REMO: RETHINKING MODALITY ROLES IN VISION-LANGUAGE MODELS FOR GENERAL VISUAL PERCEPTION AND REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite remarkable strides in multimodal understanding, the general visual perception of Vision-Language Models (VLMs) remains a critical weakness. Existing solutions typically rely on resource-heavy knowledge distillation or latency-intensive external visual tools, neither of which fundamentally improves the VLM's intrinsic perceptual mechanisms. In this paper, we present **ReMo**, a novel framework that **Re**thinks the complementary **Mo**dality roles of language and vision to achieve intrinsic visual enhancement in VLMs. Our core insight is that language should actively orchestrate visual perception rather than merely respond to it. ReMo first leverages the expansive reasoning capabilities of Large Language Models (LLMs) to generate linguistic priors, dynamically extracting reasoning-aware guidance from cross-modal attention to localize task-relevant visual evidence. However, since cross-modal alignment often discards fine-grained visual details, we introduce a structural complementary mechanism. By exploiting the innate relationship modeling of Vision Transformers (ViTs), ReMo actively recovers vital visual contexts that are typically overlooked by the LLM. By jointly harmonizing language-guided visual search and ViT-driven visual recovery alongside visual evidence reinforcement, our framework significantly elevates the general visual perception and reasoning capacities of VLMs.

1 INTRODUCTION

Vision-Language Models (VLMs) seamlessly integrate visual and linguistic modalities, serving as the cornerstone for artificial agents to perceive and interact with the real world. Beyond acting as key enablers for emerging paradigms, such as embodied AI (Ma et al., 2026) and world models (Xia et al., 2026), VLMs are rapidly propelling advancements across diverse downstream domains, including robotics (Shao et al., 2025), biomedicine (Sun et al., 2026), autonomous driving (Tian et al., 2025), and visual mathematical reasoning (Guan et al., 2026). This ubiquitous adoption underscores a critical imperative: enhancing the visual perception and reasoning capabilities of VLMs is essential for reliably executing increasingly complex, real-world tasks.

Notably, a pronounced capability gap persists between LLMs (Team et al., 2026; Xu et al., 2026a) and their multimodal counterparts, VLMs. Modern text-based LLMs have made remarkable strides in high-level reasoning, harnessing vast knowledge bases and sophisticated logic to independently tackle competitive programming challenges like the International Collegiate Programming Contest(ICPC) (Xu et al., 2026b). In contrast, despite inheriting these strong linguistic foundations, VLMs remain surprisingly fragile in visual perception (Tong et al., 2024). They frequently suffer from perceptual biases and hallucinated judgments, even when confronted with elementary tasks such as fine-grained recognition, geometric counting, and basic spatial comparison (Fig. 1). This stark discrepancy prompts a critical research question: *Are the internal reasoning and knowledge capabilities of LLMs fully leveraged in current VLMs?* To bridge this gap, existing methodologies for enhancing visual understanding primarily branch into two directions.

The first direction distills knowledge (Kim et al., 2026; Yin et al., 2026; Wu et al., 2026) from stronger models and scales up training data to alleviate the limitations of fine-grained visual perception. However, this paradigm incurs prohibitive data-generation costs and low training efficiency.

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.

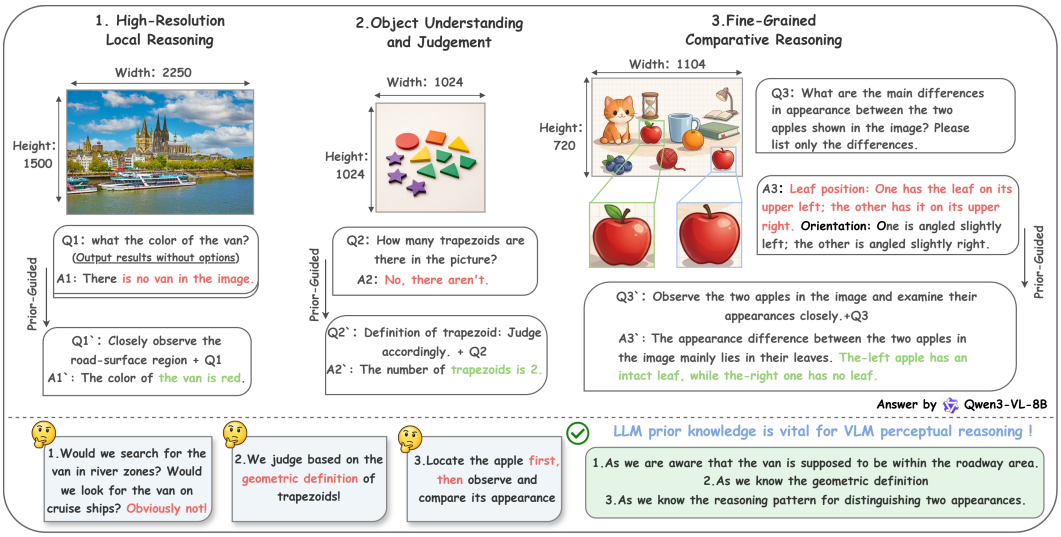


Figure 1: Contextual knowledge enhances visual perception and reasoning. Human visual understanding seamlessly integrates visual cues with prior knowledge, motivating our design to actively leverage linguistic priors for guided visual perception and reasoning in VLMs.

Furthermore, it often yields diminishing returns at scale and risks compromising model generalization. Fundamentally, it assumes that the teacher model possesses impeccable visual understanding. Yet, even state-of-the-art multimodal models like Gemini 3 Pro achieve only 69.23% accuracy on VIA-Bench (Hou et al., 2026), substantially lagging behind the 93.30% human baseline. This raises a critical question: can knowledge distillation alone catalyze a fundamental leap in visual perception? A parallel line of work adopts the "Think with Images" paradigm (Zheng et al., 2026; Zhang et al., 2025a; Huang et al., 2026), incorporating multi-step reinforcement learning with visual tools (e.g., cropping and localization). While effective on specific tasks, these external tool invocations introduce severe inference bottlenecks. More importantly, because they rely on external operations rather than modifying the model's intrinsic visual representations, they fail to fundamentally strengthen the core perception mechanisms required for diverse downstream tasks.

To overcome these bottlenecks, we propose ReMo, a knowledge-guided visual perception framework that rethinks the synergy between language and vision modalities in VLMs: *given the rich world knowledge and reasoning capabilities of LLMs, can language actively orchestrate visual perception rather than merely serving as a passive recipient of visual inputs?* Historically, leveraging LLM reasoning to facilitate visual understanding was a primary motivation behind early VLMs (Alayrac et al., 2022). Cognitively, humans rarely analyze isolated images; instead, we seamlessly integrate prior knowledge and experience to guide our gaze. Inspired by this, we first employ a lightweight language model to extract task-relevant priors and reasoning cues from the input prompt, translating them into visual search guidance. Unlike "Think with Images" which relies on external tools, we directly enhance the perception process within the model's internal architecture. Specifically, we harness the cross-modal attention in the LLM's intermediate layers to localize query-relevant regions. To ensure reliability, we introduce a denoising mechanism that suppresses attention sinks and spurious activations, yielding highly robust visual-linguistic associations.

Nevertheless, regions highlighted by cross-modal attention often lack complete visual context. While LLMs excel at semantic modeling, Vision Transformers (ViTs) are inherently superior at capturing dense structural and relational geometries within an image. We therefore project the identified key regions back into the ViT token space, exploiting the ViT's internal self-attention relationships to natively recover critical visual context overlooked during cross-modal alignment. Conceptually, this acts as the inverse of recent *VLM visual token compression* techniques: rather than aggressively pruning tokens at the cost of perceptual detail, we actively reconstruct and augment key visual evidence potentially missed by the LLM. Finally, we dynamically reweight these enriched visual regions to amplify their impact on subsequent reasoning.

Ultimately, ReMo unlocks the complementary strengths of VLM architectures: *utilizing language priors to determine "what to look at", exploiting the ViT to recover "what else needs to be seen", and applying dynamic reweighting to reinforce "how the visual evidence is leveraged"*. Consequently, we fundamentally elevate the visual perception and reasoning capabilities of VLMs without the overhead of external visual tools or the massive cost of knowledge distillation.

2 RELATED WORKS

Vision-Language Models. The rapid evolution of Large Language Models (LLMs) has catalyzed remarkable breakthroughs in Vision-Language Models (VLMs). Pioneering efforts like Flamingo (Alayrac et al., 2022) first demonstrated the efficacy of conditioning pretrained LLMs on visual inputs. Subsequently, LLaVA(Liu et al., 2023) established the prevailing paradigm: bridging a vision encoder and an LLM via a lightweight projector through multimodal instruction tuning. While proprietary titans (e.g., GPT-5(Singh et al., 2025), Claude 4, and Gemini 3 Pro) continue to push the boundaries of general multimodal intelligence, the open-source ecosystem has witnessed equally explosive growth. Rather than incremental updates, recent open-source models have introduced structural innovations: LLaVA-OneVision series (Li et al., 2024; An et al., 2026) unifies understanding across diverse visual formats; the Qwen-VL family (Qwen2.5-VL Bai et al. (2025), Qwen3-VL Bai et al. (2025)) revolutionizes fine-grained localization and long-context spatiotemporal modeling via native dynamic-resolution encoding; and architectures like InternVL3/3.5 (Zhu et al., 2025; Wang et al., 2025a) and Molmo2 (Clark et al., 2026) push the envelope through native multimodal pre-training, reinforcement learning, and optimized visual token utilization.

Visual Understanding Enhancement in VLMs. To bolster the visual perception of VLMs, recent research predominantly diverges into two trajectories. The first relies on knowledge distillation, transferring granular visual priors, reasoning traces, or complementary representations from powerful teacher models to target VLMs, as explored in LLaVA-MoD (Shu et al., 2025), MoVE-KD (Cao et al., 2025), and others (Kim et al., 2026; Yin et al., 2026; Wu et al., 2026; Chen et al., 2026; Wei et al., 2026; Lee et al., 2026; Yoon et al., 2026). The second trajectory adopts a "think-with-images" paradigm (Huang et al., 2026; Zheng et al., 2026; Zhang et al., 2025a; Su et al., 2026; Lai et al., 2026; Li & Peng, 2026), empowering VLMs to iteratively invoke external visual tools, such as cropping, zooming, and regional grounding, to gather supplementary evidence during inference. While these paradigms have significantly elevated VLM capabilities, their dependence on expensive large-scale distillation pipelines or latency-intensive external tool invocations highlights a critical need for more intrinsic and efficient perception enhancement strategies, a fundamental gap that our framework aims to bridge.

3 METHODOLOGY

Overview of ReMo. As illustrated in Fig. 2, ReMo operates in three stages: first, it leverages task-relevant knowledge and reasoning cues to localize query-specific visual evidence; next, it exploits the visual relationships encoded within the ViT to recover potentially missing contextual evidence; finally, it reweights the resulting visual tokens to optimize subsequent reasoning. In essence, ReMo mirrors an intuitive perception pipeline: *what to look at (Sec. 3.1), what else needs to be seen (Sec. 3.2), and how to leverage the visual evidence (Sec. 3.3).*

3.1 LANGUAGE-GUIDED VISUAL LOCALIZATION: WHAT TO LOOK AT

Knowledge Extraction. Inspired by human visual cognition, where individuals rely on goals, prior knowledge, and experiential expectations to rapidly localize task-relevant regions, we posit that the language component within a VLM should function beyond mere text generation. Its embedded world knowledge and reasoning capabilities serve a dual purpose: providing top-down cognitive guidance to proactively steer visual attention, and supplying essential semantic context (e.g., category attributes, definitional rules) to enrich visual comprehension. To realize this, we repurpose the VLM's inherent linguistic priors to actively drive the search for visual evidence. Concretely, we extract task-specific knowledge priors and key semantic anchors from the input prompt, subsequently leveraging cross-modal attention mechanisms to precisely localize the corresponding visual regions.

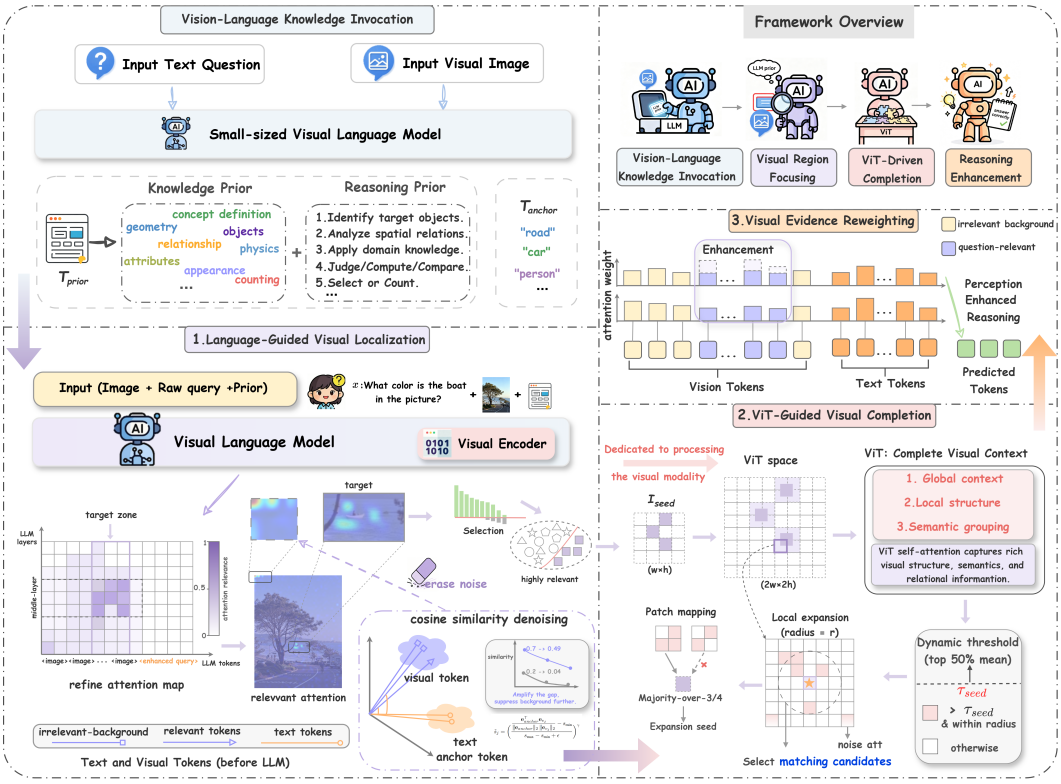


Figure 2: **Overview of ReMo Framework.** ReMo enhances VLM visual understanding through a three-stage pipeline: (1) language-guided visual localization to pinpoint task-relevant regions, (2) ViT-guided structural completion to recover missing contextual details, and (3) attention logit reweighting to optimize downstream reasoning.

Formally, given an input query Q, an image I, and a predefined vision-language prior prompt P_{vlk}, we deploy a lightweight 4B VLM (Bai et al., 2025) to extract task-pertinent knowledge priors and salient semantic anchors:

$$[T_{\text{prior}},\, T_{\text{anchor}}] = VLM(Q,\, I,\, P_{\text{vlk}}), \quad Q^{+} = [Q;\, T_{\text{prior}}]\,, \tag{1}$$

where T_{prior} denotes the relevant domain knowledge and reasoning cues required for the task, and T_{anchor} represents the salient semantic anchors extracted from the prompt. Subsequently, this enriched textual context is integrated alongside the original question into the VLM, enabling these generated priors and anchors to actively drive subsequent cross-modal reasoning. Crucially, the objective of this design is not to arbitrarily lengthen the textual prompt, but rather to recast abstract task instructions into *semantic constraints directly actionable for visual localization*, thereby establishing explicit search targets for visual evidence extraction.

Relevant Region Localization. To capitalize on the robust cross-modal correspondences captured within the intermediate layers of the VLM, we measure the attention from the augmented textual sequence to the visual tokens, aggregating these weights across designated intermediate layers to derive reliable relevance scores. Note that the semantic anchors T_{anchor} are excluded from this initial localization step and reserved for subsequent cosine-based denoising.

Formally, let $T = \{t_1, \ldots, t_M\}$ denote the textual tokens of the augmented input Q^{+}, and $V = \{v_1, \ldots, v_N\}$ represent the visual tokens. Within the l-th intermediate layer, let $A_{m,i}^{(l,h)}$ denote the attention weight assigned by textual token t_m to visual token v_i under the h-th attention head, where textual tokens serve as Queries and visual tokens act as Keys. We aggregate these attention weights across heads, textual tokens, and the designated set of intermediate layers to yield an initial cross-

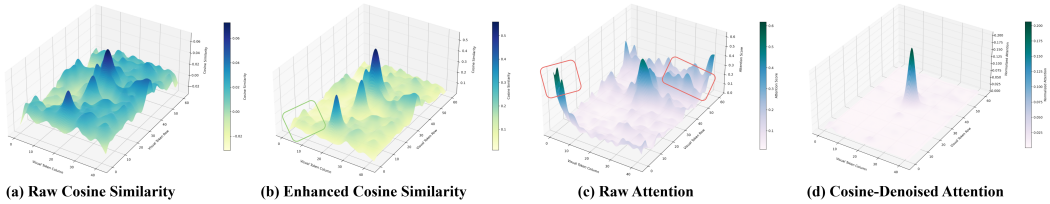


Figure 3: Visualization of cosine-denoised attention for more effective visual focus. Note: For ease of visualization, (c) and (d) show normalized and Gaussian-smoothed relative attention values.

modal relevance score a_i for each visual token:

$$a_i = \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} \mathrm{Agg}_m \left(\frac{1}{H} \sum_{h=1}^{H} A_{m,i}^{(l,h)} \right), \tag{2}$$

where L denotes the set of selected intermediate layers, H is the number of attention heads, and Aggm$(\cdot)$ signifies the aggregation operator over textual tokens. The score a_i quantifies the cross-modal alignment between visual token v_i and task-relevant textual context. The rationale for selecting intermediate layers is shown in **Appendix** Fig. 7, which compares layer-wise attention patterns.

Region Denoising. Although intermediate-layer cross-modal attention highlights coarsely relevant regions, directly using raw attention weights for visual localization remains suboptimal. First, cross-modal attention maps inherently suffer from background noise and attention-sink artifacts (Qiu et al., 2026), which erroneously assign high response values to uninformative visual tokens. Second, as attention distributions directly govern how information is aggregated during generation, refining these scores is vital for establishing a reliable foundation to select visual candidates. To this end, we introduce a lightweight semantic denoising mechanism designed to suppress spurious responses while amplifying task-critical visual evidence.

Rather than altering the internal attention architecture, we leverage explicit semantic alignment between input visual tokens and the semantic anchors T_{anchor}. Let $\mathbf{e}_{\text{anchor}}$ denote the embedding of the semantic anchors, and $\mathbf{e}_{v_j}$ represent the embedding of the j-th visual token prior to LLM injection. We first compute their cosine similarity: $s_j = (\mathbf{e}_{\text{anchor}}^{\top} \mathbf{e}_{v_j}) / (\|\mathbf{e}_{\text{anchor}}\|_2 \|\mathbf{e}_{v_j}\|_2)$. To compensate for the limited dynamic range often exhibited by raw cosine similarities, we normalize s_j bounded by the token-wide extremums and apply a power-law transformation for contrast enhancement:

$$\tilde{s}_j = \left(\frac{s_j - s_{\min}}{s_{\max} - s_{\min} + \epsilon} \right)^{\gamma}, \tag{3}$$

where $s_{\min}$ and $s_{\max}$ denote the minimum and maximum cosine similarities across all visual tokens, ϵ is a small constant for numerical stability, and γ governs the non-linear contrast stretching.

The resulting similarity score $\tilde{s}_j$ is then employed to reweight the multi-layer attention map, effectively suppressing background noise while concentrating attention on question-relevant regions:

$$\tilde{a}_j = a_j \times (1 + \alpha \tilde{s}_j), \tag{4}$$

where α represents the intensity scaling coefficient. This denoising effect is visualized in Fig. 3. Finally, based on the refined relevance scores $\tilde{a}_j$, we isolate the top-performing visual tokens to form the initial visual seeds:

$$\mathcal{I}_{\text{seed}} = \left\{ j \,\middle|\, \tilde{a}_j > \frac{1}{\lfloor K_{top} \rfloor} \sum_{t=1}^{\lfloor K_{top} \rfloor} \tilde{a}_{(t)} \right\}, \tag{5}$$

where $\tilde{a}_{(t)}$ denotes the t-th largest relevance score, and K_{top} denotes the number of top-ranked tokens used to estimate the relevance threshold. The specific setting is detailed in the Appendix.

3.2 ViT-Guided Visual Completion: What Else Needs to Be Seen

Although the visual seeds $\mathcal{I}_{\text{seed}}$ pinpoint task-relevant regions, they often fail to capture the complete visual evidence required for rigorous reasoning. This limitation stems from a fundamental architec-

tural trade-off: while the LLM excels at high-level cross-modal semantic alignment over downsampled, compressed visual tokens, its attention maps frequently discard localized spatial context and subtle visual structures. In contrast, ViTs natively preserve rich intra-modal spatial and structural dependencies. To bridge LLM attention's semantic abstraction gap, we project localized seeds into the ViT space, leveraging intrinsic visual affinities to recover overlooked contextual evidence.

Formally, given the LLM-space visual seeds $\mathcal{I}_{\text{seed}}$, we first map them to their corresponding ViT patch locations via the model's structural mapping rule: $\mathcal{S} = \mathcal{M}(\mathcal{I}_{\text{seed}})$. For each ViT seed $s \in \mathcal{S}$, we execute an attention-guided local diffusion process to identify visually and spatially correlated tokens. Let $B \in \mathbb{R}^{N_{\text{vit}} \times N_{\text{vit}}}$ denote the aggregated ViT self-attention matrix. We enforce an adaptive affinity threshold combined with a spatial proximity constraint to isolate valid contextual candidates:

$$\mathcal{C}_s = \left\{ j \;\middle|\; \|p_j - p_s\|_2 \le r \;\wedge\; B_{j,s} > \frac{1}{\lceil N_{\text{vit}}/2 \rceil} \sum_{k=1}^{\lceil N_{\text{vit}}/2 \rceil} \left(B_{:,s}\right)_{(k)} \right\}, \tag{6}$$

where $B_{j,s}$ measures the visual affinity from seed token s to candidate ViT token j, p_j and p_s represent their 2D spatial coordinates, r defines the maximum local diffusion radius, and $\left(B_{:,s}\right)_{(k)}$ denotes the k-th largest entry in column s of the ViT attention matrix B. The threshold is dynamically calibrated per seed by averaging the upper quartile of its affinity distribution across all ViT tokens. The logical conjunction $\wedge$ ensures that candidate tokens are both spatially adjacent to the seed and strongly correlated in the visual feature space, effectively preventing unconstrained global token expansion and noise propagation.

The recovered ViT context tokens are subsequently re-projected to the LLM visual token space. Given the standard 2×2 patch merging scheme where each compressed LLM token aggregates a local grid of four ViT tokens, we adopt a majority-voting rule to promote candidate LLM tokens:

$$\mathcal{I}_{\text{comp}} = \left\{ i \;\middle|\; \frac{\left|\mathcal{M}^{-1}(i) \cap \bigcup_{s \in \mathcal{S}} \mathcal{C}_s\right|}{|\mathcal{M}^{-1}(i)|} \ge \eta \right\}, \qquad \eta = 0.75, \tag{7}$$

where $\mathcal{M}^{-1}(i)$ is the set of four ViT tokens corresponding to the i-th LLM visual token. Setting $\eta = 0.75$ mandates that at least three of the four constituent ViT tokens are activated before the corresponding LLM token is included. Finally, the completed visual evidence set is constructed via:

$$\mathcal{I}_{\text{evidence}} = \mathcal{I}_{\text{seed}} \cup \mathcal{I}_{\text{comp}}. \tag{8}$$

By exploiting the dense structural relationships preserved within the ViT backbone, this completion process organically expands the initial semantic seeds, retrieving vital visual details missed by LLM cross-modal attention without incurring unconstrained token redundancy.

3.3 Visual Evidence Reweighting: How to Leverage the Evidence

While the preceding stages extract a consolidated set of visual evidence $\mathcal{I}_{\text{evidence}}$ characterized by both semantic relevance and structural completeness, effectively incorporating this spatial prior into the downstream reasoning phase remains non-trivial. Rather than resorting to destructive hard cropping or computationally expensive multi-stage re-encoding, we propose a lightweight, non-invasive attention modulation strategy. This design addresses two key limitations of hard cropping: first, explicit cropping severs spatial dependencies between local details and global context; second, external cropping pipelines disrupt end-to-end generation and incur heavy inference latency. In contrast, attention biasing non-invasively prioritizes visual evidence while fully preserving global context and architectural simplicity.

Formally, given the aggregated evidence set $\mathcal{I}_{\text{evidence}}$, we modulate the standard cross-modal attention calculation by introducing an additive positive bias to the attention logits prior to normalization:

$$\tilde{A}_{ij} = \text{Softmax}\left(\frac{\mathbf{q}_i^\top \mathbf{k}_j}{\sqrt{d}} + \lambda \mathbb{I}\left[j \in \mathcal{I}_{\text{evidence}}\right] \right), \tag{9}$$

where $\mathbf{q}_i$ and $\mathbf{k}_j$ represent the query and key representations of the i-th text token and j-th visual token, respectively, d denotes the feature dimension, and $\mathbb{I}[\cdot]$ is the indicator function. The hyperparameter $\lambda > 0$ controls the amplification intensity applied to the prioritized evidence. Under this formulation, visual tokens belonging to $\mathcal{I}_{\text{evidence}}$ receive a constant logit boost, encouraging the model to allocate higher attention weights to task-critical regions during text generation without ignoring background context.

Table 1: **Results on High-Resolution Benchmarks.** Comparisons on Vstar Bench and HR-Bench.

	Vstar-Bench			HR-Bench4K			HR-Bench8K		
	Average	*Attribute*	*Spatial*	*Average*	*Single*	*Cross*	*Average*	*Single*	*Cross*
Proprietary and Open-Source Models									
GPT-4o	67.5	72.2	60.5	65.0	66.8	63.3	59.6	60.8	58.5
Gemini-2.5-Pro	79.1	86.8	68.4	83.9	85.5	82.3	81.5	83.0	80.0
Gemini-3-Flash	72.3	64.4	77.3	77.5	81.5	74.0	73.7	75.8	71.8
Qwen2.5-VL-72B	84.8	90.8	80.9	79.4	88.8	70.0	76.3	84.3	68.3
Qwen3-VL-235B-A22B	80.6	82.9	79.1	83.0	89.0	77.0	80.4	83.0	77.3
Kimi-K2.5(1T)	85.9	87.0	84.2	81.7	87.3	76.0	75.1	82.8	67.5
Our Models									
Qwen3-VL-4B	82.2	86.9	75.0	76.1	87.8	64.5	72.5	84.5	60.5
ReMo-4B	86.4+4.2	89.5+3.4	81.6+6.6	82.9+6.8	90.0+2.2	75.8+11.3	78.4+5.9	87.5+3.0	69.3+8.8
Qwen3-VL-8B	83.2	86.1	78.9	76.3	92.0	60.5	71.1	83.0	59.3
ReMo-8B	89.5+6.3	91.3+5.2	86.8+7.9	83.2+6.9	92.8+0.8	73.5+13	79.2+8.1	87.8+4.8	70.5+11.2
Kimi-VL-16B	62.3	59.1	67.1	58.0	63.5	52.5	55.1	60.3	50.0
ReMo-16B	72.8+10.5	70.4+11.3	76.3+9.2	66.5+8.5	71.3+7.5	61.8+9.3	63.8+8.7	68.5+8.2	59.0+9.0

4 EXPERIMENTS

Baselines and Benchmarks. To comprehensively evaluate the effectiveness of our method, we conduct experiments on **Qwen3-VL-4B**, **Qwen3-VL-8B** (Bai et al., 2025) and **Kimi-VL-16B-A3B** (Team et al., 2025), covering both **Dense and Mixture-of-Experts (MoE)** architectures. We further evaluate our method on a diverse set of benchmarks. (1) **Vstar** (Wu & Xie, 2024) evaluates fine-grained visual perception and reasoning in challenging visual scenes. (2) **HR-Bench-4K** (Wang et al., 2025b) evaluates visual understanding under high-resolution inputs. (3) **HR-Bench-8K** Wang et al. (2025b) further assesses fine-grained visual understanding with higher-resolution inputs. (4) **MMStar** (Chen et al., 2024) provides a comprehensive evaluation of general multimodal understanding across diverse visual tasks. (5) **MME-RealWorld-Lite** (Zhang et al., 2025b) evaluates visual perception and reasoning in real-world scenarios. (6) **HallusionBench** (Guan et al., 2024) measures the model's susceptibility to visual and language hallucinations. (7) **POPE** (Li et al., 2023) evaluates object hallucination and the reliability of visual grounding.

4.1 RESULTS

High-Resolution Benchmarks. Table 1 presents the performance of ReMo on three challenging high-resolution visual understanding benchmarks: VSTAR, HR-Bench-4K, and HR-Bench-8K. Overall, ReMo consistently elevates accuracy across diverse VLM backbones, model scales, and network topologies, demonstrating that its benefits are architecturally agnostic and easily adaptable to mainstream VLMs. Specifically, on Qwen3-VL-4B, ReMo yields notable performance boosts of 4.2%, 6.8%, and 5.9% on VSTAR, HR-Bench-4K, and HR-Bench-8K, respectively. When scaled up to the larger Qwen3-VL-8B backbone, these gains further expand to 6.3%, 6.9%, and 8.1%. Furthermore, on the MoE-based Kimi-VL-16B-A3B, ReMo delivers remarkable accuracy leaps of 10.5%, 8.5%, and 8.7% across the three benchmarks. These robust improvements across both dense and sparse MoE models highlight ReMo's strong architectural generality, confirming its ability to effectively unlock the complementary cross-modal representations embedded in diverse VLMs.

Unlike "Think with Images" paradigms relying on explicit cropping or re-encoding, ReMo selects and utilizes visual evidence natively within internal attention space. Our findings confirm that pairing language-guided localization with ViT-guided completion mitigates visual redundancy while recovering fine-grained details for high-resolution VLM reasoning.

General Perception and Reasoning Benchmarks. To evaluate ReMo across broad visual perception and reasoning tasks, we extend our evaluation to MMStar and MME-RealWorld-Lite. These benchmarks encompass a wide spectrum of visual competencies, including relational understanding, attribute recognition, commonsense reasoning, and open-world scene understanding. As shown in Table 2, ReMo yields consistent performance gains across all VLM backbones. On MMStar, ReMo boosts Qwen3-VL-4B, Qwen3-VL-8B, and Kimi-VL-16B-A3B by 1.87%, 2.54%, and 3.27%, re-

Table 2: **Results on General Perception and Reasoning Benchmarks.** The left shows results on the **MMStar** benchmark, and the right shows results on the **MME-RealWorld-Lite** benchmark.

			Perception					Reasoning			
Models	**MMStar**	**Overall**	OCR	Remote Sen.	Diag. & Tab.	Monitoring	Auto. Driv.	OCR	Diag. & Tab.	Monitoring	Auto. Driv.
LLaVA-OV-7B	61.7	43.7	80.0	40.0	56.0	31.7	39.4	65.0	33.0	38.0	32.0
LLaVA-OV-72B	66.7	48.7	79.2	50.7	67.0	37.9	40.0	76.0	41.0	38.7	39.3
InternVL3-38B	71.5	51.0	85.6	56.0	71.0	42.6	40.0	77.0	45.0	47.3	35.0
InternVL3-78B	72.5	52.3	87.6	54.7	77.0	42.6	36.6	76.0	56.0	46.0	40.3
Qwen2.5-VL-32B	69.5	45.6	87.2	40.7	83.0	29.5	40.7	74.0	60.0	27.3	29.5
Qwen2.5-VL-72B	70.8	43.7	90.8	34.0	87.0	27.9	30.6	74.0	61.0	26.7	25.5
Qwen3-VL-32B	77.7	52.0	91.6	47.3	96.0	36.1	42.9	76.0	77.0	42.7	30.0
Our Models											
Qwen3-VL-4B	60.93	47.1	90.8	44.7	87.0	34.8	32.6	72.0	64.0	43.4	24.3
ReMo-4B	62.80+1.87	55.1+8.0	94.8	54.0	91.0	46.7	42.3	77.0	70.0	54.0	30.8
Qwen3-VL-8B	64.33	48.6	92.8	57.3	87.0	36.4	31.4	73.0	70.0	39.3	25.0
ReMo-8B	66.87+2.54	57.9+9.3	96.0	68.7	92.0	49.8	42.0	79.0	78.0	52.7	33.5
Kimi-VL-16B	60.00	38.3	75.2	39.3	47.0	21.6	31.1	68.0	28.0	28.7	31.0
ReMo-16B	63.27+3.27	45.5+7.2	80.4	47.3	52.0	29.2	37.7	77.0	39.0	40.0	37.3

Table 3: **Results on multimodal hallucination-understanding benchmarks.** The left part presents results on the **HallusionBench** benchmark. The right part shows the main results on the **POPE** benchmark, where Accuracy (%), Precision (%) are reported for its three splits.

		Random		Popular		Adversarial	
Models	**HallusionBench**	Acc	Prec	Acc	Prec	Acc	Prec
Qwen3-VL-4B	68.5	91.40	98.29	89.70	94.54	88.43	91.93
ReMo-4B	69.6	92.27	98.40	91.13	95.77	89.10	92.53
Qwen3-VL-8B	74.2	91.93	97.72	89.07	91.74	86.90	87.68
ReMo-8B	75.1	92.73	98.34	90.10	92.51	87.83	88.74
Kimi-VL-16B	64.6	89.10	98.92	88.03	96.34	87.27	94.58
ReMo-16B	66.2	90.43	99.11	89.60	97.12	88.43	95.57

spectively. On MME-RealWorld-Lite, these gains reach 8.0%, 9.3%, and 7.2%. These uniform improvements across diverse model families confirm that ReMo is architecturally agnostic and effectively leverages existing cross-modal representations in VLMs.

Notably, the markedly higher gains on MME-RealWorld-Lite highlight ReMo's efficacy in complex, open-ended real-world scenarios that demand joint reasoning over global scene semantics and local details. By synergizing language priors for semantic target localization with ViT structural geometries for context completion, ReMo allows models to systematically organize and exploit visual evidence. These results demonstrate that combining language guidance with intra-modal visual structure provides a highly generalizable foundation for multimodal perception and reasoning.

Multimodal Hallucination Benchmarks. We further evaluate ReMo on HallusionBench and POPE to examine its capacity to mitigate multimodal hallucinations. As shown in Table 3, ReMo consistently enhances performance across all evaluated VLMs on HallusionBench, yielding gains of 1.1%, 0.9%, and 1.6% for Qwen3-VL-4B, Qwen3-VL-8B, and Kimi-VL-16B-A3B, respectively. While these improvements are relatively conservative compared to those on high-resolution benchmarks, primarily because HallusionBench features a high proportion of structured tabular images where spatial evidence completion offers less utility, the steady gains across all backbones nevertheless confirm ReMo's robustness in improving visual reliability. ReMo demonstrates consistent accuracy improvements on POPE, confirming that its benefits generalize effectively to object-level hallucination reduction. These findings underscore that synergizing language-guided evidence localization with ViT-based context recovery provides a reliable defense against visual hallucinations.

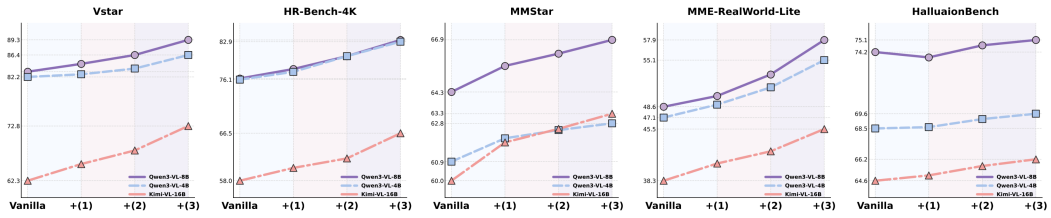


Figure 4: Progressive ablation study of Vanilla Model + (1) knowledge priors + (2) cross-modal localization with denoising (reweighting enabled) + (3) ViT-guided completion (reweighting enabled).

4.2 Ablation Study

To dissect the contribution of each stage, we conduct a progressive ablation study following ReMo's execution pipeline (Fig. 4). Because localization (Sec. 3.1) and completion (Sec. 3.2) merely select tokens, attention reweighting (Sec. 3.3) is required to drive generation. Consequently, **reweighting is enabled across all ablated variants rather than evaluated as a standalone component; its hyperparameter sensitivity is detailed in the Appendix Table 4**. With attention reweighting enabled, we incrementally introduce: (1) knowledge priors, (2) cross-modal localization with denoising, and (3) ViT-guided completion. The results confirm a clear trajectory of progressive refinement. Knowledge priors furnish task guidance and semantic definitions; cross-modal localization with cosine denoising filters out background noise to isolate relevant regions; and ViT-guided completion leverages spatial structures to recover overlooked visual details. Overall, ReMo succeeds through vision-language synergy: language priors dictate *what to look at*, localization with denoising isolates visual evidence, ViT completion recovers *what else needs to be seen*, and attention logit reweighting optimizes downstream reasoning to reinforce *how the visual evidence is leveraged*. See **Appendix Fig. 5** for visualization of ReMo visual focusing and ViT-guided completion.

4.3 Further Analysis

ReMo vs. Thinking Variant. We further contrast ReMo with the Thinking Variant to clarify their perceptual mechanisms. Instead of lengthening reasoning chains, ReMo leverages linguistic priors to actively guide visual evidence selection. On perception-heavy tasks, Qwen3-VL-8B-Instruct scores 83.2 on VStar and 76.3 on HR-Bench, exceeding the Thinking variant (77.5 and 72.4). ReMo, which relies solely on linguistic priors, further elevates performance to 84.7 and 78.0 without dedicated prompt optimization. This reveals that for visual perception, scaling reasoning compute is less effective than active evidence acquisition. Given that Thinking variants dominate complex reasoning tasks (e.g., math), ReMo and chain-of-thought (CoT) are fundamentally orthogonal: ReMo ensures reliable visual evidence for downstream CoT reasoning, offering a viable path for unified multi-modal foundation models.

5 Conclusion

In this work, we introduced ReMo, a novel framework that rethinks the complementary roles of language and vision modalities to achieve intrinsic visual enhancement in VLMs. ReMo unifies task-guided target localization and visual evidence recovery into a cohesive and lightweight pipeline: it harnesses linguistic reasoning cues to pinpoint semantically salient regions, applies cosine-calibrated denoising to eliminate background artifacts, exploits ViT structural dependencies to restore missing spatial details, and softly modulates attention logits to prioritize key visual evidence during generation. Extensive experiments on fine-grained high-resolution benchmarks, open-world reasoning tasks, and multimodal hallucination suites demonstrate that ReMo consistently enhances perceptual accuracy and structural fidelity without sacrificing architectural simplicity. Ultimately, this work provides a versatile perspective on seamlessly coordinating top-down language guidance with bottom-up visual structure for advanced multimodal perception and reasoning.

AI USE STATEMENT

During the preparation of this manuscript, the authors utilized generative AI tools solely for [language polishing of the manuscript]. Generative AI was not used for [research ideation, algorithm design, data collection, experimental analysis, or figure generation]. The authors have thoroughly reviewed and edited all content and accept full responsibility for the integrity, claims, and artifacts presented in this work.

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. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, et al. Llava-onevision-2: Towards next-generation perceptual intelligence. *arXiv preprint arXiv:2605.25979*, 2026.

Jyoti Aneja, Michael Harrison, Neel Joshi, Tyler LaBonte, John Langford, and Eduardo Salinas. Phi-4-reasoning-vision-15b technical report. *arXiv preprint arXiv:2603.03975*, 2026.

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.

Jiajun Cao, Yuan Zhang, Tao Huang, Ming Lu, Qizhe Zhang, Ruichuan An, Ningning Ma, and Shanghang Zhang. Move-kd: Knowledge distillation for vlms with mixture of visual encoders. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19846–19856. IEEE, 2025.

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, 2024.

Lin Chen, Xiaoke Zhao, Kun Ding, Weiwei Feng, Changtao Miao, Zili Wang, Wenxuan Guo, Ying Wang, Kaiyuan Zheng, Bo Zhang, et al. Beyond next-token alignment: Distilling multimodal large language models via token interactions. *arXiv preprint arXiv:2602.09483*, 2026.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Rohun Tripathi, Sangho Lee, Mohammadreza Salehi, Jason Ren, Chris Dongjoo Kim, Yinuo Yang, et al. Molmo2: Open weights and data for vision-language models with video understanding and grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28652–28668, 2026.

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.

Tongkun Guan, Zhibo Yang, Jianqiang Wan, Mingkun Yang, Zhentao Guo, Zijian Hu, Ruilin Luo, Ruizhe Chen, Songtao Jiang, Peng Wang, et al. Codepercept: Code-grounded visual stem perception for mllms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33542–33552, 2026.

Wenjin Hou, Wei Liu, Han Hu, Xiaoxiao Sun, Serena Yeung-Levy, and Hehe Fan. Seeing is believing? a benchmark for multimodal large language models on visual illusions and anomalies. *arXiv preprint arXiv:2602.01816*, 2026.

Wenxuan Huang, Bohan Jia, Shaosheng Cao, Zheyu Ye, Zhe Xu, Yao Hu, Shaohui Lin, et al. Vision-r1: Incentivizing reasoning capability in multimodal large language models. In *International Conference on Learning Representations*, volume 2026, pp. 63794–63812, 2026.

Jiwan Kim, Kibum Kim, Sangwoo Seo, and Chanyoung Park. Compodistill: Attention distillation for compositional reasoning in multimodal llms. In *International Conference on Learning Representations*, volume 2026, pp. 71814–71844, 2026.

Xin Lai, Junyi Li, Wei Li, Tao Liu, Tianjian Li, and Hengshuang Zhao. Mini-o3: Scaling up reasoning patterns and interaction turns for visual search. In *International Conference on Learning Representations*, volume 2026, pp. 76722–76746, 2026.

Byung-Kwan Lee, Yu-Chiang Frank Wang, and Ryo Hachiuma. Masking teacher and reinforcing student for distilling vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10126–10141, 2026.

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.

Geng Li and Yuxin Peng. Bvs: Bayesian visual search with multimodal large language model for fine-grained perception. *arXiv preprint arXiv:2607.03184*, 2026.

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.

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

Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. A survey on vision–language–action models for embodied ai. *IEEE Transactions on Neural Networks and Learning Systems*, 2026.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, et al. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. *Advances in Neural Information Processing Systems*, 38:100092–100118, 2026.

Rui Shao, Wei Li, Lingsen Zhang, Renshan Zhang, Zhiyang Liu, Ran Chen, and Liqiang Nie. Large vlm-based vision-language-action models for robotic manipulation: A survey. *arXiv preprint arXiv:2508.13073*, 2025.

Fangxun Shu, Yue Liao, Lei Zhang, Le Zhuo, Chenning Xu, Guanghao Zhang, Haonan Shi, Weilong Dai, Zhelun Yu, Wanggui He, et al. Llava-mod: Making llava tiny via moe-knowledge distillation. In *International Conference on Learning Representations*, volume 2025, pp. 9386–9404, 2025.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel space reasoning via curiosity-driven reinforcement learning. *Advances in Neural Information Processing Systems*, 38:8222–8251, 2026.

Yanpeng Sun, Shan Zhang, Wei Tang, Aotian Chen, Piotr Koniusz, Kai Zou, Yuan Xue, and Anton Hengel. Math blind: Failures in diagram understanding undermine reasoning in mllms. In *International Conference on Learning Representations*, volume 2026, pp. 124385–124422, 2026.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y Charles, et al. Kimi k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

Zhen Tian, Zhihao Lin, Dezong Zhao, Wenjing Zhao, David Flynn, Shuja Ansari, and Chongfeng Wei. Evaluating scenario-based decision-making for interactive autonomous driving using rational criteria: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 2025.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9568–9578. IEEE, 2024.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025a.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7907–7915, 2025b.

Lai Wei, Liangbo He, Jun Lan, Lingzhong Dong, Yutong Cai, Siyuan Li, Huijia Zhu, Weiqiang Wang, Linghe Kong, Yue Wang, et al. Zooming without zooming: Region-to-image distillation for fine-grained multimodal perception. *arXiv preprint arXiv:2602.11858*, 2026.

Linquan Wu, Tianxiang Jiang, Yifei Dong, Haoyu Yang, Fengji Zhang, Shichaang Meng, Ai Xuan, Linqi Song, and Jacky Keung. Lavit: Aligning latent visual thoughts for multi-modal reasoning. *arXiv preprint arXiv:2601.10129*, 2026.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13084–13094. IEEE, 2024.

Tianze Xia, Yongkang Li, Lijun Zhou, Jingfeng Yao, Kaixin Xiong, Haiyang Sun, Bing Wang, Kun Ma, Guang Chen, Hangjun Ye, et al. Drivelaw: Unifying planning and video generation in a latent driving world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 39701–39712, 2026.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026a.

Shiyi Xu, Hu Yiwen, Yingqian Min, Zhipeng Chen, Xin Zhao, and Ji-Rong Wen. Icpc-eval: Probing the frontiers of llm reasoning with competitive programming contests. *Advances in Neural Information Processing Systems*, 38, 2026b.

Qixiang Yin, Huanjin Yao, Yuchen Cai, Jianghao Chen, Ziyi Wang, Min Yang, Fei Su, and Zhicheng Zhao. H-opd: Confidence aware heterogeneous multi-teacher multimodal on-policy distillation. *arXiv preprint arXiv:2607.02592*, 2026.

Hee Suk Yoon, Eunseop Yoon, Jaehyun Jang, SooHwan Eom, Ji Woo Hong, Mark Hasegawa-Johnson, Qi Dai, Chong Luo, and Chang D Yoo. Decomposed on-policy distillation for vision-language reasoning: Steering gradients for visual grounding. *arXiv preprint arXiv:2606.00564*, 2026.

Hanxun Yu, Xuan Qu, Yuxin Wang, Jianke Zhu, and Lei Ke. Unlocking dense metric depth estimation in vlms. *arXiv preprint arXiv:2605.15876*, 2026.

Xintong Zhang, Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaowen Zhang, Yang Liu, Tao Yuan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, et al. Adaptive chain-of-focus reasoning via dynamic visual search and zooming for efficient vlms. *arXiv preprint arXiv:2505.15436*, 2025a.

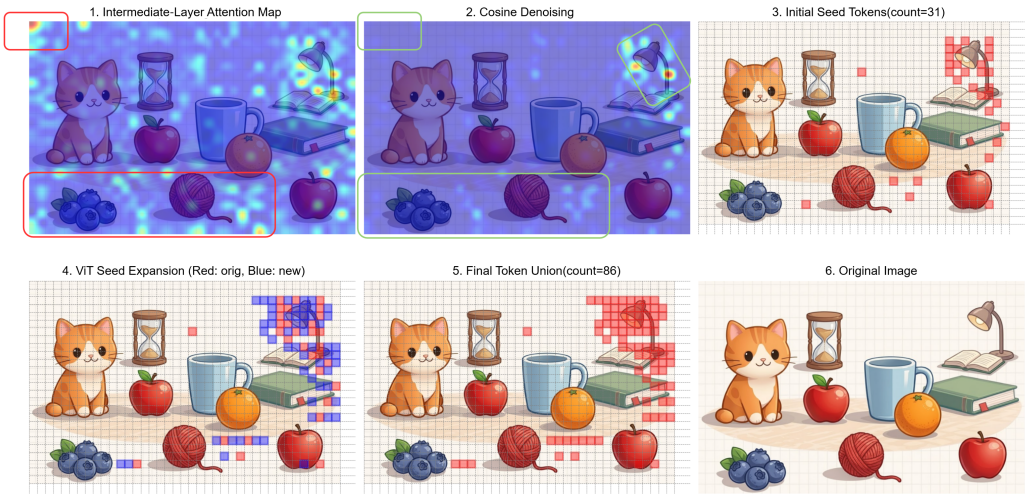
YiFan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? In *International Conference on Learning Representations*, volume 2025, pp. 89655–89701, 2025b.

Ziwei Zheng, Minghao Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, and Chao Shen. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 126775–126798, 2026.

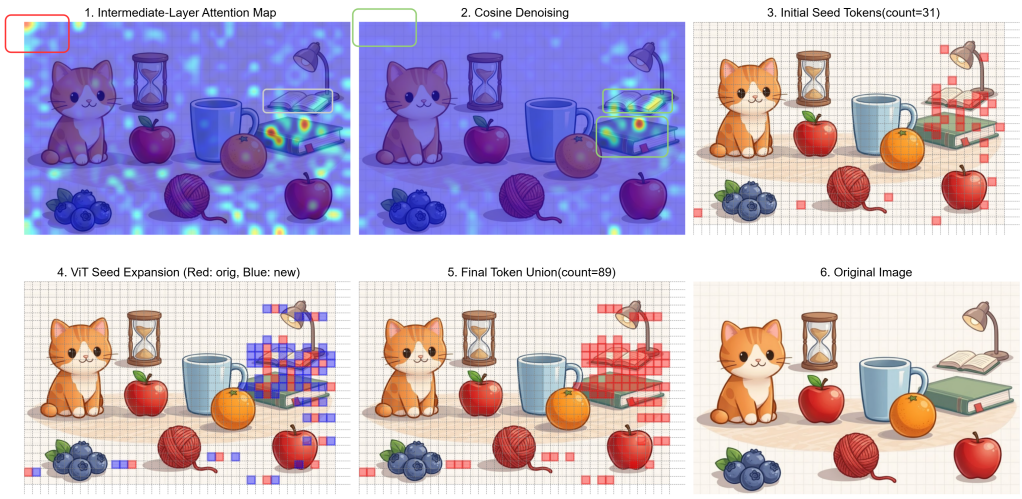
Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A APPENDIX

Visualization of ReMo Visual Focusing and ViT-based Token Completion



(a) Visualization of the Token Selection Pipeline for the Lamp



(b) Visualization of the Token Selection Pipeline for the Book

Figure 5: Visualization of the token selection process in ReMo on two representative examples, *Lamp* and *Book*. The middle-layer attention map provides a coarse localization signal, which is further refined by cosine-based denoising. The resulting initial seed tokens are then expanded using ViT-based spatial completion, where red and blue tokens denote the initial seeds and newly added tokens, respectively. The final union provides a more spatially complete set of task-relevant visual tokens while remaining concentrated around the localized evidence regions.

To provide a more intuitive view of how ReMo selects and completes visual evidence, we visualize the token selection process on two representative examples, Lamp and Book. As shown in Fig. 5, the middle-layer attention map first provides a coarse indication of the potentially relevant regions. However, the raw attention distribution remains spatially diffuse and contains noticeable responses outside the task-relevant regions. After cosine-based denoising, these noisy responses are substantially suppressed, yielding a more concentrated localization signal around the semantically relevant areas. Based on the denoised attention, ReMo selects an initial set of 31 seed tokens, which captures

Table 4: Parameter sensitivity results of the hyperparameter λ in the reweight attention mechanism in Eq. (9).

	Vstar			HR-Bench4K			HR-Bench8K		
Models	$\lambda = 1$	$\lambda = 1.5$	$\lambda = 2.5$	$\lambda = 1$	$\lambda = 1.5$	$\lambda = 2.5$	$\lambda = 1$	$\lambda = 1.5$	$\lambda = 2.5$
ReMo-4B	85.3	86.4	83.8	82.1	82.9	80.8	77.9	78.4	77.3
ReMo-8B	87.4	89.5	85.3	82.6	83.2	80.6	77.6	79.2	76.8
ReMo-16B	71.2	72.8	69.1	65.3	66.5	64.1	63.0	63.8	62.4

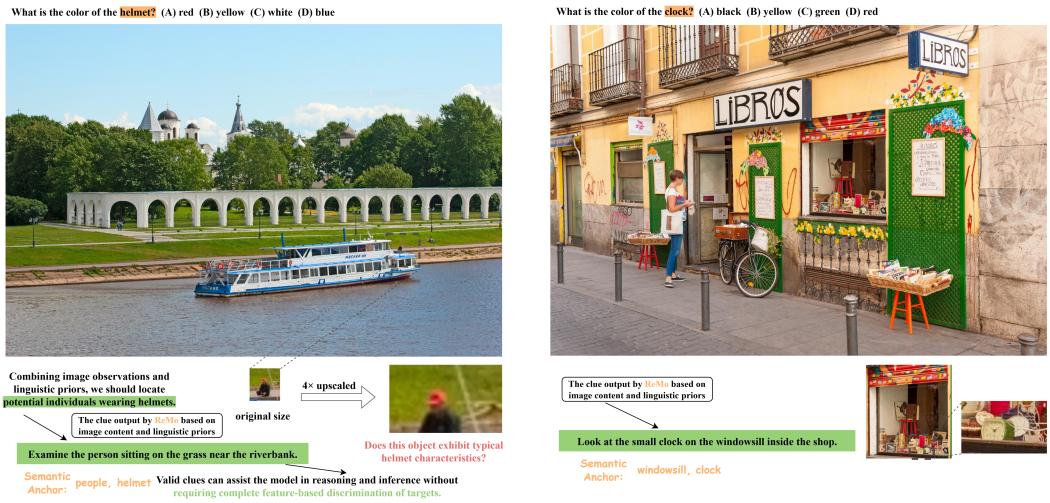


Figure 6: Visualization of how lightweight VLMs exploit global visual semantics and linguistic priors to generate actionable visual cues for fine-grained perception.

the core evidence but may still leave gaps within the localized regions. We then leverage the spatial modeling capability of the ViT to expand these seeds to their neighboring visual tokens, where the red tokens denote the initial seeds and the blue tokens denote newly recovered tokens. This process effectively fills the local spatial gaps while preserving the original visual context, resulting in a more complete set of task-relevant visual tokens. Finally, the union of the seed and newly completed tokens forms the visual evidence used by ReMo, increasing the selected tokens from 31 to 86 and 89 in the two examples, respectively. Importantly, the added tokens are not distributed uniformly over the image; instead, they remain concentrated around the initially localized regions, *indicating that ViT-based completion serves to recover missing local visual evidence rather than simply increasing the overall token density*. These visualizations also illustrate the complementary roles of the two stages: language-guided attention identifies where to look, cosine denoising refines the reliability of this localization, and ViT-based completion determines what additional visual tokens should be retained to ensure sufficient spatial coverage. Overall, the examples provide qualitative evidence that ReMo transforms a noisy and incomplete attention signal into a compact yet more spatially complete representation of the visual evidence required for downstream reasoning.

Hyperparameter Sensitivity of λ

We further investigate the sensitivity of the evidence-aware attention bias λ in Eq. (9). As shown in the table 4, the performance first improves as λ increases and reaches its highest value at $\lambda = 1.5$, while both smaller and larger values lead to inferior performance. This trend can be attributed to the nonlinear effect of the additive bias under softmax normalization. Specifically, for an evidence token, the bias changes its unnormalized attention weight from $\exp(s_j)$ to $\exp(s_j + \lambda) = e^{\lambda} \exp(s_j)$, effectively multiplying its relative weight by e^{λ} against non-evidence tokens. Thus, $\lambda = 1.0, 1.5, 2.5$, correspond to relative amplification factors of approximately $2.72, 4.48, 12.18$, respectively. A relatively small value such as $\lambda = 1.0$ may therefore provide

insufficient emphasis on the identified evidence, limiting its ability to redirect attention toward task-relevant regions. Increasing λ to 1.5 provides a stronger yet balanced preference for visual evidence, allowing the model to better exploit the localized and completed evidence while still preserving sufficient attention to the surrounding global context. In contrast, further increasing the bias to 2.5 can excessively concentrate attention on the selected evidence tokens. This may suppress complementary contextual information and introduce an overly sharp attention distribution, ultimately degrading visual reasoning performance. These results suggest that a moderate evidence bias is preferable **for balancing local evidence emphasis and global contextual preservation**, with $\lambda = 1.5$ providing the most effective trade-off in our experiments.

Language Priors Enable Effective Visual Cue Extraction from Global Semantics

The examples in Fig. 6 illustrate why a lightweight VLM can be sufficient for extracting visual cues in ReMo, without requiring strong fine-grained perception. The key observation is that cue extraction and fine-grained visual recognition are different requirements. A lightweight VLM primarily needs to identify the overall scene, salient objects, and coarse spatial relationships, which can then be combined with the input question and linguistic prior knowledge to infer where the relevant visual evidence is likely to appear. In the first example, although the helmet is visually small and difficult to recognize directly, the VLM can identify the scene as a riverside environment and the presence of an individual near the riverbank. Combined with the question and the prior that helmets are typically associated with people, the resulting cue directs the model to "examine the person sitting on the grass near the riverbank." The subsequent enlarged crop confirms that the localized region indeed contains the helmet. Similarly, in the second example, the VLM only needs to recognize the overall bookstore-like scene and the presence of objects displayed around the storefront. By incorporating the question about the clock and relevant linguistic knowledge, it generates the more specific cue "look at the small clock on the windowsill inside the shop," which successfully leads to the target region despite the clock being visually inconspicuous in the original image.

These examples suggest two complementary properties of lightweight VLMs for visual cue extraction. First, capability alignment: vision-language pretraining commonly emphasizes image–text alignment and holistic semantic description, making lightweight VLMs relatively effective at understanding global scene content even when their fine-grained visual recognition remains limited. Second, task feasibility: the cue generation stage does not require the model to precisely identify the target itself. Instead, the model only needs to combine coarse visual observations with the textual question, while the LLM contributes semantic priors and reasoning to infer a plausible location or visual property to inspect. Therefore, the required capability is substantially weaker than directly solving the original fine-grained perception task. This observation motivates our design of using a lightweight VLM as a cue extractor rather than a direct visual answerer, allowing its strong global semantic understanding and the LLM's linguistic prior knowledge to jointly produce actionable visual guidance for subsequent token-level visual focusing.

Layer-wise Visualization of Intermediate-Layer Visual Attention The choice of intermediate VLM layers in Eq. (2) is motivated by the distinct spatial characteristics of visual attention across network depth. Fig. 7 visualizes the attention maps from representative early, middle, and late layers for an image containing an airplane. In the early layers (e.g., layers 1, 4, and 11), the attention responses are relatively diffuse and distributed across broad image regions, suggesting that these layers primarily capture low-level visual patterns and coarse image structures, which provide limited spatial discrimination for identifying task-relevant regions. In contrast, the middle layers (e.g., layers 15, 17, 18, and 21) exhibit a clearer spatial concentration around the airplane, producing a more distinct visual focus while still retaining sufficient spatial coverage over the surrounding scene. This behavior makes intermediate-layer attention particularly suitable as an initial localization signal for identifying relevant visual tokens. In the late layers (e.g., layers 27, 30, 33, and 35), the attention becomes increasingly sparse and exhibits isolated high-response points at locations that are not consistently aligned with the target object, indicating that deeper representations may contain more task- or language-dependent concentration and reduced spatial fidelity for token-level localization. Therefore, rather than directly relying on a single layer, we aggregate attention from multiple intermediate layers to obtain a more stable and spatially informative localization map. This provides a principled basis for using intermediate-layer attention in Eq. (2): the middle layers offer a favorable balance between semantic abstraction and spatial localization, making them more suitable for identifying the visual regions that should be retained for subsequent token selection.

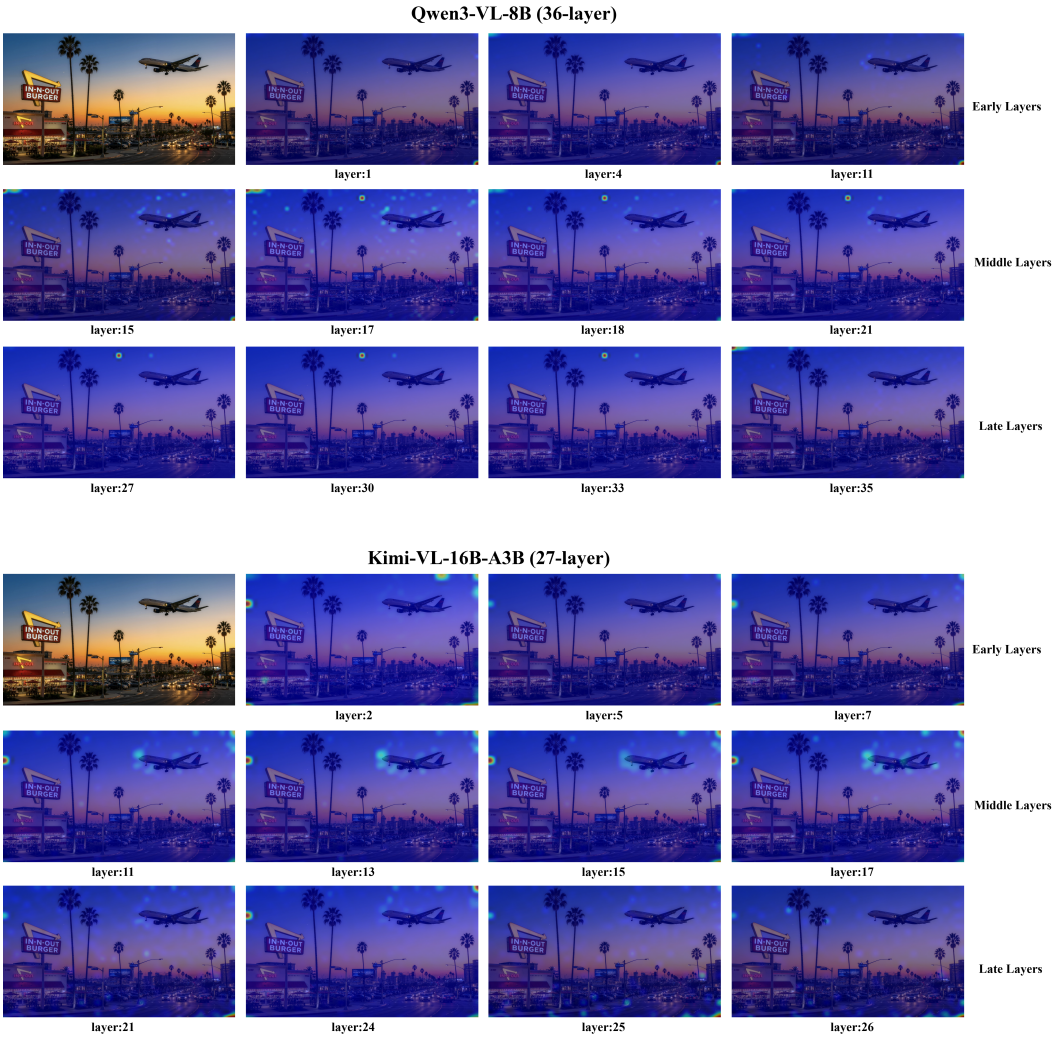


Figure 7: Layer-wise visualization of visual attention across early, middle, and late VLM layers. Intermediate layers exhibit clearer spatial concentration on the task-relevant airplane region, providing a more suitable basis for visual token localization.

Shape Perception under Challenging Visual Conditions

Beyond the mainstream VLM perception benchmarks, we further construct a dedicated Shape Perception Test to more directly examine the models' ability to recognize and distinguish geometric shapes under challenging visual conditions. The test contains 215 manually generated and annotated hard cases, with particular emphasis on two factors that can interfere with shape perception: viewpoint/angle variations and lighting and shadow effects. As illustrated in Fig. 8, these factors can substantially alter the apparent contour and local visual cues of objects, making seemingly simple shape-counting questions difficult for VLMs. For example, the trapezoids and triangles in the examples exhibit different orientations, perspectives, and shadow patterns, which can lead the vanilla model to misidentify the target shapes. In contrast, ReMo yields more accurate predictions on these cases, demonstrating that its knowledge invocation, visual focusing, and completion enhancement mechanisms jointly and effectively improve shape discrimination performance.

Quantitatively, ReMo consistently improves shape perception accuracy across all three evaluated VLMs. On Qwen3-VL-4B, the accuracy increases from 33.95% to 40.47%, corresponding to an absolute improvement of 6.52 percentage points. On Qwen3-VL-8B, the accuracy improves from

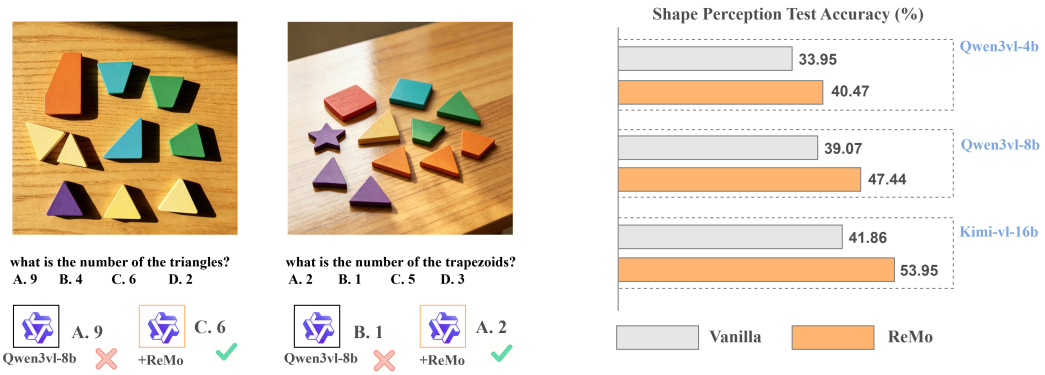


Figure 8: Evaluation on a dedicated shape perception test containing 215 challenging cases with viewpoint, angle, illumination, and shadow variations. Gray and orange bars denote the performance of the vanilla models and ReMo, respectively. ReMo consistently improves shape perception accuracy across Qwen3-VL-4B, Qwen3-VL-8B, and Kimi-VL-16B.

39.07% to 47.44%, yielding a gain of 8.37 percentage points. On Kimi-VL-16B, the accuracy increases from 41.86% to 53.95%, corresponding to a gain of 12.09 percentage points. The improvements are consistently observed across the three evaluated models, including both different model scales and architectures. These results show that ReMo can also provide measurable gains on this dedicated shape perception test, complementing the evaluation on standard VLM benchmarks and providing an additional assessment of visual perception under challenging geometric and appearance variations.

Experimental Implementation Details

As ReMo is a training-free visual perception enhancement method, it requires no additional training, parameter updates, or optimization. All experiments are conducted on NVIDIA H20 GPUs with 96GB memory. For all evaluated VLMs, we use their default system prompts and preserve the original decoding strategies and inference configurations, including the temperature and other generation parameters, without additional tuning unless explicitly specified. This ensures that different methods are compared under consistent inference settings. For the hyperparameters in Eqs. (4) and Eq. (6), we set γ= 2 in Eq. (4) to further amplify relative differences, and $\epsilon = 10^{-8}$, intensity scaling coefficien α is empirically set to 3. In Eq. (5), We set $K_{\text{top}} = \lfloor 0.05N \rfloor$, corresponding to the top 5% of visual tokens. The performance is relatively stable for ratios of 5%–10%, while increasing the ratio to 20% decreases Vstar accuracy by 2.8 points. This indicates that an excessively large seed set may introduce irrelevant tokens and subsequently propagate noise during visual completion. We therefore adopt 5% as a conservative setting. In Eq. (6), the radius r is set to 6. This choice roughly corresponds to a 12 × 12 patch region with patches of size 16 × 16, a region of approximately **196 × 196 pixels**, which is sufficiently large for effective relevant-token search while avoiding the introduction of excessive noise.

For benchmark evaluation, we follow the evaluation protocols of the corresponding datasets and prior works. We note that the evaluation procedures and result reporting of MMStar and HallusionBench differ in certain aspects from those adopted in some official technical reports. Consequently, the baseline results reproduced in our experiments may not exactly match the numbers reported in those reports. To ensure a consistent evaluation pipeline with our experimental setting, we implement and run our own baseline evaluation scripts for both benchmarks, which will be provided in the supplementary material. The reproduced baseline results are broadly consistent with those reported in the corresponding original papers (Yu et al., 2026)and (Aneja et al., 2026), providing further evidence that our evaluation implementations are properly aligned with the established protocols. **Corresponding code is available in the supplementary materials**.