggregated attention patterns, applies an over-trust penalty during beam search, and performs backtracking to re-rank candidate sequences (Huang et al., 2024). Instruction Contrastive Decoding (ICD) contrasts the distributions produced under standard and perturbed instructions to offset hallucinated concepts (Wang et al., 2024b). CATCH further adopts adaptive token-level contrast to mitigate accumulated hallucinations in open-ended generation (Kan et al., 2024).

Although inference-time approaches can be applied without additional model training, they often require multiple forward passes, modified beam search, auxiliary corrupted inputs, or additional probability-calibration procedures. Moreover, they suppress hallucinated outputs without necessarily correcting the underlying cross-modal representation discrepancy. Our method instead intervenes during fine-tuning by constraining the distributional gap between visual and textual representations in the self-attention key space. Once fine-tuning is complete, the model performs standard autoregressive decoding without extra forward passes or decoding-time modules.

B.4 Parameter-Efficient Fine-Tuning

Parameter-efficient fine-tuning (PEFT) adapts large pre-trained models by optimizing a small subset of parameters while keeping most backbone parameters frozen. Low-rank adaptation is particularly attractive for MLLMs because it substantially reduces the memory and computational cost of downstream adaptation.

AdaLoRA improves conventional low-rank adaptation by dynamically allocating the available rank budget according to the estimated importance of different parameter matrices (Zhang et al., 2023). Instead of assigning the same rank to all adapted layers, it prunes less important components and allocates more capacity to parameters that contribute more strongly to the target task. This adaptive allocation improves parameter efficiency but does not explicitly distinguish between visual and textual token representations.

DoRA decomposes a pre-trained weight matrix into magnitude and direction components and applies low-rank adaptation primarily to its directional component (Liu et al., 2024c). This decomposition improves the learning capacity and stability of low-rank adaptation, narrowing the gap between PEFT and full-parameter fine-tuning. Nevertheless, DoRA remains modality-agnostic: its parameterization does not explicitly model the distributional discrepancy between visual and textual tokens in an MLLM.

Our method is complementary to these PEFT techniques. Rather than improving low-rank parameterization solely through adaptive rank allocation or weight decomposition, we introduce modality-aware adaptation interfaces in the self-attention key projection. Specifically, modality-specific low-rank branches process visual and textual tokens separately, while a distribution-alignment constraint reduces the cross-modal discrepancy in the resulting key representations. This design preserves the parameter efficiency of low-rank adaptation while explicitly addressing modality imbalance inside the decoder.

B.5 Multimodal Hallucination Benchmarks

The development of hallucination benchmarks has enabled systematic evaluation of whether MLLMs faithfully ground their outputs in visual inputs. Existing benchmarks differ in their evaluation formats, hallucination categories, and levels of diagnostic granularity.

POPE is a polling-based benchmark designed to evaluate object hallucination in large vision-language models (Li et al., 2023). It formulates object-existence evaluation as binary questions and constructs samples using random, popular, and adversarial negative-object selection strategies. This design makes POPE a widely adopted benchmark for measuring whether a model incorrectly claims that an absent object appears in an image. However, POPE primarily focuses on object existence and therefore captures only one dimension of multimodal hallucination.

HallusionBench evaluates both language hallucination and visual illusion in large vision-language models (Guan et al., 2023). It contains examples in which linguistic reasoning and visual perception are entangled, making it possible to diagnose whether an incorrect response originates from visual misunderstanding, language-based reasoning errors, or their interaction. Compared with object-existence benchmarks, HallusionBench covers more complex multimodal reasoning and perception scenarios.

AMBER is an LLM-free, multidimensional benchmark for evaluating hallucination in MLLMs (Wang et al., 2023). It assesses several hallucination dimensions, including object existence, attributes, relations, and discriminative judgments. Its LLM-free evaluation design reduces dependence on proprietary language models as automatic judges and provides a broader characterization of multimodal hallucination behavior.

These benchmarks provide complementary perspectives: POPE focuses on object-existence hallucination, HallusionBench examines entangled visual illusion and language hallucination, and AMBER evaluates multiple hallucination dimensions. We use these benchmarks to assess whether aligning modality-specific representations in the attention key space improves visual grounding across different evaluation settings.

Table 11: Cross-modal key distribution alignment at the **late layer** for LLaVA-1.5-7B, Qwen2.5-VL-7B, and Qwen3-VL-8B on MMBench-EN and MMMU. MMD and JS divergence quantify the representation gap between image and text features in the projected key space.

MMD

	LLaVA-1.5-7B		Qwen2.5-VL-7B		Qwen3-VL-8B	
Setting	MMB-EN	MMMU	MMB-EN	MMMU	MMB-EN	MMMU
Base	0.16	0.22	0.04	0.23	0.02	0.02
Ours	0.07	0.07	0.03	0.21	0.01	0.01
Δ	0.09	0.15	0.01	0.02	0.01	0.01

JS

	LLaVA-1.5-7B		Qwen2.5-VL-7B		Qwen3-VL-8B	
Setting	MMB-EN	MMMU	MMB-EN	MMMU	MMB-EN	MMMU
Base	0.39	0.39	0.31	0.27	0.47	0.35
Ours	0.32	0.36	0.28	0.16	0.41	0.31
Δ	0.07	0.03	0.03	0.09	0.06	0.04

B.6 Cross-Modal and Distribution Alignment

Cross-modal representation alignment and distribution matching are widely used in multimodal learning, domain adaptation, and domain generalization. Their central goal is to reduce discrepancies between representations originating from different modalities or data domains while preserving task-relevant semantic information.

Prior methods include optimal-transport-based alignment, which seeks a structure-preserving correspondence between representation sets (Chen et al., 2025), and kernel-based statistical distances such as maximum mean discrepancy (MMD), which measures differences between distributions through their embeddings in a reproducing kernel Hilbert space (Gretton et al., 2012; Sun et al., 2024a). In addition, Song et al. (2024) study semi-supervised alignment from a distributional perspective and discuss training paradigms that improve cross-modal semantic consistency under weak pairing or limited alignment data.

Most existing distribution-alignment methods operate on global multimodal embeddings, encoder outputs, or task-level feature representations. However, modality imbalance can also emerge internally within the decoder, where visual and textual tokens interact through self-attention. In particular, if visual keys and textual keys occupy substantially different regions of the representation space, their attention compatibility with subsequent query tokens may become systematically imbalanced.

Our work introduces distribution alignment directly into the self-attention key space of MLLMs. Rather than enforcing only sample-wise feature similarity, we constrain the modality-level distributional discrepancy between visual and textual keys during parameter-efficient fine-tuning. This formulation targets the internal mechanism through which modality-specific representations influence attention allocation, while avoiding additional computation or architectural changes at inference time.

C Additional Analysis and Experimental Results

C.1 Cross-modal Key-space Divergence at Middle and Late Layers

Table 11 complements the analysis of the first layer in the main text by reporting the divergence between the visual and textual key distributions in the middle and late layers. Consistent with the main-text results, MaLoRA reduces cross-modal discrepancy under both MMD and Jensen–Shannon (JS) divergence across most model–dataset pairs on MMBench-EN and MMMU for all three backbones. At the middle layer, the reduction remains clearly visible, indicating that the improvement in visual–textual alignment is preserved beyond shallow representations after several layers of multimodal interaction. A similar trend is also observed at the late layer: although the absolute divergence values and the magnitude of improvement vary more across settings, MaLoRA still lowers cross-modal divergence in most cases. Overall, these results are consistent with the early-layer findings and suggest that the benefit of MaLoRA is not restricted to a single layer but remains observable across different stages of the network.

C.2 Image-Text Irrelevant Experiments

Beyond these general benchmarks, we also construct and use an image-text irrelevant test suite to evaluate model robustness under text-visual conflict conditions (i.e., when distracting text is irrelevant to visual evidence). Specifically, we adopt the image-text irrelevant subsets from Deng et al.

Table 12: Performance (%) on the irrelevant benchmark subsets (`vqav2_irrelevant`, `openphish_irrelevant`, and `docvqa_irrelevant`) for Qwen2.5-VL-7B and Qwen3-VL-8B under four fine-tuning settings: Base, LoRA, QLoRA, and MaLoRA.

	Qwen2.5-VL-7B				Qwen3-VL-8B			
Datasets	Base	LoRA	QLoRA	MaLoRA	Base	LoRA	QLoRA	MaLoRA
vqav2_irrelevant	68.4	71	58	71.8	70.7	71.8	66.9	72.2
openphish_irrelevant	1.6	2.48	2.04	5.24	0.91	1.88	0.4	3.04
docvqa_irrelevant	91.6	91.2	83.3	91.9	92.8	92.6	91.5	93.5

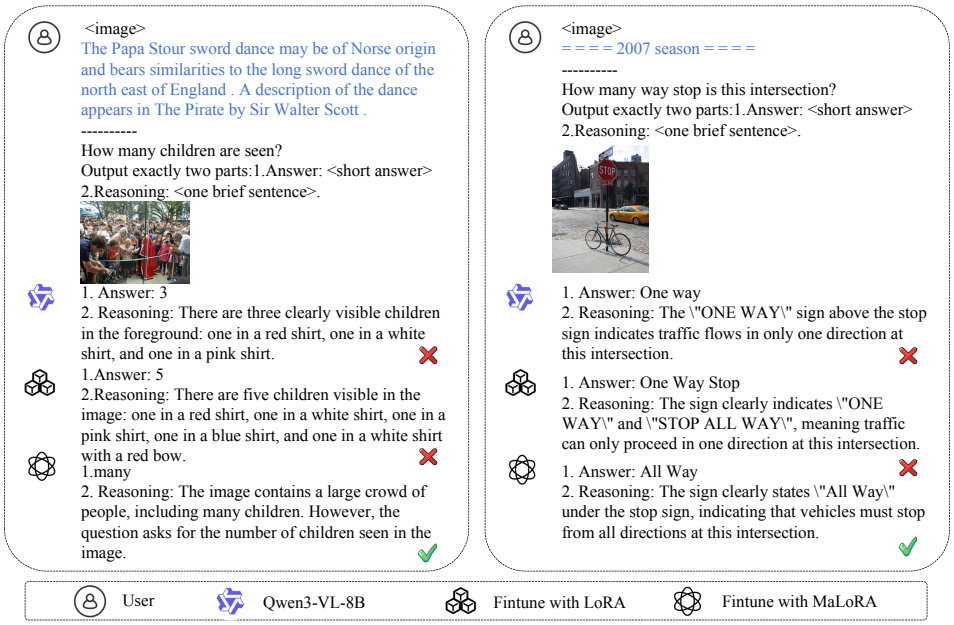


Figure 5: Case studies on the irrelevant setting. When irrelevant textual distractors are added to the prompt, the base Qwen3-VL-8B and LoRA-finetuned models are more likely to be misled, while the MaLoRA-finetuned model remains grounded in visual evidence and produces correct answers.

(2025) study as test data, and unify them into the test processed format, including three subsets: vqav2_irrelevant, docvqa_irrelevant, and openphish_irrelevant, with evaluation sizes of 1,000, 1,000, and 5,000, respectively. The corresponding results are reported in Table 12.

From Table 12, MaLoRA achieves the best performance across most settings, with the clearest gains on openphish_irrelevant (e.g., 5.24 vs. 2.48/2.04 for Qwen2.5-VL-7B, and 3.04 vs. 1.88/0.40 for Qwen3-VL-8B), indicating stronger robustness under severe irrelevant-text interference. On vqav2_irrelevant and docvqa_irrelevant, improvements are still consistent, suggesting that our method improves robustness while preserving strong baseline capability on relatively easier irrelevant scenarios.

As shown in Figure 5, we present two representative case studies under the irrelevant setting, where irrelevant textual distractors are inserted into the input and the model is expected to answer based on visual evidence. The results indicate that both the base Qwen3-VL-8B model and the LoRA-finetuned variant are more easily biased by textual priors, leading to errors such as incorrect counting and semantic misinterpretation of traffic signs. In contrast, the MaLoRA-finetuned model consistently ignores irrelevant text, focuses on key visual cues, and produces correct answers. These qualitative examples demonstrate that our method achieves stronger robustness and better vision-grounded alignment under text–vision conflict conditions.

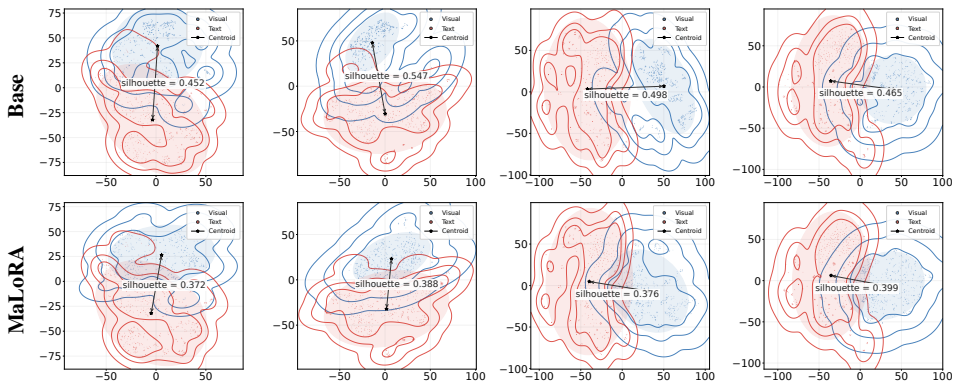


Figure 6: t-SNE visualization of hidden representations for Base and MaLoRA. From left to right, the four columns correspond to Qwen3-VL-8B on MMMU, Qwen3-VL-8B on MMBench-EN, Qwen2.5-VL-7B on MMMU, and Qwen2.5-VL-7B on MMBench-EN, respectively. The top and bottom rows show the results of Base and MaLoRA, respectively.

Table 13: Comparison of the Base model (without task-specific fine-tuning), LoRA, and our method on MMStar across different sizes of the Qwen2.5-VL model family.

Method	Qwen2.5-VL-3B	Qwen2.5-VL-7B	Qwen2.5-VL-32B	Qwen2.5-VL-72B
Base	55.67	65.67	65.00	65.33
LoRA	57.33	61.67	70.00	71.33
Ours	58.67	66.67	71.67	73.67

C.3 Modality Gap Metric

We use the modality gap metric to quantify the alignment between visual and textual representations in Figure 1. Here we provide the formal definition. For a given transformer layer l, let $\{\mathbf{v}_i\}_{i=1}^{N_v} \subset \mathbb{R}^d$ and $\{\mathbf{t}_j\}_{j=1}^{N_t} \subset \mathbb{R}^d$ denote the key vectors extracted from visual tokens and text tokens, respectively. We first apply ℓ_2 normalization to each vector:

$$\hat{\mathbf{v}}_i = \frac{\mathbf{v}_i}{\|\mathbf{v}_i\|_2}, \quad \hat{\mathbf{t}}_j = \frac{\mathbf{t}_j}{\|\mathbf{t}_j\|_2}. \tag{8}$$

The modality gap at layer l is then defined as the Euclidean distance between the centroids of the two normalized distributions:

$$\mathcal{G}^{(l)} = \left\| \frac{1}{N_v} \sum_{i=1}^{N_v} \hat{\mathbf{v}}_i \; - \; \frac{1}{N_t} \sum_{j=1}^{N_t} \hat{\mathbf{t}}_j \right\|_2. \tag{9}$$

A smaller $\mathcal{G}^{(l)}$ indicates that the visual and textual representations are more closely aligned in the attention key space at layer l. Since all vectors lie on the unit hypersphere after normalization, $\mathcal{G}^{(l)} \in [0, 2]$.

C.4 Additional t-SNE Visualization

Figure 6 presents additional t-SNE visualizations for Qwen3-VL-8B and Qwen2.5-VL-7B on MMMU and MMBench-EN. Across all four model–dataset pairs, the Base model shows a visible separation between visual and text representations, reflecting persistent cross-modal discrepancy in the key space. After fine-tuning with MaLoRA, the two modalities become noticeably closer and more mixed. This qualitative evidence further supports our main conclusion that MaLoRA consistently alleviates cross-modal key-space misalignment across different backbones and benchmarks.

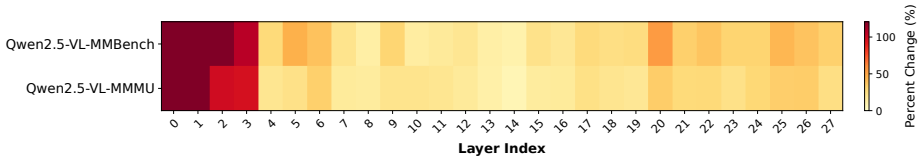


Figure 7: Layer-wise relative change (%) in aggregated Text→Visual attention after MaLoRA fine-tuning, measured with respect to the corresponding base model on Qwen2.5-VL-7B across MMMU and MMBench. Positive values indicate that textual queries allocate a larger proportion of attention to visual keys after fine-tuning.

Table 14: Training and inference costs on Qwen2.5-VL-7B under identical hardware and batch settings. Peak GPU Mem denotes the maximum GPU memory used during training, Step Time is the average wall-clock time per optimization step, Params Added is the proportion of trainable parameters relative to the base model, and Inference Overhead is the per-sample latency relative to Base, measured over 5 GPU runs with 20 output tokens.

Method	Peak GPU Mem (GB)	Step Time (s)	Params Added	Inference Overhead
Base	16.87	4.450	0.000%	+0.00%
QLoRA	16.10	6.645	1.991%	+200.12%
LoRA	25.07	4.469	1.991%	+0.00%
LoRA r72	28.61	4.546	2.240%	+0.08%
DoRA	21.90	10.792	2.007%	+0.00%
AdaLoRA	25.87	5.634	0.6130%	+83.43%
MaLoRA(Ours)	**25.49**	**6.381**	**2.234%**	**+30.75%**

C.5 Layer-wise Text-to-Visual Attention Change Analysis

To examine whether the reduction in key-space divergence translates into functionally meaningful changes in attention allocation, we measure the layer-wise relative change in aggregated Text→Visual attention after MaLoRA fine-tuning. If aligning visual and textual keys indeed makes visual evidence more accessible to textual queries, we should observe increased attention from text tokens to visual keys.

As shown in Figure 7, MaLoRA consistently increases Text→Visual attention across all three backbones. For LLaVA-1.5-7B, the increase is pronounced in both early and middle layers on MMMU and MMBench. Similarly, Qwen2.5-VL-7B exhibit strong attention strengthening in shallow layers (layers 0–5), where cross-modal interactions are first formed.

Together with the cross-modal divergence analysis in Table 3 and Appendix Table 11, these results support the proposed mechanism: reducing the geometric mismatch between visual and textual keys in the projected key space leads to greater attention allocation from text tokens to visual evidence during decoding.

C.6 Scaling Analysis

To assess whether key-space alignment generalizes across model capacities, we evaluate MaLoRA on four scales of Qwen2.5-VL (3B, 7B, 32B, and 72B) using MMStar, a benchmark that stresses fine-grained visual reasoning. As reported in Table 13, MaLoRA consistently outperforms both the Base checkpoint and standard LoRA at every scale. The absolute improvements over LoRA range from +1.34 (3B) to +5.00 (7B), with the largest gain observed at 7B. Notably, at this scale, LoRA fine-tuning degrades performance below the Base model (61.67 vs. 65.67), whereas MaLoRA surpasses both (66.67). These results indicate that the benefit of key-space alignment is not restricted to a narrow capacity regime, but scales effectively from small to large multimodal models.

C.7 Training and Inference Cost

We measure the practical overhead of MaLoRA on Qwen2.5-VL-7B under identical hardware and batch settings, comparing it with standard LoRA, QLoRA, DoRA, and AdaLoRA. As reported in Table 14, MaLoRA introduces 2.234% trainable parameters, comparable to LoRA r64 (1.991%). Its peak GPU memory increases only marginally from 25.07 GB to 25.49 GB (+1.7%), remaining well below LoRA r72 (28.61 GB). The per-step training time rises from 4.469 s to 6.381 s (+42.8%) due to token-dependent gating and the MMD/Gram regularization objectives; nevertheless, MaLoRA remains faster than QLoRA (6.645 s) and substantially faster than DoRA (10.792 s).

At inference, LoRA and DoRA introduce almost no latency overhead because their updates can be merged into base weights. MaLoRA incurs a +30.75% overhead because its token-wise gating is evaluated online and cannot be fully merged. This is markedly lower than AdaLoRA (+83.43%) and QLoRA (+200.12%). Moreover, because the reported latency is measured with only 20 output tokens, the relative overhead of MaLoRA is expected to decrease as generation length increases, since fixed prefill and execution costs are amortized over longer sequences. Overall, MaLoRA retains the parameter efficiency and memory footprint of LoRA while offering a practical trade-off between computational cost and improved visual grounding.

D Evaluation Prompts

To ensure reproducibility and facilitate future research, we provide here the complete set of prompts used to evaluate our model across all benchmarks. These prompts were consistently applied during inference to maintain fairness and comparability.

D.1 General Understanding

MMBench-EN

$< image >$
Question: question
Options:
options
Please select the correct answer from the options above.

D.2 Expert Reasoning

MMStar

$< image >$
Question: question
Options:
options
Please select the correct answer from the options above.

SimpleVQA

$< image >$
Question: question
Answer the question using a single word or phrase.

D.3 MATH REASONING

WeMath

$< image >$
Now, we require you to solve a multiple-choice math question. Please briefly describe your thought process and provide the final answer(option).
Question: question
Option: options
Answer the question using a single word or phrase, strictly in the format of A, B, C, D.

DynaMath | MathVision

$< image >$
Question: question
Answer the question using a single word or phrase.

D.4 OCR QA

OCRVQA | TextVQA | ST-VQA

$< image >$
Question: question
Answer the question using a single word or phrase.

D.5 STRUCTURED QA

DocVQA | ChartQA

$< image >$
Question: question
Answer the question using a single word or phrase.

D.6 GUI GROUNDING

RICO-ScreenQA

$< image >$
Question: question
Answer the question using a single word or phrase.

You are an impartial evaluator. Given a question, the ground truth answer, and a model's prediction, judge whether the prediction is semantically correct (equivalent to or appropriately conveys the ground truth).
Question: question
Ground Truth Answer: answer
Model Prediction: prediction
Is the model's prediction correct? Answer with ONLY "YES" or "NO". Answer:

D.7 DOMAIN-SPECIFIC

RSVQA

$< image >$
Question: question
Answer the question using a single word or phrase.

VQA-RAD

$< image >$
Question: question
Answer the question using a single word or phrase.