IMAGE KEYS AND VALUES ARE ASYMMETRIC IN MLLM TEXT-TO-IMAGE ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) integrate visual information via text-to-image (T2I) attention in the decoder. While recent studies identify various T2I attention phenomena, the underlying signals driving this attention remain underexplored. We investigate this through a structural analysis of image Key and Value representations, decomposing each into a visual prior and an instance-aware residual. This uncovers two Key–Value asymmetries: within each layer, image Keys are prior-dominated while Values are residual-dominated; across layers, the residual grows in Values but fades in deep-layer Keys. A cross-distribution analysis further shows the Key prior is input-agnostic, retaining high similarity under extreme distributional shifts. This reveals a routing–content gap: deep layers carry rich instance-aware content in Values, yet route attention through Keys dominated by an input-agnostic prior. Notably, within the same model family, the larger model exhibits a greater instance-aware component in its image Keys, providing complementary support for residual amplification. We therefore propose `Image Key Steering`, a training-free approach that amplifies the instance-aware component of deep-layer image Keys. Across four MLLMs and ten benchmarks, `Image Key Steering` consistently improves base-model performance, with the largest gains on tasks requiring precise visual grounding.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) have rapidly advanced multimodal understanding (Alayrac et al., 2022; Yin et al., 2024). By coupling vision encoders (Radford et al., 2021; Dosovitskiy et al., 2021; Zhai et al., 2023) with large language models (Touvron et al., 2023a;b; Grattafiori et al., 2024; Brown et al., 2020; Chowdhery et al., 2023), architectures like LLaVA (Liu et al., 2023), VILA (Lin et al., 2024), and Qwen-VL (Wang et al., 2024a) achieve remarkable performance across diverse tasks, from visual question answering (Antol et al., 2015; Hudson & Manning, 2019; Singh et al., 2019; Mathew et al., 2021) to cross-modal reasoning (Yue et al., 2024; Lu et al., 2022). At their core lies the decoder's attention mechanism (Vaswani et al., 2017), where text tokens attend to image tokens to ground generation in visual evidence. Understanding the signals governing this text-to-image (T2I) attention is therefore central to understanding MLLM behavior.

Recent works document macroscopic T2I attention phenomena, such as attention sinks (Kang et al., 2025b; Zhang et al., 2025b; Choi et al., 2026; Fan et al., 2026), head-level specialization (Bi et al., 2025; Kim et al., 2025; Basile et al., 2025), and layer-wise cross-modal information flow (Jiang et al., 2025; Wu et al., 2026; Zhang et al., 2025c). However, a fundamental question remains underexplored: *what underlying signals in image token representations drive MLLM T2I attention?* Answering this reveals the representation-level mechanisms governing T2I attention and enables targeted intervention. We investigate this through a structural analysis of image Key and Value representations: Keys interact with text Queries to determine the attention weights, whereas Values supply the representations combined by those weights. Specifically, in §3.1, we apply an ANOVA-style decomposition (Song & Zhong, 2023; Doshi et al., 2026) at each decoder layer, separating each representation into a shared ***visual prior*** and a per-image ***instance-aware residual***. This uncovers two Key–Value asymmetries:

- **Within each layer**, image Keys are dominated by the visual prior while image Values are dominated by the instance-aware residual. Thus, T2I attention routing relies heavily on shared signals, while

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.

the Values being aggregated remain strongly instance-aware (see Fig. 1). A cross-distribution analysis further confirms the Key prior is ***input-agnostic***, retaining high similarity even under extreme distributional shifts (see Fig. 2).
- **Across layers**, the routing–content split widens with depth: the residual energy in image Values grows substantially in deeper layers, while that of image Keys fades (see Fig. 3). Deep layers thus carry rich instance-aware content in Values yet route attention through Keys dominated by an input-agnostic prior. This divergence originates within the language decoder rather than the vision components (see Appendix §S5) and offers a plausible explanation for why deep-layer image tokens can be pruned with negligible performance loss (Lin et al., 2025) (see Appendix §S6).

Notably, within the same model family, the larger model exhibits a greater instance-aware component in its image Keys on average (see §3.2), providing complementary support for residual amplification. Together, these observations suggest that strengthening the fading Key residual may narrow the routing–content gap and better utilize the rich visual content already encoded in Values. We therefore propose `Image Key Steering` (see §4), a training-free inference-time method that amplifies this residual in deep layers while preserving the shared visual prior. It requires only a lightweight calibration pass to estimate the visual prior, adding no parameters or fine-tuning. Across multiple MLLM architectures and a broad suite of vision-language benchmarks, `Image Key Steering` delivers consistent gains, particularly on tasks demanding precise visual grounding. Overall, enhancing the semantic expressiveness of image Keys offers a promising new direction for improving vision-language understanding. Our contributions are as follows:

- **We identify a routing–content gap in T2I attention through two Key–Value asymmetries.** Within each layer, image Keys are prior-dominated while Values are residual-dominated, with the Key prior being input-agnostic and the Value residual being instance-aware. Across layers, the instance-aware Key component fades just as Values grow richest. Thus, MLLMs carry fine-grained visual content in deep layers yet route attention through signals dominated by an input-agnostic prior. A within-family comparison further shows that the larger model exhibits a greater instance-aware Key component on average, providing complementary support for residual amplification.
- **We propose `Image Key Steering`, a training-free inference-time method that directly closes this gap.** It amplifies the instance-aware component of image Keys in deep layers via a lightweight calibration pass, requiring no additional parameters, no fine-tuning, and integrating seamlessly into existing MLLM pipelines. The visual prior required for calibration can be estimated from a small set of unlabeled images, with performance robust to the calibration distribution.
- **We validate `Image Key Steering` across four MLLMs and ten benchmarks spanning three task categories.** It consistently improves the base models and is competitive with or outperforms inference-time baselines across most benchmarks, with the largest gains on tasks demanding precise visual grounding (*e.g.*, a 35% reduction in object hallucination, CHAIR_I: $15.2 \rightarrow 9.8$, on LLaVA-1.5-7B). The gains hold across both 7B and 13B scales and across fixed-length (*e.g.*, LLaVA) and dynamic-resolution visual tokenization (*e.g.*, Qwen2-VL).

2 Related Work

2.1 Text-to-Image Attention in Multimodal Large Language Models

Understanding how MLLMs allocate text-to-image (T2I) attention has become important for interpreting and improving cross-modal generation. Existing studies can be broadly organized along three granularities: *Token-level*, *Head-level*, and *Layer-level*. *Token-level* studies focus on the attention weights image tokens receive and on how these weights can be modulated at the token or region level to steer model behavior. For example, redistributing attention away from sink tokens to semantically relevant ones (Kang et al., 2025b; Zhang et al., 2025b), or amplifying attention on task-relevant regions can improve perception and reduce hallucination (Liu et al., 2024a; Zhang et al., 2025a; Liu et al., 2026). *Head-level* studies identify that a small subset of attention heads carries the bulk of image-to-text information transfer, with head selection governed by input semantics rather than low-level appearance (Kim et al., 2025). *Layer-level* studies examine where visual information propagates through the decoder, with findings spanning concentration of cross-modal information flow in middle layers (Kaduri et al., 2025; Jiang et al., 2025), redundancy of visual tokens in deep layers that enables efficient token compression (Chen et al., 2024; Lin et al., 2025), and stage-wise functional decomposition of the decoding process (Yu & Lee, 2025). Despite their substantial

advances, these studies largely remain at the level of observable attention patterns, focusing on *where* attention is allocated across tokens, heads, or layers. Much less understood is *what* signals in image token representations shape this allocation. We address this gap by opening up the image Key and Value projections themselves and decomposing each into an input-agnostic visual prior and an instance-aware residual. This decomposition uncovers a routing–content gap between Keys and Values that is orthogonal to weight-level analyses, and we show that directly acting on it yields consistent improvements in vision-language understanding.

2.2 Decomposition of Representations and Attention

Classical decomposition methods, such as PCA, SVD, and non-negative matrix factorization, have long been used to probe the structure of high-dimensional representations (Jolliffe & Cadima, 2016; Lee & Seung, 1999; Saxe et al., 2019). Recent work applies this perspective to transformers in two complementary ways: decomposing token representations into interpretable factors, and decomposing attention directly. For example, hidden states in language models have been analyzed in terms of global, positional, and contextual components (Song & Zhong, 2023), and functional ANOVA has been adapted to neural representations to isolate additive and interaction effects of input factors (Märtens & Yau, 2020). Decomposition has also been applied directly to attention: singular value decomposition of the query-key interaction matrix has been used to study how self-attention evolves across layers in vision transformers (Pan et al., 2024), while bi-orthogonal factorization combines ANOVA-style decomposition of token activations with SVD of $QK^\top$ to disentangle positional and content contributions to attention (Doshi et al., 2026). These methods provide a natural lens for studying the signals encoded in image token representations. Building on this perspective, our work leverages the philosophy of factorization as a mathematical tool to uncover the underlying signals of T2I attention of MLLMs.

3 Structural Analysis of Image Key and Value Representations

Image Keys determine how text Queries route attention over image tokens, whereas image Values carry the visual content retrieved by that routing. We first formalize T2I attention and then decompose the image Key and Value representations at each decoder layer into a visual prior and an instance-aware residual (see §3.1). Examining their within-layer composition and depth-wise evolution on LLaVA-1.5-7B (Liu et al., 2023) reveals two Key–Value asymmetries (see §3.2): ❶ within each layer, image Keys are dominated by the prior while image Values are dominated by the instance-aware residual, with cross-distribution analysis further showing that the Key prior is input-agnostic; and ❷ across layers, the instance-aware residual grows steadily in Values but fades in deep-layer Keys. Together, these asymmetries produce a widening gap between the representations that route attention and those that carry visual content. The same patterns also hold generally for other MLLMs, such as Qwen2VL-7B (Wang et al., 2024a) (Appendix §S4).

Preliminary: Text-to-Image Attention. Consider the l-th decoder layer of an MLLM. The input sequence comprises N_i image tokens followed by N_t text tokens, forming the concatenated representation $\mathbf{X}^{(l)} = [\mathbf{X}^{(l)}_{\text{img}}; \mathbf{X}^{(l)}_{\text{text}}] \in \mathbb{R}^{(N_i+N_t)\times d}$, where d is the hidden dimension. The self-attention mechanism computes Query, Key, and Value projections as:

$$\mathbf{Q}^{(l)} = \mathbf{X}^{(l)}\mathbf{W}^{(l)}_Q, \quad \mathbf{K}^{(l)} = \mathbf{X}^{(l)}\mathbf{W}^{(l)}_K, \quad \mathbf{V}^{(l)} = \mathbf{X}^{(l)}\mathbf{W}^{(l)}_V, \tag{1}$$

where $\mathbf{W}^{(l)}_Q, \mathbf{W}^{(l)}_K, \mathbf{W}^{(l)}_V \in \mathbb{R}^{d\times d}$ are the projection matrices.

Focusing on the text-to-image (T2I) interaction, the image-conditional attention from text tokens (as queries) to image tokens (as keys) is defined as:

$$\mathbf{A}^{(l)}_{\text{T2I}} = \text{softmax}\left(\frac{\mathbf{Q}^{(l)}_{\text{text}}\left(\mathbf{K}^{(l)}_{\text{img}}\right)^\top}{\sqrt{d}}\right), \tag{2}$$

where $\mathbf{Q}^{(l)}_{\text{text}} \in \mathbb{R}^{N_t\times d}$ and $\mathbf{K}^{(l)}_{\text{img}} \in \mathbb{R}^{N_i\times d}$ are the Query and Key projections of the text and image tokens, respectively.

The corresponding image-conditional Value readout is:

$$\mathbf{O}^{(l)}_{\text{img}\to\text{text}} = \mathbf{A}^{(l)}_{\text{T2I}}\,\mathbf{V}^{(l)}_{\text{img}}, \tag{3}$$

with $\mathbf{V}^{(l)}_{\text{img}} \in \mathbb{R}^{N_i \times d}$. Image Keys thus govern *where* text tokens attend (Eq. 2), while image Values determine *what* visual content is retrieved (Eq. 3). In the native decoder, the actual visual contribution is further scaled by each text Query's total attention mass on image tokens.

3.1 Decomposing Image Keys and Values

With the roles of image Keys and Values established, we now ask what signals in their representations drive T2I attention. Inspired by ANOVA-style factorizations of neural representations (Song & Zhong, 2023; Doshi et al., 2026), we decompose image Keys and Values at each decoder layer into a visual prior and an instance-aware residual. Unlike the three-way decomposition of Song & Zhong (2023), we do not separate positional effects from the global mean, as positional structure is itself shared across images and is naturally absorbed into the prior term. This also avoids the complication that a per-position decomposition is not well-defined for MLLMs with highly variable image token lengths (*e.g.*, Qwen2VL).

Definition: Visual Prior and Instance-aware Residual. Let $\mathbf{h}^{(l)}(\mathbf{x}) \in \mathbb{R}^{N_i \times d}$ denote the image-side Key or Value projection at decoder layer l for input image $\mathbf{x}$, where $l \in \{0, 1, \ldots, L-1\}$ and L is the total number of decoder layers. Given a calibration set $\mathcal{D}_{\text{cal}}$, we decompose it as

$$\mathbf{h}^{(l)}(\mathbf{x}) = \boldsymbol{\mu}^{(l)} + \mathbf{r}^{(l)}(\mathbf{x}), \tag{4}$$

where $\boldsymbol{\mu}^{(l)}$ is the *visual prior*, capturing the component shared across images in $\mathcal{D}_{\text{cal}}$, and $\mathbf{r}^{(l)}(\mathbf{x})$ is the *instance-aware residual*, capturing the image-specific deviation from this prior:

$$\mathbf{r}^{(l)}(\mathbf{x}) = \mathbf{h}^{(l)}(\mathbf{x}) - \boldsymbol{\mu}^{(l)}. \tag{5}$$

We apply this decomposition separately to image Keys and image Values, yielding their respective priors $\boldsymbol{\mu}^{(l)}_K$ and $\boldsymbol{\mu}^{(l)}_V$, and residuals $\mathbf{r}^{(l)}_K(\mathbf{x})$ and $\mathbf{r}^{(l)}_V(\mathbf{x})$.

Estimating the visual prior. We estimate $\boldsymbol{\mu}^{(l)}$ from a calibration set $\mathcal{D}_{\text{cal}}$ of images sampled from the target benchmark (*e.g.*, MME (Fu et al., 2025), ScienceQA (Lu et al., 2022)), covering a broad range of visual content (*e.g.*, objects, scenes, landmarks, OCR, artworks, scientific diagrams, *etc*). *Case A.* For models with a fixed image token length N_i (*e.g.*, LLaVA (Liu et al., 2023)), the visual prior is computed per token position as:

$$\boldsymbol{\mu}^{(l)} = \frac{1}{|\mathcal{D}_{\text{cal}}|} \sum_{\mathbf{x}\in\mathcal{D}_{\text{cal}}} \mathbf{h}^{(l)}(\mathbf{x}) \;\in\; \mathbb{R}^{N_i \times d}. \tag{6}$$

Case B. For models with variable image token lengths (*e.g.*, Qwen2VL (Wang et al., 2024a)), we first apply global average pooling over image tokens before averaging across images:

$$\boldsymbol{\mu}^{(l)} = \frac{1}{|\mathcal{D}_{\text{cal}}|} \sum_{\mathbf{x}\in\mathcal{D}_{\text{cal}}} \mathrm{GAP}\Big(\mathbf{h}^{(l)}(\mathbf{x})\Big) \;\in\; \mathbb{R}^{d}, \tag{7}$$

where $\mathrm{GAP}(\cdot)$ denotes Global Average Pooling, which averages over the token dimension.

Energy and its fraction. We measure the magnitude of each component as its mean squared energy. We further confirm the functional relevance of this decomposition through actual-Query analysis and a norm-matched residual control (Appendices §S3 and §S7.1). Specifically, the prior energy and the residual energy at layer l are

$$E^{(l)}_{\text{prior}} \;=\; \big\|\boldsymbol{\mu}^{(l)}\big\|_F^2, \qquad E^{(l)}_{\text{res}} \;=\; \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{cal}}}\big\|\mathbf{r}^{(l)}(\mathbf{x})\big\|_F^2, \tag{8}$$

and the corresponding *prior-energy* and *residual-energy fractions* are

$$\eta^{(l)}_{\text{prior}} \;=\; \frac{E^{(l)}_{\text{prior}}}{\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{cal}}}\big\|\mathbf{h}^{(l)}(\mathbf{x})\big\|_F^2}, \qquad \eta^{(l)}_{\text{res}} \;=\; 1 - \eta^{(l)}_{\text{prior}}. \tag{9}$$

We apply this decomposition separately to the Key and Value projections, denoting their priors as $\boldsymbol{\mu}^{(l)}_K$ and $\boldsymbol{\mu}^{(l)}_V$, their prior and residual energies as $E^{(l)}_{\text{prior},K}, E^{(l)}_{\text{res},K}$ and $E^{(l)}_{\text{prior},V}, E^{(l)}_{\text{res},V}$, and the corresponding prior-energy fractions as $\eta^{(l)}_{\text{prior},K}$ and $\eta^{(l)}_{\text{prior},V}$, respectively.

3.2 Two Key–Value Asymmetries in T2I Attention: Composition and Evolution



Figure 1: **Energy fractions of image Key and Value projections across decoder layers on LLaVA-1.5-7B.** Results are shown on MME (*Left two*) and ScienceQA (*Right two*). Image Keys are dominated by the input-agnostic prior across all layers on both benchmarks, while image Values are dominated by the instance-aware residual, growing with depth. Results on Qwen2VL-7B (Wang et al., 2024a) are shown in Appendix §S4.

Asymmetry 1: Within each layer, Keys are prior-dominated while Values are residual-dominated. The first asymmetry concerns what each projection is made of at any given layer. Across all decoder layers, the prior-energy fraction of image Keys, $\eta^{(l)}_{\text{prior},K}$, remains above 78% on both benchmarks (see Fig. 1, 1st and 3rd panels). This means that roughly four-fifths of the energy carried in Key representations comes from a single component shared across all input images. Image Values exhibit the opposite composition. As shown in Fig. 1 (2nd and 4th panels), the residual-energy fraction $\eta^{(l)}_{\text{res},V}$ dominates throughout and reaches 72–76% in deeper layers on both benchmarks. The signals that govern *where* text tokens attend are thus largely shared from one input to the next, while the signals that determine *what* is retrieved are strongly instance-aware.

The above analysis reveals that image Keys are dominated by a prior shared across the calibration set $\mathcal{D}_{\text{cal}}$. A crucial follow-up question is whether this Key prior is simply a statistical artifact of the input image distribution, or if it reflects an *input-agnostic* structural bias—an intrinsic property governing how the model routes attention. We examine this by re-estimating the prior on two input distributions that carry no natural-image statistics: Gaussian-noise images and solid random-color images. We then measure the cosine similarity between these estimates and the natural-image estimate. As shown in Fig. 2, the Key prior $\boldsymbol{\mu}^{(l)}_K$ remains nearly invariant across all three distributions, with cosine similarity above 0.90 at every layer. The Value prior $\boldsymbol{\mu}^{(l)}_V$, in contrast, varies substantially under the same intervention. Its cosine similarity drops below 0.70 in most layers and as low as 0.55 between natural and solid-color images, more than 0.35 below the corresponding Key value.

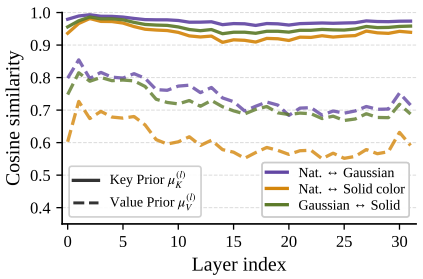

Figure 2: **Cross-distribution analysis of image Key and Value priors on LLaVA-1.5-7B.** The Key prior $\boldsymbol{\mu}^{(l)}_K$ remains stable across all input distributions, while the Value prior $\boldsymbol{\mu}^{(l)}_V$ varies substantially, revealing the Key prior as an input-agnostic property.

This observation confirms that the Key prior is indeed ***input-agnostic***: it reflects a routing structure the model imposes regardless of input content, not statistical regularities of any particular image distribution. The Value prior, by contrast, tracks the calibration distribution and behaves as a distribution-level visual statistic. We therefore interpret the dominant Key prior as a stable routing scaffold rather than as a defect by itself. Our method preserves this prior and instead targets the layer-wise attenuation of the instance-aware Key residual established next.

Asymmetry 2: Across layers, instance-aware content accumulates in Values but fades in deep-layer Keys. The second asymmetry concerns how each projection evolves with depth. As shown in Fig. 3 (2nd and 4th panels), the instance-aware residual energy of image Values, $E^{(l)}_{\text{res},V}$, grows nearly monotonically from shallow to deep layers on both benchmarks, indicating that Values progressively accumulate richer instance-aware content. Image Keys follow a strikingly different trajectory. As shown in Fig. 3 (1st and 3rd panels), $E^{(l)}_{\text{res},K}$ on LLaVA-1.5-7B rises modestly through the early-to-

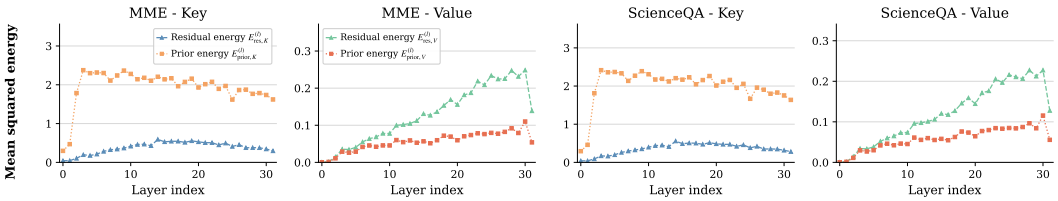

Figure 3: **Absolute energies of image Key and Value projections across decoder layers on LLaVA-1.5-7B.** Results are shown on MME (*Left two*) and ScienceQA (*Right two*). The prior energy dominates Key representations across all layers, while the residual energy stays small, rising to a peak around layer 14 and declining in deeper layers. The residual energy of image Values grows steadily with depth. Both benchmarks exhibit consistent patterns despite their different image distributions.

middle layers, peaks around layer 14, and then declines through the deeper layers. Notably, layer 14 has also been independently identified as a critical layer for T2I attention in prior work (Zhang et al., 2025a; Kang et al., 2025a), corroborating that the peak of instance-aware Key energy coincides with the layer where T2I attention is most semantically informative. Beyond this peak, the instance-aware residual of Keys actively fades in precisely the layers where Values are at their richest.

Together, these asymmetries reveal a routing–content gap: image Values grow increasingly instance-aware with depth, while the Keys that route attention to them remain prior-dominated and lose instance-aware signal. This gap emerges in the language decoder rather than the vision components (see Appendix §S5). A within-family comparison offers complementary evidence: LLaVA-OneVision-7B (Li et al., 2024) exhibits a higher mean Key residual fraction than its 0.5B variant in the early, middle, and deep thirds of the decoder (Fig. 4). This observation further motivates residual amplification as a way to narrow the routing–content gap. Query-level analysis further shows that the prior dominates overall routing, whereas the residual more closely captures the allocation among image tokens (see Appendix §S3).

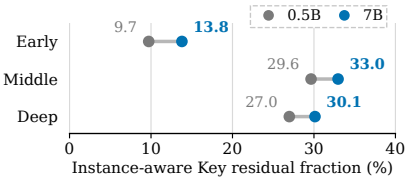

Figure 4: **Within-family comparison of instance-aware Key residuals.** Early, middle, and deep denote the first, middle, and final thirds of normalized decoder depth, respectively. Points show the mean residual-energy fraction within each region.

Further Discussion. Our two asymmetries offer a possible explanation for recent visual token pruning methods Chen et al. (2024); Lin et al. (2025); Fan et al. (2026); Han et al. (2025), which downsample image tokens in early layers and withdraw them in deep layers with little performance loss. In early-to-middle layers, image Keys are prior-dominated and only a small residual distinguishes tokens. In deep layers, this residual further diminishes, so the routing channel carries little instance-aware information to lose, even though Values remain rich. Pruning exploits this redundancy to reduce compute, whereas `Image Key Steering` (§4) takes the opposite route and strengthens the instance-aware component of Keys directly.

4 Image Key Steering

The analysis in §3 suggests a direct way to narrow the routing–content gap: strengthen the instance-aware Key residual where it fades, while preserving the prior. We therefore propose `Image Key Steering` (see Fig. S1), a training-free method that amplifies this residual in deep decoder layers.

Recall from Eq. 4 that the image Key projection at layer l admits the decomposition:

$$\mathbf{h}_K^{(l)}(\mathbf{x}) \;=\; \boldsymbol{\mu}_K^{(l)} \;+\; \mathbf{r}_K^{(l)}(\mathbf{x}), \tag{10}$$

where $\boldsymbol{\mu}_K^{(l)}$ is the input-agnostic visual prior and $\mathbf{r}_K^{(l)}(\mathbf{x})$ is the instance-aware residual capturing the per-image deviation from this prior. Image Key Steering modifies the image Key at each selected layer $l \in \mathcal{L}$ as:

$$\tilde{\mathbf{h}}_K^{(l)}(\mathbf{x}) \;=\; \boldsymbol{\mu}_K^{(l)} \;+\; \alpha \cdot \mathbf{r}_K^{(l)}(\mathbf{x}), \qquad \alpha > 1, \quad l \in \mathcal{L}, \tag{11}$$

where $\alpha = 1$ recovers the unmodified model and $\alpha > 1$ injects an additional dose of instance-aware signal into the Keys that route T2I attention. The modified Keys $\tilde{\mathbf{h}}_K^{(l)}(\mathbf{x})$ then enter the native decoder

attention, while text-side Keys, all Values, and the remaining model parameters are kept unchanged. In practice, the prior $\boldsymbol{\mu}_K^{(l)}$ is estimated from a small unlabeled image set with a single forward pass; Appendix §S7 provides a controlled calibration-distribution robustness check.

The form of Eq. 11 reflects three design choices, each grounded in the analysis of §3.2 and empirically validated in ablation studies in §5.3 and Appendix §S7. ***(i) Steering Keys, not Values.*** Image Values are already residual-dominated and grow richer with depth (Asymmetries 1 and 2), motivating an intervention on the fading Key residual. Value-only steering yields smaller gains than Key-only steering, and joint Key–Value steering provides no consistent additional benefit across the tested benchmarks (see Table 3). ***(ii) Amplifying the residual, not suppressing the prior.*** Closing the routing–content gap requires increasing the relative share of the instance-aware residual in the Key. Eq. 11 achieves this by scaling up the instance-aware residual while keeping the visual prior intact, since the visual prior serves as a stable, input-agnostic routing scaffold (see Fig. 2); suppressing the prior, by contrast, disrupts this scaffold and degrades performance (see Table 4). ***(iii) Targeting deep layers.*** The Key residual energy peaks in the middle and declines through deeper layers (Asymmetry 2), precisely the regime where Values are richest. Therefore, we restrict our intervention to this deeper range where the Key–Value gap is largest; mid- or shallow-layer steering substantially underperforms or corrupts the model (see Table 5).

Setting: why training-free? A natural extension is to learn α or per-layer residual transformations. We keep `Image Key Steering` training-free to provide a clean mechanistic validation of the analysis in §3 and to avoid the non-trivial challenge of formulating a learning objective that amplifies the residual without distorting the input-agnostic prior.

5 Experiments

5.1 Experimental Setup

Implementation details. As mentioned in §3.1, `Image Key Steering` applies to both fixed-length and dynamic-resolution MLLMs. We evaluate four base models across two scales and tokenization regimes: LLaVA-1.5-7B/13B (Liu et al., 2023) and VILA-13B (Lin et al., 2024) use fixed-length visual tokens, while Qwen2VL-7B (Wang et al., 2024a) uses resolution-dependent visual tokens. For each model, we select a steering layer set $\mathcal{L}$ once from the representation analysis and fix it across all benchmarks. We select category-level α values on randomly sampled 10% subsets of MME, CV-Bench^{2D}, and POPE for general understanding, vision-centric reasoning, and hallucination, respectively, then freeze each value across all benchmarks in its category. We show in Appendix §S7 that performance is robust across a wide range of α. We estimate the visual prior $\boldsymbol{\mu}_K^{(l)}$ from a small unlabeled set with a single forward pass. See Appendix §S2 for selection and calibration protocols.

Baselines. We compare `Image Key Steering` against eight recent inference-time methods that span the main families of training-free interventions for MLLMs: contrastive-decoding approaches that contrast outputs from perturbed visual inputs or leverage auxiliary distributions (*i.e.*, VCD (Leng et al., 2024), SID (Huo et al., 2025), ONLY (Wan et al., 2025), AGLA (An et al., 2025)), attention-editing approaches that directly modify attention patterns or activations (*i.e.*, OPERA (Huang et al., 2024), TAME (Tang et al., 2025), EVAS (Zhang et al., 2025b)), and visual-sink mitigation (*i.e.*, Vissink (Kang et al., 2025b)). This selection covers the dominant strategies for improving MLLM generation at inference without retraining or fine-tuning.

Benchmarks. We evaluate on a broad suite of vision-language benchmarks organized into three categories. *General vision-language understanding* (MME (Fu et al., 2025), LLaVA^{W} (Liu et al., 2023), MM-Vet (Yu et al., 2023), MMB^{en} (Liu et al., 2024b), SQA^{I} (Lu et al., 2022), SEED^{I} (Li et al., 2023a)) measures broad multimodal competence across perception, reasoning, and knowledge-intensive QA. *Vision-centric reasoning* (MMVP (Tong et al., 2024b), CV-Bench (Tong et al., 2024a)) targets fine-grained visual perception where success requires attending to subtle image details. *Visual hallucination* (CHAIR (Rohrbach et al., 2018), POPE (Li et al., 2023b)) measures how faithfully the model grounds its outputs in the actual image content. These three categories examine whether amplifying instance-aware signals in image Keys improves general competence, fine-grained perception, and visual grounding. Further details are provided in Appendix §S1.

Table 1: Evaluation results on **general vision-language and vision-centric tasks**. **Bold** indicates the best result within each model group, underline indicates the second best. All baseline results are reproduced by us using the best hyperparameters reported in their original papers.

Method	General Vision-Language						Vision-Centric	
	MME	SQA^I	MMB^{en}	$SEED^I$	$LLaVA^W$	MM-Vet	MMVP	CV-Bench^{2D}
7B Models								
LLaVA-1.5-7B	1504.8	66.8	64.3	58.6	59.2	29.9	3.33	56.8
+ VCD[CVPR24](Leng et al., 2024)	1488.5	63.3	61.4	58.9	59.2	29.2	10.0	53.6
+ TAME[ICLR25](Tang et al., 2025)	1518.4	69.4	65.3	59.4	59.5	30.0	11.3	61.2
+ EVAS[EMNLP25](Zhang et al., 2025b)	1515.4	67.5	**65.8**	60.2	59.6	31.7	10.7	61.3
+ Vissink[ICLR25](Kang et al., 2025b)	1513.8	67.3	65.4	60.7	59.9	31.0	9.33	57.6
+ Ours	**1520.6**	**69.6**	65.1	**61.1**	**60.9**	**32.2**	**12.0**	**61.4**
Qwen2VL-7B	1684.6	74.1	83.0	77.9	85.3	56.9	52.0	63.5
+ Vissink[ICLR25](Kang et al., 2025b)	1688.5	**74.3**	83.3	78.1	85.7	57.1	**55.3**	63.6
+ Ours	**1690.9**	74.2	**83.5**	**78.8**	**86.3**	**57.3**	54.7	**64.4**
13B Models								
LLaVA-1.5-13B	1529.9	71.6	67.7	61.6	66.7	34.9	24.7	58.2
+ VCD[CVPR24](Leng et al., 2024)	1521.6	69.3	67.4	61.8	65.7	34.0	27.3	60.7
+ EVAS[EMNLP25](Zhang et al., 2025b)	1537.6	72.1	68.2	62.1	66.9	35.4	30.7	61.7
+ Vissink[ICLR25](Kang et al., 2025b)	1534.3	72.2	68.1	62.3	67.2	35.1	28.0	59.6
+ Ours	**1556.6**	**73.0**	**68.3**	**62.9**	**67.6**	**35.8**	**31.3**	**62.6**
VILA-13B	1552.6	73.7	70.3	62.8	72.7	36.4	22.7	58.6
+ Vissink[ICLR25](Kang et al., 2025b)	1555.6	74.7	**71.7**	63.0	72.4	37.1	28.0	59.7
+ Ours	**1578.6**	**75.1**	71.4	**63.8**	**74.3**	**37.3**	**31.3**	**61.3**

5.2 Main Results

Table 1 and Table 2 summarize the performance on six general vision-language benchmarks, two vision-centric benchmarks, and two hallucination benchmarks across four MLLMs with different model scales and visual tokenization schemes. These results lead to three key observations. ***Image Key Steering consistently improves general vision-language performance.*** Across diverse MLLM architectures and scales, our method improves over the corresponding base model and achieves competitive or superior results compared to existing inference-time interventions across most benchmarks. For example, on LLaVA-1.5-13B our method raises MME from 1529.9 to 1556.6 and $SEED^I$ from 61.6 to 62.9, outperforming all competing approaches. The gains cover a broad range of capabilities, including multimodal reasoning, instruction following, visual question answering, and fine-grained visual understanding. Importantly, the improvements hold for both 7B and 13B models, as well as for fixed-resolution models (*i.e.*, LLaVA-1.5 and VILA) and dynamic-resolution models (*i.e.*, Qwen2-VL), suggesting that the Key-side routing–content gap identified in §3 is not specific to a particular architecture or scale. ***Vision-centric tasks see particularly notable improvements.*** On benchmarks that demand precise visual grounding, Image Key Steering improves LLaVA-1.5-7B on MMVP from 3.33 to 12.0 and LLaVA-1.5-13B from 24.7 to 31.3, consistent with the prediction that fine-grained perception is most directly limited by the routing capacity of image Keys (see Appendix §S8 for qualitative examples). ***The benefits of stronger visual grounding further translate to hallucination benchmarks.*** By closing the routing–content gap, `Image Key Steering` enables the model to attend more faithfully to visual evidence during generation, which directly reduces object fabrication. On LLaVA-1.5-7B, our method attains the highest POPE accuracy (*i.e.*, 86.9) and the lowest instance-level hallucination (*i.e.*, 9.8) among all baselines while achieving the highest recall (*i.e.*, 79.4), indicating that the reduction does not come at the cost of missing real objects. The same trend holds at larger scale, where on LLaVA-1.5-13B our method further reduces instance-level hallucination to 9.5 on CHAIR while improving POPE accuracy to 87.5.

Table 2: Hallucination evaluation on **POPE and CHAIR**. ↑/↓ indicate higher/lower is better. All baseline results are reproduced by us using the best hyperparameters reported in their original papers.

Method	POPE		CHAIR		
	F1↑	Acc↑	C_S↓	C_I↓	Recall↑
LLaVA-1.5-7B	85.4	84.0	51.0	15.2	75.2
+ OPERA[CVPR24](Huang et al., 2024)	84.2	85.2	47.0	14.6	78.5
+ VCD[CVPR24](Leng et al., 2024)	85.3	85.0	51.0	14.9	77.2
+ SID[ICLR25](Huo et al., 2025)	85.6	85.8	44.2	12.2	73.0
+ AGLA[CVPR25](An et al., 2025)	84.6	85.5	**43.0**	14.1	78.9
+ ONLY[ICCV25](Wan et al., 2025)	85.5	85.1	49.8	14.3	75.9
+ Vissink[ICLR25](Kang et al., 2025b)	**86.0**	86.5	52.4	14.5	79.1
+ Ours	85.9	**86.9**	44.2	**9.8**	**79.4**
LLaVA-1.5-13B	85.9	85.5	47.2	13.6	81.4
+ Vissink[ICLR25](Kang et al., 2025b)	86.2	86.6	44.0	11.9	81.7
+ Ours	**86.9**	**87.5**	**40.8**	**9.5**	**82.5**
Qwen2VL-7B	87.0	87.5	30.4	8.4	**72.5**
+ Vissink[ICLR25](Kang et al., 2025b)	87.4	88.2	30.0	8.2	71.5
+ Ours	**88.0**	**88.7**	**28.4**	**8.1**	71.7

5.3 Ablation Studies

Steering Keys $vs.$ Values. Asymmetry 1 places the routing–content gap primarily on the Key side, so our design in §4 amplifies the Key residual only. Table 3 supports this choice: amplifying Keys yields gains across all three benchmarks (*e.g.*, MME: $1504.8 \rightarrow 1520.6$, MMVP: $3.33 \rightarrow 12.0$), while Value-only amplification produces smaller gains. Compared with Key-only steering, joint amplification slightly improves MME ($1522.3\ vs.\ 1520.6$) but lowers MMVP ($11.3\ vs.\ 12.0$) and POPE ($86.5\ vs.\ 86.9$). Thus, adding Value steering provides no consistent further benefit across these benchmarks.

Table 3: Ablation on Key-side, Value-side, and joint amplification. All variants use $\mathcal{L} = \{23, \ldots, 31\}$ on LLaVA-1.5-7B.

Method	MME	MMVP	POPE Acc↑
Baseline	1504.8	3.33	84.0
Keys only	<u>1520.6</u>	**12.0**	**86.9**
Values only	1505.4	4.6	84.2
Keys & Values	**1522.3**	<u>11.3</u>	<u>86.5</u>

Amplifying the residual, not suppressing the prior. Eq. 11 amplifies the instance-aware residual $\mathbf{r}_K^{(l)}$ while keeping the prior $\boldsymbol{\mu}_K^{(l)}$ intact. Table 4 compares it against all alternative ways of perturbing the image Key. Changing the prior alone degrades performance in the tested settings, whether $\boldsymbol{\mu}_K^{(l)}$ is suppressed or amplified, confirming that the prior acts as a stable, input-agnostic routing scaffold (Fig. 2). Joint scaling yields a small gain on MME but no improvement on MM-Vet (*i.e.*, $29.9\ vs.\ 29.9$), and a much smaller gain than ours on CV-Bench^{2D} (*i.e.*, $+1.5\ vs.$ our $+4.6$), consistent with multiplicative attention sharpening that helps perception decisions but not fine-grained grounding. Selectively amplifying the residual delivers consistent gains across all three benchmarks, with the largest improvements on the grounding-sensitive MM-Vet and CV-Bench^{2D}, indicating that the instance-aware component is an effective target for closing the routing–content gap.

Table 4: Comparison of all paths to modifying the image Key in deep layers. α scales the residual $\mathbf{r}_K^{(l)}$ and β scales the prior $\boldsymbol{\mu}_K^{(l)}$. All variants use LLaVA-1.5-7B with $\mathcal{L} = \{23, \ldots, 31\}$.

Method	MME	MM-Vet	CV-Bench^{2D}
Baseline ($\alpha = 1, \beta = 1$)	1504.8	29.9	56.8
Suppress prior ($\alpha = 1, \beta = 0.5$)	1501.2	29.6	56.4
Suppress prior ($\alpha = 1, \beta = 0$)	1463.4	27.1	52.6
Amplify prior ($\alpha = 1, \beta = 1.5$)	1494.0	28.0	56.1
Joint scaling ($\alpha = 1.5, \beta = 1.5$)	**1524.4**	<u>29.9</u>	<u>58.3</u>
Ours ($\alpha = 1.5, \beta = 1$)	<u>1520.6</u>	**32.2**	**61.4**

Sensitivity to $\mathcal{L}$. Asymmetry 2 identifies deep layers as the regime where the Key residual fades, but Values are richest, motivating our restriction of $\mathcal{L}$ to this range. Table 5 confirms this prediction: deep-layer steering clearly outperforms mid-layer steering (*e.g.*, MME: $1556.6\ vs.\ 1359.6$), while shallow-layer steering corrupts the model entirely. The shallow-layer collapse further reinforces our analysis: shallow Keys are heavily prior-dominated (Asymmetry 1), so multiplicative amplification distorts the input-agnostic routing scaffold rather than enriching the residual, and this distortion compounds through the entire decoder.

Table 5: Sensitivity to $\mathcal{L}$ on LLaVA-1.5-13B with $\alpha = 1.8$. ✗ denotes model corruption.

$\mathcal{L}$	MME	POPE Acc
Shallow ($\{0, \ldots, 11\}$)	✗	✗
Mid ($\{12, \ldots, 23\}$)	1359.6	72.33
Deep ($\{24, \ldots, 39\}$)	**1556.6**	**87.48**

6 Conclusion

In this work, we investigate the underlying signals within image token representations that drive text-to-image (T2I) attention in multimodal large language models (MLLMs). Through an ANOVA-style decomposition of image Keys and Values across decoder layers, we identify two consistent Key–Value asymmetries. Within each layer, image Keys are dominated by an input-agnostic visual prior while Values are dominated by an instance-aware residual. Across layers, this asymmetry widens with depth: the instance-aware residual grows steadily in Values but fades in deep-layer Keys. Together, these asymmetries reveal a systematic gap between routing and content on the Key side of T2I attention. Building on this insight, we propose `Image Key Steering` as a training-free intervention that amplifies the instance-aware residual of image Keys in deep layers. We demonstrate that this approach consistently improves vision-language understanding across diverse benchmarks, particularly for visual grounding and hallucination reduction. Ultimately, our results suggest that the instance-aware component of image Keys is an important factor in T2I attention, offering a representation-level perspective for improving vision-language understanding in MLLMs.

AI USE STATEMENT

In preparing this manuscript, generative AI tools were used for language editing. All AI-assisted edits were reviewed and verified by the authors, who take responsibility for the final content, claims, and artifacts.

ETHICS STATEMENT

This work introduces no new training data or model weights and primarily inherits the ethical profile of the underlying MLLMs. Improved visual grounding can benefit reliability-critical applications, but it may also amplify dual-use risks such as surveillance or automated visual data mining. We discuss these risks and responsible deployment considerations in Appendix §S9 and Appendix §S12.

REPRODUCIBILITY STATEMENT

The method is fully specified in §4. Model-specific hyperparameters, calibration details, benchmark protocols, compute resources, and additional ablations are documented in Appendices §S1–S11.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022.

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *CVPR*, 2025.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015.

Lorenzo Basile, Valentino Maiorca, Diego Doimo, Francesco Locatello, and Alberto Cazzaniga. Head pursuit: Probing attention specialization in multimodal transformers. In *NeurIPS*, 2025.

Jing Bi, Junjia Guo, Yunlong Tang, Lianggong Bruce Wen, Zhang Liu, Bingjie Wang, and Chenliang Xu. Unveiling visual perception in language models: An attention head analysis approach. In *CVPR*, 2025.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *ECCV*, 2024.

Jiho Choi, Jaemin Kim, Sanghwan Kim, Seunghoon Hong, and Jin-Hwi Park. When sinks help or hurt: Unified framework for attention sink in large vision-language models. *arXiv preprint arXiv:2604.03316*, 2026.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of machine learning research*, 24(240):1–113, 2023.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. In *NeurIPS*, 2023.

Fenil R Doshi, Thomas Fel, Talia Konkle, and George Alvarez. Bi-orthogonal factor decomposition for vision transformers. *arXiv preprint arXiv:2601.05328*, 2026.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Yingqi Fan, Junlong Tong, Anhao Zhao, and Xiaoyu Shen. What do visual tokens really encode? uncovering sparsity and redundancy in multimodal large language models. *arXiv preprint arXiv:2603.00510*, 2026.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *NeurIPS Datasets and Benchmarks Track*, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Jiayi Han, Liang Du, Yiwen Wu, Guanming Liang, Xiangguo Zhou, Weibo Zheng, Donghong Han, and Zixun Sun. Adav: Adaptive text-visual redirection for vision-language models. In *ACL Findings*, 2025.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022.

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 *CVPR*, 2024.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019.

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-introspective decoding: Alleviating hallucinations for large vision-language models. In *ICLR*, 2025.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *CVPR*, 2025.

Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the image? a deep-dive into the vision of vision language models. In *CVPR*, 2025.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. Your large vision-language model only needs a few attention heads for visual grounding. In *CVPR*, 2025a.

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

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *CVPR*, 2023.

Jinyeong Kim, Seil Kang, Jiwoo Park, Junhyeok Kim, and Seong Jae Hwang. Interpreting attention heads for image-to-text information flow in large vision-language models. *arXiv preprint arXiv:2509.17588*, 2025.

Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *nature*, 401(6755):788–791, 1999.

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 *CVPR*, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023a.

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

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. VILA: on pre-training for visual language models. In *CVPR*, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

Zhihang Lin, Mingbao Lin, Luxi Lin, and Rongrong Ji. Boosting multimodal large language models with visual tokens withdrawal for rapid inference. In *AAAI*, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *ECCV*, 2024a.

Shuliang Liu, Songbo Yang, Dong Fang, Sihang Jia, Yuqi Tang, Lingfeng Su, Ruoshui Peng, Yibo Yan, Xin Zou, and Xuming Hu. Vision-language introspection: Mitigating overconfident hallucinations in mllms via interpretable bi-causal steering. *arXiv preprint arXiv:2601.05159*, 2026.

Yiyang Liu, James Chenhao Liang, Ruixiang Tang, Yugyung Lee, Majid Rabbani, Sohail A. Dianat, Raghuveer Rao, Lifu Huang, Dongfang Liu, Qifan Wang, and Cheng Han. Re-imagining multimodal instruction tuning: A representation view. In *ICLR*, 2025.

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 *ECCV*, 2024b.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, 2022.

Kaspar Märtens and Christopher Yau. Neural decomposition: Functional ANOVA with variational autoencoders. In *AISTATS*, 2020.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, 2021.

Xu Pan, Aaron Philip, Ziqian Xie, and Odelia Schwartz. Dissecting query-key interaction in vision transformers. In *NeurIPS*, 2024.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *EMNLP*, 2018.

Andrew M Saxe, James L McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23): 11537–11546, 2019.

Ying Shen, Zhiyang Xu, Qifan Wang, Yu Cheng, Wenpeng Yin, and Lifu Huang. Multimodal instruction tuning with conditional mixture of lora. In *ACL*, 2024.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, 2019.

Jiajun Song and Yiqiao Zhong. Uncovering hidden geometry in transformers via disentangling position and context. *arXiv preprint arXiv:2310.04861*, 2023.

Feilong Tang, Zile Huang, Chengzhi Liu, Qiang Sun, Harry Yang, and Ser-Nam Lim. Intervening anchor token: Decoding strategy in alleviating hallucinations for mllms. In *ICLR*, 2025.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *NeurIPS*, 2024a.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR*, 2024b.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

Zifu Wan, Ce Zhang, Silong Yong, Martin Q Ma, Simon Stepputtis, Louis-Philippe Morency, Deva Ramanan, Katia Sycara, and Yaqi Xie. Only: One-layer intervention sufficiently mitigates hallucinations in large vision-language models. In *ICCV*, 2025.

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

Taowen Wang, Yiyang Liu, James Chenhao Liang, Junhan Zhao, Yiming Cui, Yuning Mao, Shaoliang Nie, Jiahao Liu, Fuli Feng, Zenglin Xu, et al. M2pt: Multimodal prompt tuning for zero-shot instruction learning. In *EMNLP*, 2024b.

Hongxuan Wu, Yukun Zhang, and Xueqing Zhou. How vision becomes language: A layer-wise information-theoretic analysis of multimodal reasoning. *arXiv preprint arXiv:2602.15580*, 2026.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D Manning, and Christopher Potts. Reft: Representation finetuning for language models. In *NeurIPS*, 2024.

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

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Zhuoran Yu and Yong Jae Lee. How multimodal llms solve image tasks: A lens on visual grounding, task reasoning, and answer decoding. In *COLM*, 2025.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. Mllms know where to look: Training-free perception of small visual details with multimodal llms. In *ICLR*, 2025a.

Xiaofeng Zhang, Yihao Quan, Chen Shen, Chaochen Gu, Xiaosong Yuan, Shaotian Yan, Jiawei Cao, Hao Cheng, Kaijie Wu, and Jieping Ye. Shallow focus, deep fixes: Enhancing shallow layers vision attention sinks to alleviate hallucination in lvlms. In *EMNLP*, 2025b.

Xuange Zhang, Dengjie Li, Bo Liu, Zenghao Bao, Yao Zhou, Baisong Yang, Zhongying Liu, Yujie Zhong, and Tongtong Yuan. Layer-wise vision injection with disentangled attention for efficient lvlms. In *ICCV*, 2025c.

SUMMARY OF THE APPENDIX

This appendix contains additional experimental results and discussions of our ICLR 2027 submission: *Image Keys and Values are Asymmetric in MLLM Text-to-Image Attention*, organized as follows:

- §S1 provides detailed **descriptions of all evaluation benchmarks** used in our experiments.
- §S2 explains the **implementation details** of Image Key Steering.
- §S3 provides an **actual-Query logit and attention-distribution analysis**.
- §S4 presents the T2I decomposition analysis under **dynamic resolution**.
- §S5 investigates the **origin of the Key–Value asymmetry** in the visual frontend versus the decoder.
- §S6 discusses how the **routing–content gap relates to visual token pruning**.
- §S7 provides **ablation experiments** on the amplification factor, calibration distribution, and a norm-matched cross-image residual control.
- §S8 provides the **qualitative analysis** of our method.
- §S9 includes additional discussions on **ethics concerns**.
- §S10 shows related **asset licenses and usage terms** of our work.
- §S11 discusses the **reproducibility** of our approach.
- §S12 discusses the **social impact, limitations**, and directions for **future work**.

S1 BENCHMARK DETAILS

General vision-language understanding. We use six benchmarks to comprehensively assess multimodal competence across perception, reasoning, and knowledge-intensive question answering.

- **MME** (Fu et al., 2025) probes fourteen distinct capabilities ranging from object and scene recognition to OCR and commonsense reasoning; we report scores on its perception subset.
- **LLaVA-Bench (In-the-Wild)** (Liu et al., 2023) presents sixty questions over twenty-four images drawn from varied contexts such as paintings, memes, and outdoor scenes, with responses scored by GPT-4 to capture open-ended generalization ability.
- **MM-Vet** (Yu et al., 2023) similarly uses GPT-4 to judge the accuracy and helpfulness of model answers across 218 questions over 200 images.
- **MMBench-English** (Liu et al., 2024b) contains approximately 3,000 multiple-choice questions spanning twenty ability dimensions; each question is tested under circular permutations of the answer choices, with GPT-4 used to align free-form predictions to the option labels.
- **ScienceQA** (Lu et al., 2022) is a large-scale multiple-choice dataset covering elementary and high school science across three subjects and 379 skills; we evaluate on the subset of questions that include image context.
- **SEED-Bench** (Li et al., 2023a) assesses spatial and semantic image understanding through multiple-choice questions; we use its image-only partition.

Vision-centric reasoning. Two benchmarks specifically target fine-grained visual perception, where correctly answering questions demands sensitivity to subtle image-level details rather than reliance on language priors.

- **MMVP** (Tong et al., 2024b) is built around CLIP-blind image pairs—visually distinct images that the CLIP encoder maps to nearly identical representations—and comprises 300 question-answer pairs designed to expose failures in low-level visual grounding.
- **CV-Bench** (Tong et al., 2024a) repurposes ground-truth annotations from ADE20k, COCO, and OMNI3D into 2,638 multiple-choice questions that test both 2D understanding (spatial relationships and object counting) and 3D understanding (depth ordering and relative distance estimation).

Visual hallucination. Two complementary benchmarks quantify how faithfully model outputs are grounded in the visual input.

- **CHAIR** (Rohrbach et al., 2018) evaluates object hallucination in free-form captions by comparing the objects mentioned in generated descriptions against the ground-truth annotations. It reports two complementary metrics: CHAIR_I measures the fraction of hallucinated objects among all objects mentioned, while CHAIR_S measures the fraction of captions that contain at least one hallucinated object. Formally,

$$\text{CHAIR}_I = \frac{|\{\text{halluc. objects}\}|}{|\{\text{mentioned objects}\}|}, \quad \text{CHAIR}_S = \frac{|\{\text{halluc. captions}\}|}{|\{\text{all captions}\}|}. \tag{12}$$

  Following the standard evaluation protocol established by OPERA (Huang et al., 2024) and adopted by subsequent work (Zhang et al., 2025b), we evaluate on the same 500 images sampled from the MSCOCO val2014 split (Lin et al., 2014), which is annotated with 80 object categories. For each image, we use the fixed query prompt *"Please describe this image in detail."* to elicit free-form captions from the model.
- **POPE** (Li et al., 2023b) queries models with binary yes/no questions about object presence across 500 images from MSCOCO, A-OKVQA, and GQA. Negative samples are constructed under three strategies—random, popular, and adversarial—and performance is reported in terms of accuracy, precision, recall, and F1 score.

GPT-4 Evaluation. LLaVA-Bench (In-the-Wild) and MM-Vet both rely on GPT-4 as an automated judge. As `gpt-4-0613`, which is commonly used in prior work, **has been deprecated**, we instead use `gpt-4o-2024-11-20` as the evaluator, with a maximum output length of 1,024 tokens.

S2 Implementation Details

Models and Tokenization. We evaluate `Image Key Steering` on four representative MLLMs spanning two parameter scales and two visual tokenization regimes. LLaVA-1.5-7B, LLaVA-1.5-13B (Liu et al., 2023), and VILA-13B (Lin et al., 2024) encode each image into a fixed number of visual tokens, whereas Qwen2VL-7B (Wang et al., 2024a) produces a variable number of visual tokens depending on the input resolution. This selection allows us to examine whether our intervention generalizes across both fixed-length and dynamic-resolution tokenization paradigms, as well as across model capacities.

Steering Layer Set $\mathcal{L}$. For each model, the steering layer set $\mathcal{L}$ is selected once from the layer-wise representation analysis in §3.2, without using benchmark performance, and is then frozen across all benchmarks. Concretely, for LLaVA-1.5-7B (32 decoder layers in total) we use $\mathcal{L} = \{23, \ldots, 31\}$; for LLaVA-1.5-13B (40 layers) we use $\mathcal{L} = \{24, \ldots, 39\}$; for VILA-13B (40 layers) we use $\mathcal{L} = \{22, \ldots, 39\}$; and for Qwen2VL-7B (28 layers) we use $\mathcal{L} = \{21, \ldots, 27\}$.

Amplification Factor α. The amplification factor α is selected at the category level rather than per benchmark. For each model, we use a randomly sampled 10% subset of MME, CV-Bench^{2D}, and POPE to select α for general vision-language understanding, vision-centric reasoning, and visual hallucination, respectively, using each benchmark's primary metric. Once selected, α is frozen and applied unchanged to every benchmark in the corresponding category; no benchmark receives an individually tuned value. The specific values used per model are as follows. For LLaVA-1.5-7B, we use $\alpha = 1.5$ for general understanding and vision-centric reasoning, and $\alpha = 2.4$ for visual hallucination. For LLaVA-1.5-13B, α falls a relatively small range $[1.8, 2.0]$ across all three categories. For Qwen2VL-7B, α also falls in a relatively small range $[1.1, 1.3]$ across all categories. For VILA-13B, we use $\alpha = 4.0$ across all categories. Table S1 evaluates the selected LLaVA-1.5-7B configurations on the remaining 90% of the three selection benchmarks, while Table S3 shows robustness over a broad range of α.

Table S1: Evaluation on the held-out 90% of each α-selection benchmark. The 10% subsets used for selection are excluded from these scores.

Category	Benchmark / Metric	Selected α	Base	+ Steering
General vision-language	MME ↑	1.5	1507.3	**1522.4**
Vision-centric reasoning	CV-Bench^{2D} ↑	1.5	56.6	**61.7**
Visual hallucination	POPE Acc ↑	2.4	83.5	**86.6**

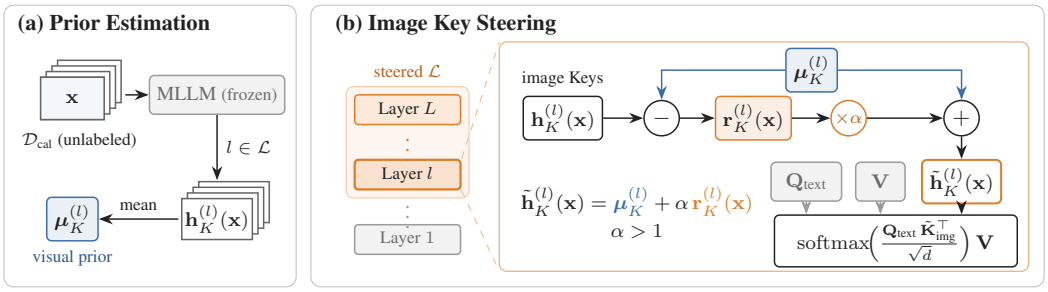

Figure S1: **Overview of `Image Key Steering`. (a)** The visual prior $\boldsymbol{\mu}_K^{(l)}$ is estimated once, offline, by averaging the image Key projections $\mathbf{h}_K^{(l)}(\mathbf{x})$ of a frozen MLLM over a small unlabeled image set $\mathcal{D}_{\text{cal}}$ (a single forward pass per image). **(b)** At inference, in each deep layer $l \in \mathcal{L}$, the instance-aware residual $\mathbf{r}_K^{(l)}(\mathbf{x}) = \mathbf{h}_K^{(l)}(\mathbf{x}) - \boldsymbol{\mu}_K^{(l)}$ is amplified by $\alpha > 1$ and added back to the prior (Eq. 11). The steered image Keys then enter text-to-image attention. Text Queries, all Values, all other layers and all weights are unchanged (gray), so the method is training-free.

The improvements persist when the selection examples are removed from evaluation: $+15.1$ points on MME, $+5.1$ on CV-Bench^{2D}, and $+3.1$ on POPE accuracy. Separately, Table 3 uses one fixed configuration, $\alpha = 1.5$ and $\mathcal{L} = \{23, \ldots, 31\}$, across MME, MMVP, and POPE, demonstrating that the method remains effective without benchmark-specific α values.

Visual Prior Estimation. The input-agnostic visual prior $\boldsymbol{\mu}_K^{(l)}$ at each steering layer $l \in \mathcal{L}$ is estimated from a small unlabeled image set sampled from the benchmark's own image distribution via a single forward pass, following Eq. 6 and Eq. 7. This calibration uses raw images only, never questions, answers, labels, model predictions, or evaluation metrics. Because the images come from the benchmark's own image pool, the primary protocol is unsupervised but transductive. We use 200 images for each benchmark, with the sole exception of LLaVA^{W} (Liu et al., 2023), whose complete image set contains only 24 images; in this case we use all 24. Table S4 compares task-specific and COCO calibration on three benchmarks. Table S5 extends the comparison to five additional settings across three model families, using one fixed set of 200 COCO images disjoint from all evaluation images and keeping α and $\mathcal{L}$ unchanged. We retain task-specific calibration as the primary protocol and report these task-agnostic controls separately.

Inference and Evaluation Variance. `Image Key Steering` is applied purely at inference time: it does not modify any model weights, requires no gradient computation, and introduces no additional learnable parameters. The only modification is the rescaling of the instance-aware residual of image Keys at the selected layers, as defined in Eq. 11. The computational overhead is negligible, as the prior $\boldsymbol{\mu}_K^{(l)}$ is precomputed once and steering amounts to an elementwise operation on the image Key projections. All benchmark generations use sampling-disabled decoding (`do_sample=False`; temperature 0 where applicable), so repeated generation under a fixed configuration is deterministic. For GPT-judged benchmarks, we repeat each evaluation three times; the standard deviation is at most 0.1 in most settings and does not exceed 0.2 in any setting, while the gains range from 0.9 to 1.7 points on LLaVA^{W} and from 0.4 to 2.3 points on MM-Vet. The maximum number of newly generated tokens is set to 512 for CHAIR (Rohrbach et al., 2018).

S3 Actual-Query Logit and Attention-Distribution Analysis

Representation energy alone does not determine dot-product attention: a high-energy Key component may be weakly aligned with the model's Queries, while a lower-energy component may still change the resulting distribution. We therefore repeat the prior–residual analysis after projecting the decomposed image Keys onto the model's actual post-RoPE text Queries. For an image Key, the T2I logit decomposes as

$$z_{\text{img}} = \frac{\mathbf{q}^{\top}\mathbf{K}}{\sqrt{d}} = \frac{\mathbf{q}^{\top}\boldsymbol{\mu}_K}{\sqrt{d}} + \frac{\mathbf{q}^{\top}\mathbf{r}_K}{\sqrt{d}}. \tag{13}$$

We use 32 calibration images and 16 evaluation images with disjoint identities from MMVP. We measure the squared logit contribution of each component and the resulting attention distributions. For the distribution controls, all non-image Keys remain unchanged; lower Jensen–Shannon (JS) divergence indicates greater similarity to the original attention distribution.

Table S2: Actual-Query decomposition of image-Key logits and attention distributions on LLaVA-1.5-7B. The two JS entries in each cell correspond to the prior-only and residual-only controls, respectively; lower is more similar to the original attention distribution.

Layers	Prior logit share	Prior/residual RMS	Cross term	Full-context JS	Image-conditional JS
All	97.3%	5.95×	−1.2%	0.020 / 0.521	0.280 / 0.092
23–31	97.5%	6.22×	−1.7%	0.016 / 0.550	0.259 / 0.071

Across all layers, the prior accounts for 97.3% of the squared component logits and has $5.95\times$ the residual RMS. In layers 23–31, these values increase to 97.5% and $6.22\times$, respectively, while the signed cross term remains small at -1.7%. Thus, the prior's representation-level dominance persists after projection onto the model's actual Queries rather than disappearing through weak Query alignment.

The distributional controls reveal complementary roles for the two components. The full-context attention distribution is substantially closer to the prior-only distribution than to the residual-only distribution, particularly in layers 23–31 (JS: 0.016 versus 0.550). Under the image-conditional comparison, this relationship reverses: the original allocation among image tokens is closer to the residual-only distribution than to the prior-only distribution (JS: 0.071 versus 0.259). These results support a more precise interpretation: the prior supplies the dominant full-context routing scaffold, while the smaller residual more closely captures instance-aware refinement among image tokens. This logit-level evidence complements the controlled interventions in §5.3.

S4 T2I Decomposition under Dynamic Resolution

To verify that the two Key–Value asymmetries identified in §3.2 are not artifacts of LLaVA-1.5-7B's fixed-length visual tokenization, we replicate our decomposition on Qwen2VL-7B (Wang et al., 2024a), which uses a dynamic-resolution visual encoder that produces a variable number of image tokens per input. We use the global-average-pooled prior in Eq. 7 for all Qwen2VL results, with a calibration set $\mathcal{D}_{\text{cal}}$ sampled from MME (Fu et al., 2025). Figs. S2 and S3 report the energy fractions and absolute energies across the 28 decoder layers of Qwen2VL-7B.

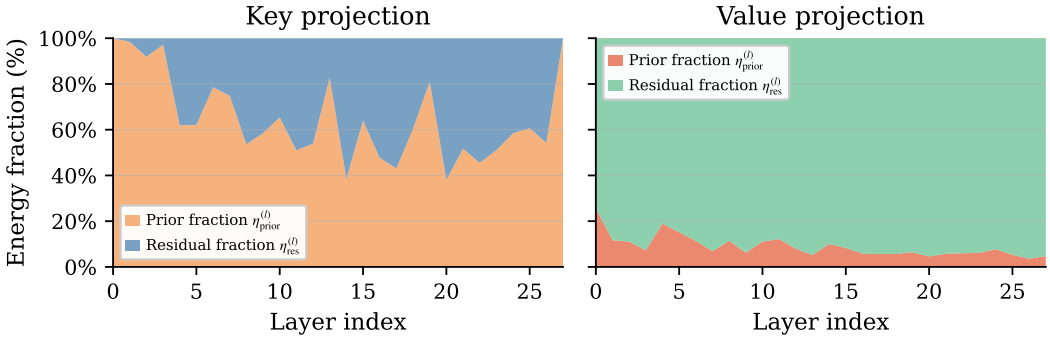

Figure S2: **Energy fractions of image Key and Value projections across decoder layers on Qwen2VL-7B.** Image Keys (*Left*) have substantially larger prior fractions than Values, with $\eta^{(l)}_{\text{prior},K}$ exceeding 95% at layers 0, 1, and 27 and reaching around 80% at intermediate peaks (13, 19). Image Values (*Right*) are overwhelmingly residual-dominated, with $\eta^{(l)}_{\text{res},V}$ above 74% at every layer and reaching 93–96% in deeper layers.

Asymmetry 1: Keys have larger prior fractions than Values. Qwen2VL-7B also exhibits a pronounced contrast between Key and Value composition. As shown in Fig. S2 (*Right*), the residual

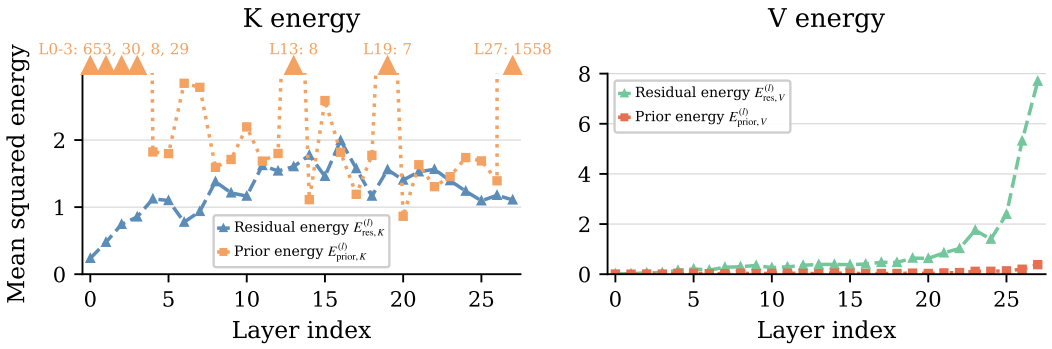

Figure S3: **Absolute energies of image Key and Value projections across decoder layers on Qwen2VL-7B.** (*Left*) The Key prior energy $E^{(l)}_{\text{prior},K}$ exhibits extreme spikes at the first few layers (L0–3: 653, 30, 8, 29), at two intermediate layers (L13: 7.67, L19: 6.61), and at the final layer (L27: 1558), while the residual energy $E^{(l)}_{\text{res},K}$ stays bounded (peaking near 2.0 at layer 16). (*Right*) The Value residual energy $E^{(l)}_{\text{res},V}$ grows nearly monotonically with depth, accelerating after layer 20 and reaching 7.70 at the final layer, while the Value prior energy stays below 0.4 throughout.

fraction of image Values exceeds 74% at every layer and climbs above 93% for $l \geq 13$, reaching 96.4% at layer 26. On the Key side (see Fig. S2, *Left*), the prior fraction $\eta^{(l)}_{\text{prior},K}$ exceeds 95% at layers 0, 1, and 27, reaches around 80% at layers 13 and 19, and lies around 40–60% across many intermediate layers. Qwen2VL uses a prior pooled across token positions, while LLaVA uses a separate prior at each token position. Across layers, Keys consistently contain a larger prior fraction than Values, while Values remain strongly residual-dominated.

Asymmetry 2 holds: instance-aware content accumulates in Values but not in deep-layer Keys. The across-layer evolution is even more pronounced on Qwen2VL-7B. As shown in Fig. S3 (*Right*), $E^{(l)}_{\text{res},V}$ grows nearly monotonically from 0.019 at layer 0 to 7.70 at layer 27, with a $\sim 9\times$ jump in the last 7 layers alone. The Key residual $E^{(l)}_{\text{res},K}$ follows a very different trajectory (see Fig. S3, *Left*): it rises to a modest peak of 1.99 at layer 16 and then drifts downward, settling near 1.1 at layer 27. As on LLaVA, the instance-aware residual of Keys plateaus and decays in precisely the layers where Values are at their richest.

Boundary and intermediate Key-prior spikes. A feature of Qwen2VL-7B that is more pronounced than in LLaVA is the presence of extreme Key-prior energy spikes at specific layers: $E^{(l=0)}_{\text{prior},K} = 653$ and $E^{(l=27)}_{\text{prior},K} = 1558$ are two to three orders of magnitude above mid-network values of 1–3, and two intermediate layers (L13: 7.67, L19: 6.61) show smaller but distinct spikes. At the boundary layers 0, 1, and 27, the residual accounts for less than 5% of total Key energy (*e.g.*, 0.0% at L0 and 0.1% at L27). At L13 and L19, the prior fraction is around 80%, with a correspondingly larger residual share. The absolute residual energy remains bounded across these layers (Fig. S3), so the largest Key-energy spikes are concentrated in the shared prior.

Summary. Qwen2VL-7B exhibits the same broad contrast between Keys and Values: Keys have substantially larger prior fractions than Values, and their residual energy declines after its mid-layer peak, while Values are overwhelmingly residual-dominated and accumulate instance-aware energy nearly monotonically with depth. These findings extend the composition and depth-wise contrasts observed on LLaVA-1.5-7B to a model with dynamic-resolution visual tokenization.

S5 Origin of the Key–Value Asymmetry

The Key–Value asymmetry documented in §3.2 shows that image Keys are prior-dominated while image Values are residual-dominated throughout the decoder. We investigate whether this asymmetry is already present in the upstream visual representations or is introduced by the decoder itself. We

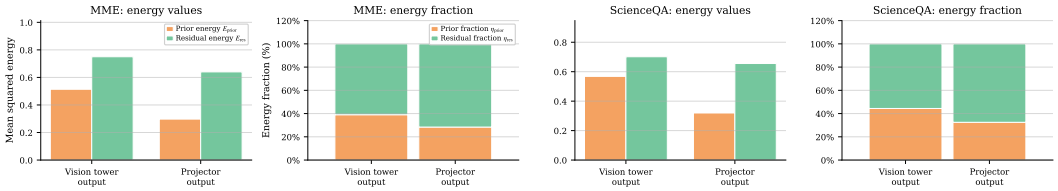

Figure S4: **Decomposition of upstream visual representations on LLaVA-1.5-7B.** We apply the prior/residual decomposition (Eq. 4) to image token representations at the vision tower output (second last ViT layer) and the projector output, on MME (*Left two*) and ScienceQA (*Right two*). At both stages and on both benchmarks, the instance-aware residual dominates the total energy (55–72%), with the projector further increasing the residual fraction. This contrasts sharply with the decoder's Key projections, where the prior accounts for over 78% of the energy across all layers (see Fig. 1), indicating that the Key–Value asymmetry originates in the decoder rather than the visual frontend.

apply the same energy decomposition of Eq. 4 to image token representations at two pre-decoder modules in LLaVA-1.5-7B, namely the second last layer output of the vision tower and the output of the multimodal projector. As shown in Fig. S4, on both MME and ScienceQA, the instance-aware residual already accounts for the majority of the total energy at the vision tower output (61.0% and 55.6%, respectively), and this fraction further increases after the projector (71.6% and 67.5%). The representations entering the decoder are thus predominantly instance-aware, and the projector amplifies rather than suppresses this property. Once these same representations pass through the decoder's Key projection $\mathbf{W}_K^{(l)}$, however, the picture reverses: the prior-energy fraction of image Keys jumps to roughly 88% at layer 0 and remains above 78% across all 32 layers. The Value projections, by contrast, preserve the residual-dominated composition of their input throughout (see Fig. 1). Since both projections receive the same input, the asymmetry is introduced specifically by $\mathbf{W}_K^{(l)}$.

Although a full explanation of this asymmetry remains an open question, our findings consistently point to the decoder's Key projection as the source of the routing–content gap, rather than the visual frontend. Image Key Steering is designed to intervene at precisely this point (§4).

S6 Connection to Visual Token Pruning

Our two asymmetries offer a possible explanation for recent visual token pruning methods (Chen et al., 2024; Lin et al., 2025; Fan et al., 2026; Han et al., 2025), which downsample image tokens in early layers and withdraw them in deep layers with little performance loss. In early-to-middle layers, image Keys are prior-dominated and only a small residual distinguishes tokens. In deep layers this residual further diminishes, so the routing channel carries little instance-aware information to lose even though Values remain rich. Pruning exploits this redundancy to reduce compute, whereas `Image Key Steering` (§4) takes the opposite route and strengthens the instance-aware component of Keys directly.

S7 Additional Ablation Experiments

This section complements the analyses of `Image Key Steering` in §5.3 with studies of amplification strength, calibration distribution, and image-specific residual correspondence under matched perturbation norms.

Sensitivity to α. The amplification factor α in Eq. 11 controls how strongly the instance-aware Key residual is boosted. Table S3 examines its effect on LLaVA-1.5-13B with $\mathcal{L} = \{24, \ldots, 39\}$. All three benchmarks exhibit a consistent trend: performance improves as α increases from 1.0 to 1.8 (*e.g.*, MME: 1529.9 → 1556.6, MM-Vet: 34.9 →

Table S3: Sensitivity to α on LLaVA-1.5-13B with $\mathcal{L} = \{24, \ldots, 39\}$. $\alpha = 1.0$ corresponds to the original model. MM-Vet results are reported as mean±std over 3 runs to account for the variance of GPT-based evaluation.

Metric	$\alpha = 1.0$	$\alpha = 1.2$	$\alpha = 1.4$	$\alpha = 1.6$	$\alpha = 1.8$	$\alpha = 2.0$
MME	1529.9	1532.6	1541.9	1545.9	**1556.6**	1536.0
POPE Acc	85.47	87.15	87.32	87.41	**87.48**	87.01
MM-Vet	$34.9_{\pm 0.1}$	$34.9_{\pm 0.1}$	$35.1_{\pm 0.2}$	$35.6_{\pm 0.1}$	$\mathbf{35.8}_{\pm 0.1}$	$35.4_{\pm 0.2}$

35.8), peaks at $\alpha = 1.8$, and drops slightly at $\alpha = 2.0$ while still remaining above the baseline. For MM-Vet, we report mean and standard deviation over 3 independent runs to account for the inherent variance of the GPT-4 judge. The gain from $\alpha = 1.0$ to the optimal $\alpha = 1.8$ is substantially larger than the observed run-to-run variation (*i.e.*, $+0.9$ $vs.$ ≤ 0.2). Overall, `Image Key Steering` is robust across a wide range of amplification strengths.

Table S4: Robustness to calibration distribution. Both sets contain 200 images; the task-agnostic set is randomly sampled from COCO val (Lin et al., 2014). All results use LLaVA-1.5-7B with $\alpha = 1.5$ and $\mathcal{L} = \{23, \ldots, 31\}$.

Calibration Set	MME	SEED^{I}	CV-Bench$^{\text{2D}}$
Task-specific	1520.6	61.1	61.4
COCO val	1518.9	60.8	61.2

Robustness to calibration distribution. Since the Key prior is input-agnostic (Fig. 2), we expect its estimate to be relatively insensitive to moderate calibration-distribution shifts. Replacing the task-specific images with 200 COCO val (Lin et al., 2014) images changes performance only slightly on these three benchmarks (*e.g.*, MME: 1520.6 $vs.$ 1518.9, CV-Bench$^{\text{2D}}$: 61.4 $vs.$ 61.2), suggesting that the prior estimate transfers across these image sources.

Table S5: Extended calibration control. The task-specific and disjoint COCO columns report Image Key Steering with the same α and $\mathcal{L}$. A single fixed set of 200 COCO images, disjoint from all evaluation images, is used across all five settings.

Model	Benchmark / Metric	Base	Task-specific	Disjoint COCO
LLaVA-1.5-7B	LLaVA^{W} ↑	59.2	60.9	60.7
LLaVA-1.5-7B	MMVP ↑	3.33	12.0	11.3
LLaVA-1.5-7B	CHAIR$_I$ ↓	15.2	9.8	9.8
Qwen2VL-7B	MME ↑	1684.6	1690.9	1693.5
VILA-13B	MME ↑	1552.6	1578.6	1577.6

We extend this check to LLaVA^{W}, MMVP, and CHAIR on LLaVA-1.5-7B, and to MME on Qwen2VL-7B and VILA-13B (Table S5). The same fixed set of 200 COCO images is used across all five settings and is disjoint from all evaluation images; α and $\mathcal{L}$ remain unchanged to isolate the calibration choice. Gains over the base model persist in every added setting: MMVP retains an improvement from 3.33 to 11.3, CHAIR$_I$ remains at 9.8, and both additional model families improve on MME. Together, Tables S4 and S5 cover eight settings across three model families, supporting the transfer of prior estimates across the tested calibration sources.

S7.1 Norm-Matched Cross-Image Residual Control

To test whether steering benefits from the residual's correspondence with the current image, we replace the added residual with one from a different image while matching the perturbation norm. For an image $\mathbf{x}_i$, we take a donor residual from $\mathbf{x}_j$, $j \neq i$, at the same layer and image-token positions, and rescale it to match the Frobenius norm of the current image's residual:

$$\widetilde{\mathbf{h}}_K^{(l)}(\mathbf{x}_i) = \mathbf{h}_K^{(l)}(\mathbf{x}_i) + (\alpha - 1)\widehat{\mathbf{r}}_K^{(l)}(\mathbf{x}_j), \qquad j \neq i, \tag{14}$$
$$\left\|\widehat{\mathbf{r}}_K^{(l)}(\mathbf{x}_j)\right\|_F = \left\|\mathbf{r}_K^{(l)}(\mathbf{x}_i)\right\|_F.$$

We use LLaVA-1.5-7B with $\alpha = 1.5$ and $\mathcal{L} = \{23, \ldots, 31\}$, keeping the calibrated prior and all other settings identical between the instance-matched and cross-image conditions. This matches the norm of the added perturbation; the resulting Key norms and attention distributions may still differ. We also include generic image-Key scaling, $\widetilde{\mathbf{h}}_K^{(l)} = 1.5\mathbf{h}_K^{(l)}$, in the same layers.

As shown in Table S6, the cross-image residual leaves both scores essentially at baseline, whereas instance-matched steering improves MME by 15.8 points and CV-Bench$^{\text{2D}}$ by 4.6 points. Generic

Table S6: Residual correspondence at matched perturbation norm on LLaVA-1.5-7B. The cross-image control replaces the added instance-matched residual with an equally large perturbation derived from another image.

Intervention	MME ↑	CV-Bench^{2D} ↑
Base	1504.8	56.8
Generic image-Key scaling	**1524.4**	58.3
Norm-matched cross-image residual	1504.9	56.8
Instance-matched residual steering	1520.6	**61.4**

scaling achieves a higher MME score but a smaller CV-Bench^{2D} gain. These results support a role for image-specific residual correspondence beyond the magnitude of the added perturbation alone.

S8 Qualitative Analysis

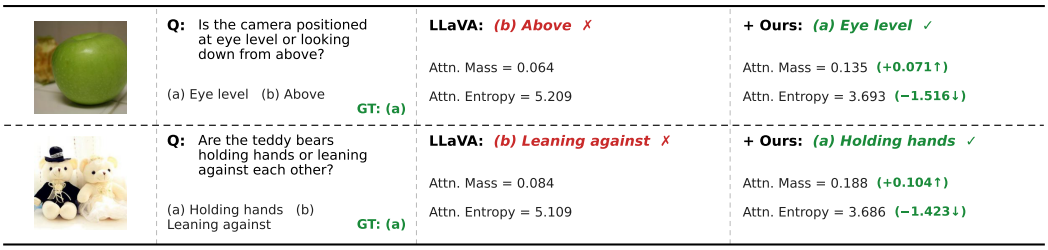

Figure S5: **Qualitative examples from MMVP on LLaVA-1.5-7B.** Each example shows the baseline prediction (LLaVA) and the prediction with `Image Key Steering` (+ Ours), with ✓/✗ indicating correctness against the ground truth (GT). We report two attention statistics averaged across all heads in the steering layers $\mathcal{L}$: Attn. Mass (total attention weight allocated to image tokens) and Attn. Entropy (per-head Shannon entropy of the T2I attention distribution over image tokens, averaged across heads). Image Key Steering increases attention mass and reduces entropy, enabling the model to attend more selectively to visual evidence and correct its prediction.

Fig. S5 shows how Image Key Steering corrects visual grounding failures on MMVP that the LLaVA-1.5-7B baseline gets wrong. Both examples require fine-grained visual discrimination: judging camera angle and recognizing a subtle spatial relationship between objects. The baseline model answers both incorrectly, with low attention mass on image tokens (0.064 and 0.084) and high entropy (5.209 and 5.109), indicating attention that is both weakly allocated to the image and diffusely spread across image tokens, rather than concentrated on relevant visual evidence. After applying Image Key Steering, the attention mass more than doubles (+0.071 and +0.104) while the entropy drops substantially (-1.516 and -1.423), and the model produces the correct answer in both cases. These shifts are consistent with the mechanism identified in §3: amplifying the instance-aware residual of image Keys strengthens the visual grounding of text queries, shifting attention from a diffuse, prior-driven pattern toward a more concentrated distribution over image tokens.

S9 Ethics Concerns

As our approach operates entirely at inference time without introducing new data or parameters, it primarily inherits the ethical profile of the foundational MLLMs. However, enhancing the fine-grained visual grounding of these models inherently amplifies their dual-use potential. While improved cross-modal alignment benefits legitimate applications, it could simultaneously facilitate malicious downstream tasks, such as invasive surveillance or unauthorized visual data mining. Furthermore, increased overall accuracy may inadvertently exacerbate automation bias, leading users to over-rely on model outputs and reducing vigilance against residual hallucinations. Mitigating these risks requires adherence to responsible deployment practices and context-specific safety evaluations.

S10 Asset Licenses and Usage Terms

We use only publicly released models and benchmarks, subject to their respective licenses and usage terms. LLaVA-1.5 (Liu et al., 2023), Qwen2-VL (Wang et al., 2024a) 7B, MMBench (Liu et al., 2024b), LLaVA-Bench (Liu et al., 2023), and CV-Bench (Tong et al., 2024a) are released under Apache-2.0. VILA (Lin et al., 2024) code is released under Apache-2.0, while its pretrained weights are released under CC BY-NC-SA 4.0. The underlying LLaMA/Vicuna-based weights used by LLaVA-1.5 and VILA are additionally governed by the applicable Meta LLaMA/Llama community license terms. MM-Vet (Yu et al., 2023) code is released under Apache-2.0, while its dataset is released under CC BY-NC 4.0. MMVP (Tong et al., 2024b) and POPE (Li et al., 2023b) are released under MIT. ScienceQA (Lu et al., 2022) is released under CC BY-NC-SA 4.0. SEED-Bench (Li et al., 2023a) is released under CC BY-NC 4.0. MSCOCO (Lin et al., 2014) annotations are released under CC BY 4.0; use of the associated images is subject to the terms of the original image sources, including the Flickr Terms of Use where applicable. MME (Fu et al., 2025) is publicly released for academic research and evaluation use.

All datasets and models used in this work are publicly available under the above licenses or usage terms. The contents of these datasets and model outputs are provided by their original creators and do not represent the views or opinions of the authors.

S11 Reproducibility

`Image Key Steering` is implemented in Pytorch (Paszke et al., 2019). Experiments are conducted on NVIDIA RTX 6000 Ada 48GB GPUs. To guarantee reproducibility, our full implementation will be available after acceptance. Implementation details are included in Appendix §S2.

S12 Social Impact and Limitations

Positive impact. This work contributes to a mechanistic understanding of text-to-image attention, a critical prerequisite for building interpretable and reliable multimodal systems. As a training-free method, it can be seamlessly integrated into existing models, offering a practical solution for environments where retraining is computationally prohibitive. The resulting improvements in visual grounding are highly beneficial for applications demanding strict factual accuracy, including assistive technologies, medical diagnostics, and document analysis. Broadly, characterizing these internal attention mechanisms facilitates the diagnosis of failure modes in deployed systems, advancing ongoing research in model transparency.

Potential negative impact. Consistent with general advancements in MLLM capabilities, our method carries potential dual-use risks, including the generation of highly grounded misinformation or the facilitation of automated surveillance. However, as discussed in §S9, this work does not introduce new training data or model weights. Instead, it applies an inference-time intervention to existing, open-source models. Because the intervention solely manipulates existing internal representations, we do not anticipate that it meaningfully shifts the overall risk profile of these systems beyond that of their foundational base models.

A promising direction is to make the intervention learnable, for example by fine-tuning α or learning per-layer residual transformations on top of the frozen backbone (Hu et al., 2022; Dettmers et al., 2023; Wu et al., 2024; Wang et al., 2024b; Shen et al., 2024; Liu et al., 2025; Khattak et al., 2023). We do not pursue this in the present work for two reasons. First, our goal is to provide a clean causal validation of the analysis in §3. A training-free intervention isolates the effect of relieving the Key-side routing–content gap from additional effects that additional training would introduce, ensuring that any observed gain can be attributed to the mechanism we identify rather than to incidental optimization effects. Second, formulating a suitable training objective is non-trivial. The desired property is to strengthen the instance-aware residual of image Keys while preserving the input-agnostic prior that governs routing structure. This property is not directly expressible as a standard supervised loss. Naive fine-tuning also risks collapsing the Key–Value asymmetry that makes the intervention meaningful. We consider the design of learnable variants as a direction for our future work.