

RAVE: Re-Allocating Visual Attention in Large Multimodal Models

Xi Leng^{1,2}, Xinhong Ma¹, Ziqiang Dong¹, Feng Zhang^{1,3},
Xiaoying Tang², Yang Yang¹, Guanjin Jiang¹

¹Qwen Business Unit of Alibaba, ²The Chinese University of Hong Kong, Shenzhen,

³Beijing Institute of Technology

{xinhong.mxh, ziqiang.dzq, chris.yang, guanjin.jiang}@alibaba-inc.com
xileng@link.cuhk.edu.cn, bit_zhangfeng@bit.edu.cn

Abstract

Large multimodal models (LMMs) inherit the self-attention mechanism of pretrained language backbones, yet standard attention can exhibit suboptimal allocation, including cross-modal misallocation between textual and visual evidence and intra-visual imbalance among visual tokens. We propose RAVE (Re-Allocating Visual Attention), a lightweight pair-gating mechanism that adds a learned query-key bias to pre-softmax attention scores over visual keys, derived from pre-ROPE query and key features. RAVE requires no architectural modification to the backbone and can be trained end-to-end with the rest of the model. Across a suite of multimodal benchmarks, RAVE improves over standard attention by 3 points on average, with the largest gains on perception-intensive tasks—including multilingual OCR, chart understanding, document VQA, and scene text VQA—where accurate visual grounding is critical.

1 Introduction

A prevalent architectural paradigm for large multimodal models (LMMs), exemplified by the LLaVA family (Liu et al., 2023a, 2024) and the Qwen-VL series (Wang et al., 2024; Bai et al., 2025a,b), projects visual features into the language token space and processes them alongside textual tokens in a unified sequence. This approach relies on the self-attention mechanism inherited from a pretrained language backbone, which we refer to as *standard attention* throughout this work. While simple and effective, this inherited mechanism is not explicitly designed to regulate modality-wise attention allocation. Recent observations suggest that such uniform treatment of visual and textual tokens can lead to suboptimal use of visual evidence during generation, weakening visual grounding in complex multimodal tasks.

This suboptimal allocation manifests at two levels. At the *cross-modal level*, the visual block as

a whole may receive insufficient attention relative to textual tokens, reflecting architectural biases inherited from language backbones. This imbalance appears both across decoder depth, where visual tokens become increasingly under-attended in deeper layers (Chen et al., 2024a; Zhang et al., 2025b), and during decoding, where attention to the visual block decreases as generated answer tokens accumulate (Sun et al., 2025; Tu et al., 2025; Yang et al., 2025; Tang et al., 2025a). At the *intra-visual level*, attention within the visual block may be unevenly distributed: visually uninformative patches can absorb disproportionate mass, forming “visual sinks” (Kang et al., 2025), while positional bias may further concentrate attention on specific spatial regions (Kuo et al., 2025). Despite their different manifestations, these phenomena point to a common limitation: standard attention lacks an explicit mechanism for regulating visual attention allocation during multimodal generation. As a result, visual evidence may be under-utilized or misallocated, ultimately undermining semantically precise visual grounding.

Existing approaches mitigate these problems through inference-time attention intervention (Tu et al., 2025; Kang et al., 2025; Tang et al., 2025a; Zhou et al., 2025), architectural redesign (Zhang et al., 2025a; Kuo et al., 2025; Kim et al., 2025), or learned visual revisiting (Sun et al., 2025; Yang et al., 2025). While effective in specific scenarios, they often rely on hand-crafted intervention heuristics with limited context awareness, incur structural departures from standard Transformer attention that may complicate integration, or require auxiliary training pipelines involving curated data and specialized generation protocols. There remains a need for a minimalist, drop-in mechanism that enables context-aware recalibration of attention allocation, while remaining compatible with standard training and inference workflows.

To this end, we propose RAVE (**Re-Allocating**

Method	Training-free	Extra supervision	Arch. change	Input-adaptive	Cross-modal	Intra-visual	Lightweight integration
ATTNREAL (Tu et al., 2025)	✓	✗	✗	✗	✓	✗	✓
VAR (Kang et al., 2025)	✓	✗	✗	<i>Partial</i>	✗	✓	✓
CAUSALMM (Zhou et al., 2025)	✓	✗	✗	<i>Partial</i>	✓	✗	✓
FARSIGHT (Tang et al., 2025a)	✓	✗	✗	<i>Partial</i>	✓	✗	✓
D-ATTN (Kuo et al., 2025)	✗	✗	✓	✓	✓	<i>Partial</i>	✗
LLAViT (Kim et al., 2025)	✗	✗	✓	✓	<i>Partial</i>	✓	✗
TVC (Sun et al., 2025)	✗	✓	✗	✓	✓	✗	<i>Partial</i>
LOOK-BACK (Yang et al., 2025)	✗	✓	✗	✓	✓	✗	<i>Partial</i>
GATED-ATTN (Qiu et al., 2025)	✗	✗	<i>Partial</i>	✓	<i>Partial</i>	<i>Partial</i>	<i>Partial</i>
RAVE (ours)	✗	✗	✗	✓	✓	✓	✓

Table 1: Positioning of RAVE against representative methods for improving visual attention allocation in LMMs.

Visual Attention), a lightweight pair-gating mechanism that corrects standard attention in a drop-in manner. RAVE adds a learned query–key bias to the pre-softmax attention scores over visual keys, derived from pre-ROPE query and key features. This enables context-adaptive recalibration of visual attention within the native attention operator, without altering the backbone architecture or disrupting standard end-to-end training.

We evaluate RAVE on a diverse suite of multimodal benchmarks covering both perception-intensive and reasoning-oriented tasks. RAVE consistently improves over standard attention, achieving an average gain of 3 points. The improvements are particularly pronounced on perception-intensive tasks, including multilingual OCR, chart understanding, document understanding, and text-rich VQA, where accurate visual grounding is critical. These results show the effectiveness of context-aware attention recalibration within standard attention for improving multimodal performance.

Our contributions are threefold:

- We revisit standard attention in LMMs and identify suboptimal visual attention allocation as a key bottleneck for effective visual grounding in multimodal generation.
- We propose RAVE, a lightweight, architecture-preserving pair-gating mechanism that modulates pre-softmax attention scores over visual keys with negligible additional parameters.
- We demonstrate that RAVE consistently outperforms standard attention across a wide range of multimodal benchmarks, achieving 3-point average gains, with particularly strong improvements on perception-intensive tasks.

2 Related Work

Attention allocation in large multimodal models. Modern large multimodal models (LMMs),

including the LLAVA family and the Qwen-VL series, typically project visual features into the language token space and process them together with textual tokens in a unified sequence using a decoder-style transformer. Under this paradigm, the inherited self-attention mechanism serves as the default interface for multimodal fusion, making visual grounding highly dependent on how attention is allocated across visual and textual tokens during generation. As a result, attention allocation has become an important lens for understanding and improving multimodal reasoning in LMMs.

Visual-attention misallocation in LMMs. Recent studies have identified two complementary forms of suboptimal visual attention allocation. **Cross-modal imbalance:** *visual tokens, as a block, are systematically under-attended relative to textual tokens.* This imbalance manifests along two axes. *Across decoder depth,* Chen et al. (2024a) show on LLAVA-class models that, despite occupying a substantial fraction of the input sequence, visual tokens receive significantly less attention than textual tokens, with the disparity increasing in deeper layers. Zhang et al. (2025b) attribute this effect to a combination of language-side bias in attention scores and layer-wise decay in activation magnitude, and show that such imbalance can severely degrade fine-grained perception. *During decoding,* a temporally indexed counterpart has been reported under different names: Sun et al. (2025) describe “visual forgetting,” where attention to the image decays as long-chain reasoning progresses; Tu et al. (2025) observe that historical output tokens increasingly dominate attention as generation unfolds; and Yang et al. (2025) report that visual attention can diminish to near zero in later reasoning stages. Tang et al. (2025a) further relate this temporal decay to the distance-dependent attenuation induced by rotary position embeddings (Su et al., 2024). Collectively, these studies suggest

that visual attention can be systematically disadvantaged by architectural and positional biases in standard attention, which can compromise reliable visual grounding and weaken fine-grained perception. ***Intra-visual imbalance:*** *within the visual block, attention is unevenly distributed across visual tokens.* This imbalance also manifests in two forms. First, Xiao et al. (2024) identify “attention sinks” in language models, where a small number of tokens absorb disproportionate attention regardless of semantic relevance. Kang et al. (2025) show that an analogous phenomenon arises in LMMs, where attention concentrates on visually uninformative tokens such as corner patches or homogeneous background regions. Tang et al. (2025a) further describe a related “attention collapse,” in which outlier tokens—arising from either modality—receive disproportionately high attention scores. Second, Kuo et al. (2025) identify a position-induced bias within the visual block: with row-major flattened image tokens, positional encoding can cause textual queries to favor spatially proximate patches in the 1D sequence, leading to attention that is spatially biased rather than semantically grounded. Consequently, even when nontrivial attention is assigned to the visual block, a substantial portion of that mass may be allocated to uninformative or systematically biased regions.

Together, these cross-modal and intra-visual imbalances point to a key limitation of standard attention: it lacks an explicit mechanism for regulating visual attention allocation during multimodal generation, a gap that RAVE aims to bridge.

Mitigating visual-attention misallocation in LMMs. Existing mitigation strategies can be grouped into three families, each operating at a different place in the pipeline. Table 1 summarizes representative methods along the axes most relevant to our discussion. ***Training-free methods intervene at inference time.*** ATTNREAL (Tu et al., 2025) recycles excessive attention from historical output tokens back to visual tokens; VAR (Kang et al., 2025) identifies visual sink tokens and redistributes their surplus attention toward more informative visual regions, reducing intra-visual attention bias; CAUSALMM (Zhou et al., 2025) applies causal adjustment and counterfactual attention interventions to mitigate modality-prior-induced hallucinations; FARSIGHT (Tang et al., 2025a) modifies the causal mask to absorb attention from outlier tokens and uses a progressively di-

minishing mask schedule to counter RoPE-induced distance-dependent attenuation during decoding. Although these methods require no additional training, they typically rely on hand-crafted intervention rules rather than learned, input-adaptive corrections. ***Architectural methods redesign multimodal attention.*** D-ATTN (Kuo et al., 2025) decomposes self-attention into modality-specific paths (vision-to-vision, text-to-vision, text-to-text) and modifies the positional treatment of visual tokens; LLAViT (Kim et al., 2025) learns separate $\{Q, K, V\}$ projections for visual tokens and enables bidirectional attention within the visual block; AKI (Wang et al., 2025) relaxes the causal mask to allow visual tokens to attend to subsequent question tokens. These methods can be effective, but they often introduce additional architectural complexity or require non-trivial re-alignment with existing LMM checkpoints. ***Training-based revisiting methods learn explicit visual reactivation behaviors.*** TVC (Sun et al., 2025) trains the model to re-attend to visual input by inserting visual embeddings and bridging prompts into long-chain reasoning data, and further applies inference-time calibration by periodically reintroducing compressed visual tokens; LOOK-BACK (Yang et al., 2025) uses supervised fine-tuning and reinforcement learning to elicit a special $\langle \text{back} \rangle$ token that redirects attention to the image. These methods can improve visual utilization, but they require auxiliary training pipelines with curated reasoning data and specialized generation protocols.

RAVE fills a distinct design space: a lightweight, learnable module placed directly inside the native attention. While it shares the gating philosophy of recent works such as GATED-ATTN (Qiu et al., 2025), RAVE shifts the intervention from the attention output to the pre-softmax attention scores associated with visual keys. This shift is fundamental: whereas GATED-ATTN modulates how attended features are integrated, RAVE modulates where attention is allocated. By directly recalibrating the competition for attention mass, RAVE provides a mechanism for regulating visual attention allocation that is relevant to both cross-modal and intra-visual misallocation, while remaining natively compatible with standard LMM training and inference.

3 Preliminaries

To analyze attention allocation in decoder-only LMMs, we adopt segment-wise attention mass, a

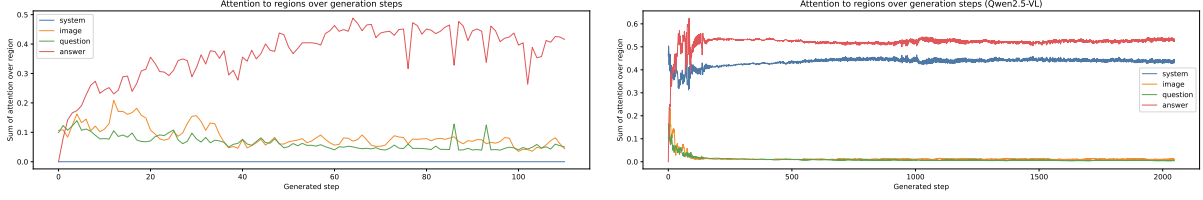


Figure 1: Attention mass assigned by an answer token to the four input segments—*system*, *image*, *question*, and *answer*—as generation proceeds, averaged over decoder layers and attention heads. **Left:** Qwen3-VL-4B-Instruct, which does not apply a default system prompt, so the *system* mass is identically zero. **Right:** Qwen2.5-VL-7B-Instruct with its default system prompt “You are a helpful assistant”.

commonly used diagnostic statistic for characterizing token-type attention patterns and visual-token utilization (Chen et al., 2024a; Tu et al., 2025; Yang et al., 2025). Following this line of analysis, we instantiate the statistic under the segment partition used in this work. Specifically, we separate the context into system-prompt tokens $\mathcal{I}_{\text{sys}}$ introduced by the chat template, visual tokens $\mathcal{I}_{\text{img}}$ produced by the vision encoder, question tokens $\mathcal{I}_{\text{que}}$ from the user instruction, and answer tokens $\mathcal{I}_{\text{ans}}$ generated autoregressively by the model. We write N for the total sequence length after generation and let $\mathcal{I}_s \subseteq \{1, \dots, N\}$ denote the set of token positions belonging to segment $s \in \{\text{sys}, \text{img}, \text{que}, \text{ans}\}$. At layer ℓ and head h , standard attention with rotary position embedding (Su et al., 2024) computes:

$$Q^{\ell,h} = \text{RoPE}(\bar{Q}^{\ell,h}), K^{\ell,h} = \text{RoPE}(\bar{K}^{\ell,h}), \quad (1)$$

$$L_{ij}^{\ell,h} = \frac{\langle Q_i^{\ell,h}, K_j^{\ell,h} \rangle}{\sqrt{d_k}}, A_{ij}^{\ell,h} = \text{softmax}_j(L_{ij}^{\ell,h} + M_{ij}) \quad (2)$$

where $\bar{Q}, \bar{K} \in \mathbb{R}^{N \times d_k}$ denote pre-ROPE query and key projections; M is the causal mask; L and A are the attention scores and probabilities. With grouped-query attention (Ainslie et al., 2023), multiple query heads share one key/value head; thus, throughout this work, “head” refers to a query head, and the gating module is instantiated at the query-head granularity in Section 5.

We characterize attention allocation across input segments from two complementary views. For an answer position $t \in \mathcal{I}_{\text{ans}}$, we first compute the *segment-wise layer-averaged attention mass*:

$$\alpha_s^{(t)} = \frac{1}{HL} \sum_{\ell=1}^L \sum_{h=1}^H \sum_{j \in \mathcal{I}_s} A_{tj}^{\ell,h}, \quad (3)$$

$$s \in \{\text{sys}, \text{img}, \text{que}, \text{ans}\},$$

which aggregates, over all decoder layers ℓ and heads h , the attention mass assigned by token t to segment s . By construction, $\sum_s \alpha_s^{(t)} = 1$ at each

decoding step, so $\{\alpha_s^{(t)}\}_s$ can be interpreted as a distribution of attention mass over input segments.

To examine how attention allocation varies across decoder layers, we also compute a *layer-resolved* variant:

$$\alpha_s^{(t),\ell} = \frac{1}{H} \sum_{h=1}^H \sum_{j \in \mathcal{I}_s} A_{tj}^{\ell,h}, \quad (4)$$

which preserves the layer dimension while using the same segment partition.

4 Motivation

Setup. We analyze the attention allocation dynamics of Qwen3-VL-4B-Instruct (Bai et al., 2025a) and Qwen2.5-VL-7B-Instruct (Bai et al., 2025b). To cover different visual grounding scenarios, we use representative probe inputs spanning *visual perception* tasks (e.g., dense OCR and chart understanding) and *visual-textual reasoning* tasks (e.g., mathematical and physics diagrams). For each example, we track both the segment-wise layer-averaged mass $\alpha_s^{(t)}$ and the layer-resolved mass $\alpha_s^{(t),\ell}$ throughout decoding.

Observation 1: Standard attention exhibits progressive visual-attention dilution. Figure 1 shows the segment-wise layer-averaged attention mass $\alpha_s^{(t)}$ on a symbolic mathematical reasoning example for Qwen3-VL-4B-Instruct and Qwen2.5-VL-7B-Instruct. In both models, attention allocation shifts as decoding proceeds: visual-token mass starts at a non-negligible level but steadily declines, whereas answer-token mass grows in parallel. This suggests a gradual transition from visually grounded generation to increasingly self-conditioned generation based on the model’s own textual history. The trend is consistent with prior reports of visual-attention dilution in different settings, including mathematical reasoning (Sun et al., 2025), hallucination snowballing (Tang et al., 2025a), and fine-grained cognition and emotion

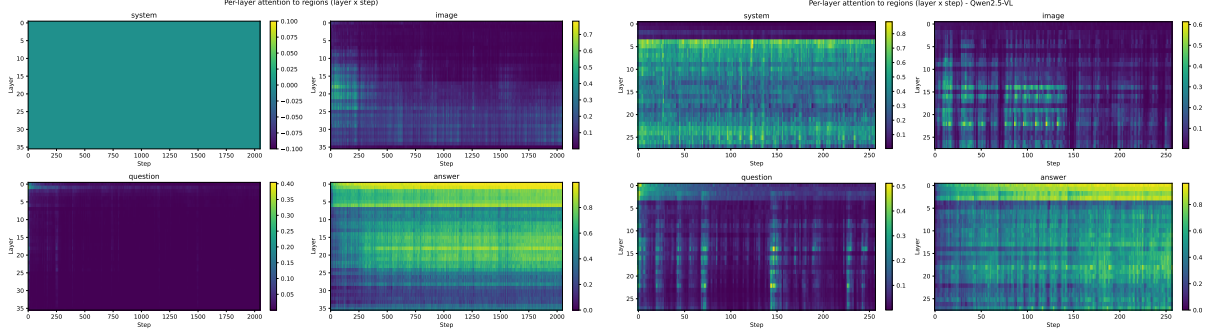


Figure 2: Per-layer attention mass (layer $\times$ decoding step) for Qwen3-VL-4B-Instruct (left) and Qwen2.5-VL-7B-Instruct (right). Each panel shows, across layers and decoding steps, the total attention mass assigned by the currently generated token to the four input segments: *system*, *image*, *question*, and *answer*.

understanding (Zhang et al., 2025b), and further indicates that visual under-allocation remains a recurring issue in recent decoder-only LMMs.

The right panel of Figure 1 illustrates a negative case associated with visual-attention dilution. In Qwen2.5-VL-7B-Instruct, visual attention decreases during generation, while nontrivial attention remains on the default system prompt “You are a helpful assistant”, which carries little task-specific content. This prompt-side concentration resembles attention-sink behavior in language models (Xiao et al., 2024). The model consequently produces unstable image-conditioned reasoning and repeats until reaching the decoding limit. By contrast, Qwen3-VL-4B-Instruct, which does not use a default system prompt in our setting, avoids this prompt-side concentration but still shows declining visual attention.

Observation 2: The dilution pattern is visible across decoder layers. The same pattern is also visible in the layer-resolved view. Figure 2 presents $\alpha_s^{(t),\ell}$ on a representative OCR-style input for the same two backbones. Despite being substantially different from the mathematical reasoning example in Figure 1, it exhibits a similar trend during decoding: visual-token mass decreases, while answer-token mass increases. The layer-resolved view further shows that this shift is a decoding-stage phenomenon that persists throughout the decoder rather than a depth-specific effect. These observations suggest that visual-attention dilution is not merely a task-specific artifact of one example, but a recurring allocation pattern in standard attention.

5 Method

Motivated by the visual-attention dilution observed in Section 4 and the intra-visual attention imbalance discussed in Section 2, we propose RAVE (Re-

Allocating Visual Attention), a lightweight pair-gating mechanism that serves as a drop-in correction to standard attention in decoder-only LMMs.

5.1 Pair gate from pre-ROPE query and key

Let $\bar{Q}^{\ell,h}, \bar{K}^{\ell,h} \in \mathbb{R}^{N \times d_k}$ denote the *pre-rotary* query and key features of layer ℓ and head h , i.e., the inputs to RoPE($\cdot$) in Eq. (1). We derive the gating signal from these pre-ROPE representations rather than from the post-ROPE Q and K . This choice is motivated by recent analyses showing that rotary position encoding can introduce systematic bias into cross-modal attention (Zhao et al., 2025; Tang et al., 2025a; Kuo et al., 2025). Using $\bar{Q}$ and $\bar{K}$ therefore helps decouple the gating signal from positional decay, allowing it to depend primarily on modality-conditioned content features.

At each layer ℓ , we attach a pair of learnable linear projections

$$W_{gq}^\ell, W_{gk}^\ell \in \mathbb{R}^{d_k \times 1}, \quad (5)$$

shared across all H query heads in that layer. For each head h , these projections produce per-token scalar scores

$$s_q^{\ell,h} = \bar{Q}^{\ell,h} W_{gq}^\ell \in \mathbb{R}^N, \quad s_k^{\ell,h} = \bar{K}^{\ell,h} W_{gk}^\ell \in \mathbb{R}^N. \quad (6)$$

We then form a pairwise gate by taking the outer product of these scores:

$$G_{ij}^{\ell,h} = \phi\left(s_{q,i}^{\ell,h} s_{k,j}^{\ell,h}\right), \quad (7)$$

where ϕ maps the pairwise score into the gating bias used for visual attention recalibration. Since each entry $G_{ij}^{\ell,h}$ is computed from the current query token and a candidate key token, the gate provides a context-aware correction for each query–key pair. The projections are shared within each layer, introducing only negligible parameter overhead. This

design preserves the original attention structure of the backbone and can therefore be seamlessly integrated into a wide range of decoder-only LMMs.

5.2 Pre-softmax additive recalibration over visual keys

Based on the standard attention logits $L_{ij}^{\ell,h}$ defined in Eq. (2), RAVE injects the pairwise gate $G_{ij}^{\ell,h}$ into the pre-softmax scores associated with visual keys:

$$\tilde{L}_{ij}^{\ell,h} = L_{ij}^{\ell,h} + G_{ij}^{\ell,h} \cdot \mathbb{1}[j \in \mathcal{I}_{\text{img}}], \quad (8)$$

where $\mathbb{1}[\cdot]$ is the indicator function. The recalibrated attention distribution is then computed as

$$\tilde{A}_{ij}^{\ell,h} = \text{softmax}_j \left(\tilde{L}_{ij}^{\ell,h} + M_{ij} \right), \quad (9)$$

where M_{ij} denotes the causal attention mask. Since the correction is added before normalization, RAVE recalibrates how visual and textual keys compete for attention probability mass within the native attention operator, rather than reweighting an already normalized attention distribution. As this correction is computed at the query-key-pair level for visual keys, it also enables redistribution within the visual-token set, providing a direct handle on intra-visual attention imbalance. Restricting the correction to visual-key positions keeps text-key logits intact, thereby limiting direct interference with the pretrained language-side attention prior.

5.3 Prefill-aware and GQA-aware partial-head application

RAVE recalibrates attention to visual keys during autoregressive decoding for each answer-position query, modulating both the strength of visual evidence incorporation and its distribution across visual tokens. The same recalibration is also applied during prefill, where question tokens adjust their attention to visual keys and refine the image-conditioned representations that subsequent answer generation builds upon.

Since RAVE is introduced into decoder-only LMMs built on a pretrained language backbone, we use a partial-head intervention to avoid over-disrupting the inherited language prior. This partition is implemented in a GQA-aware manner. Modern decoder-only LMMs commonly adopt grouped-query attention (GQA), where H_q query heads share H_{kv} key/value heads. We refer to the query heads that share the same key/value head as one GQA group. Let $r = H_q/H_{kv}$ denote the number

Algorithm 1: Forward pass of RAVE attention for a selected query head $h \in \mathcal{H}_{\text{rave}}$; ungated heads use vanilla attention.

Input: Pre-ROPE query/key features

$\bar{Q}^{\ell,h}, \bar{K}^{\ell,h} \in \mathbb{R}^{N \times d_k}$; values

$V^{\ell,h} \in \mathbb{R}^{N \times d_v}$; visual-token index set $\mathcal{I}_{\text{img}}$;

layer-shared gate weights

$W_{gq}^\ell, W_{gk}^\ell \in \mathbb{R}^{d_k \times 1}$; causal mask M .

Output: Output features $O^{\ell,h} \in \mathbb{R}^{N \times d_v}$.

1 $Q^{\ell,h} \leftarrow \text{RoPE}(\bar{Q}^{\ell,h}), \quad K^{\ell,h} \leftarrow \text{RoPE}(\bar{K}^{\ell,h});$

2 $L^{\ell,h} \leftarrow Q^{\ell,h} (K^{\ell,h})^\top / \sqrt{d_k};$

3 $s_q^{\ell,h} \leftarrow \bar{Q}^{\ell,h} W_{gq}^\ell, \quad s_k^{\ell,h} \leftarrow \bar{K}^{\ell,h} W_{gk}^\ell;$

4 $G^{\ell,h} \leftarrow \phi(s_q^{\ell,h} (s_k^{\ell,h})^\top);$

5 $\tilde{L}^{\ell,h} \leftarrow L^{\ell,h};$

6 $\tilde{L}_{:, \mathcal{I}_{\text{img}}}^{\ell,h} \leftarrow L_{:, \mathcal{I}_{\text{img}}}^{\ell,h} + G_{:, \mathcal{I}_{\text{img}}}^{\ell,h};$

7 $\tilde{A}^{\ell,h} \leftarrow \text{softmax}_j(\tilde{L}^{\ell,h} + M);$

8 $O^{\ell,h} \leftarrow \tilde{A}^{\ell,h} V^{\ell,h};$

9 **return** $O^{\ell,h};$

of query heads in each group. For each GQA group, we select a fraction p of its r query heads for RAVE and leave the remaining query heads unchanged.

Let $\mathcal{H}_{\text{rave}}$ denote the selected query-head subset, with $|\mathcal{H}_{\text{rave}}| = p H_q$. For heads in $\mathcal{H}_{\text{rave}}$, we use the recalibrated logits $\tilde{L}_{ij}^{\ell,h}$; for the remaining heads, we keep the original logits $L_{ij}^{\ell,h}$:

$$\hat{L}_{ij}^{\ell,h} = \begin{cases} \tilde{L}_{ij}^{\ell,h}, & h \in \mathcal{H}_{\text{rave}}, \\ L_{ij}^{\ell,h}, & h \notin \mathcal{H}_{\text{rave}}. \end{cases} \quad (10)$$

We use $p = 0.25$ in our main configuration and study the effect of different head ratios in Section 6.

5.4 The RAVE algorithm

We initialize W_{gq}^ℓ and W_{gk}^ℓ near zero and use tanh for ϕ in Equation (7), so that $G_{ij}^{\ell,h} \approx 0$ at the beginning of training and the pretrained backbone behavior is preserved. Since RAVE adds only two layer-wise scalar projections shared across query heads, its parameter overhead is $2Ld_k$, negligible compared with the backbone size.

Algorithm 1 summarizes the forward pass of RAVE for a selected gated query head. During training, the gate parameters are optimized jointly with the LMM under the standard next-token cross-entropy loss, without auxiliary supervision or an additional gating objective. Overall, RAVE provides an end-to-end trainable mechanism for visual attention recalibration, enabling context-aware modulation that helps mitigate both cross-modal and intra-visual attention misallocation, while keeping

Attention Method	Perception-intensive						Reasoning / Knowledge-heavy			Perc. Avg.	Reason. Avg.	Overall Avg.
	MTVQA	ChartQA	DocVQA	OCRBench	InfoVQA	TextVQA	MathVista	MMMU	MMMU-Pro			
<i>Standard Attention</i>	6.51	23.20	35.85	40.60	24.86	62.17	24.60	30.56	20.00	32.20	25.05	29.82
VAR	6.46	22.60	35.56	40.10	24.29	62.39	24.50	31.00	19.03	31.90	24.84	29.55
D-ATTN	6.23	23.92	35.14	39.40	26.21	62.18	27.90	39.33	20.29	32.18	29.17	31.18
GATED-ATTN	6.05	20.92	33.61	37.70	24.22	60.54	24.80	31.16	15.83	30.51	23.93	28.31
RAVE	9.50	26.28	43.55	44.20	28.21	65.90	25.40	32.22	20.16	36.26	25.93	32.81

Table 2: Main results under the controlled Mistral-v0.3-7B setting. The trainable variants use matched model and training configurations, whereas VAR is applied post hoc to the same trained *Standard Attention* checkpoint. Benchmarks are grouped into perception-intensive and reasoning/knowledge-heavy tasks.

the backbone architecture and standard LMM training pipeline unchanged.

6 Experiments

6.1 Setup

Experimental protocol and baselines. Following the controlled protocol of D-ATTN (Kuo et al., 2025), we aim to isolate the effect of the attention mechanism rather than pursue state-of-the-art performance through larger-scale data or model scaling. We compare RAVE with four attention baselines under matched or maximally comparable settings: *Standard Attention*, VAR, D-ATTN, and GATED-ATTN. The trainable variants use the same language backbone, vision encoder, projector, training data, and optimization schedule, allowing performance differences to be more directly attributed to the attention mechanism. *Standard Attention* retains the original causal self-attention; D-ATTN decomposes self-attention into visual-to-visual, textual-to-visual, and textual-to-textual paths; and GATED-ATTN applies learned output-side *O*-gating. As a training-free method, VAR is applied post hoc to the same trained *Standard Attention* checkpoint without additional training. Detailed construction, implementation, and comparability settings are provided in Appendix A.

Benchmarks. We evaluate on a suite of multimodal benchmarks covering both perception-intensive and reasoning-oriented capabilities using the VLMEvalKit evaluation framework (Duan et al., 2024). The perception-intensive group includes multilingual OCR, document understanding, chart understanding, and text-rich visual question answering, with MTVQA-TEST (Tang et al., 2025b), ChartQA-TEST (Masry et al., 2022), DocVQA-VAL (Mathew et al., 2021), OCRBench (Liu et al., 2023b), InfoVQA-VAL (Mathew et al., 2022), and TextVQA-VAL (Singh et al., 2019). The reasoning-oriented group includes MathVista-MINI (Lu et al., 2024), MMMU-DEV-

VAL (Yue et al., 2024), and MMMU-Pro-10c (Yue et al., 2025), which require visual reasoning, mathematical reasoning, and domain knowledge. This benchmark selection allows us to examine whether visual attention recalibration improves fine-grained visual evidence utilization while preserving general multimodal reasoning ability. For each benchmark, we report the evaluation score produced by VLMEvalKit, together with Perception Avg., Reasoning Avg., and Overall Avg. over the corresponding groups.

6.2 Main results

Table 2 summarizes the main results across the evaluated benchmarks. Under the matched Mistral-v0.3-7B setting, RAVE achieves the best overall average among the compared attention mechanisms. Since the comparisons follow a strictly matched training and evaluation protocol, the observed gains can be more directly attributed to the proposed visual attention recalibration mechanism.

The improvement of RAVE is most pronounced in perception-intensive scenarios, such as OCR, document understanding, chart understanding, and text-rich VQA, where accurate generation relies heavily on fine-grained visual evidence. This trend aligns with the design of RAVE: the query-dependent pre-softmax correction to visual-key logits helps the model dynamically modulate both the amount of visual evidence incorporated into generation and the distribution of attention among visual tokens.

On reasoning- and knowledge-heavy benchmarks, such as MathVista, MMMU, and MMMU-Pro, the trend is more nuanced. RAVE improves over Standard Attention on average, but D-ATTN remains stronger on this group. This suggests that RAVE mainly improves visual evidence allocation without substantially disrupting the reasoning and language-side capabilities inherited from the backbone. Overall, RAVE provides the clearest benefits in perception-intensive scenarios, where generation relies on fine-grained visual evidence, while pre-

Variant	Configuration				Benchmarks									
	Location	Form	p	Application	MTVQA	ChartQA	DocVQA	OCRBench	InfoVQA	TextVQA	MathVista	MMMU	MMMU-Pro	Avg.
Main configuration														
(1)	Pre.	Add.	0.25	P+D	9.50	26.28	43.55	44.20	28.21	65.90	25.40	32.22	20.16	32.81
Location variants (Pre./Post.: pre/post softmax)														
(2)	Post.	Add.	0.25	P+D	8.91	24.96	44.68	43.20	28.73	65.61	24.40	32.44	19.54	32.50
Form variants (Add./Mul.: additive/multiplicative)														
(3)	Pre.	Mul.	0.25	P+D	10.05	23.96	44.01	42.40	27.78	65.74	26.00	31.67	18.32	32.21
Head ratio variants (p : partial-head ratio)														
(4)	Pre.	Add.	0.50	P+D	10.25	23.08	42.99	41.30	28.79	65.25	26.20	32.33	19.08	32.14
(5)	Pre.	Add.	0.75	P+D	10.37	25.00	45.21	43.60	28.78	67.13	24.40	29.00	18.61	32.46
(6)	Pre.	Add.	1.00	P+D	10.27	23.52	44.42	44.40	28.06	65.97	23.60	29.78	19.19	32.13
Application variants (P+D/Dec.: prefill+decoding/decoding-only)														
(7)	Pre.	Add.	0.25	Dec.	10.15	25.28	43.36	42.70	27.98	65.28	23.10	30.78	18.79	31.94

Table 3: Ablation study of RAVE design choices across all benchmarks. Variant (1) is the default RAVE setting. Each remaining variant modifies one component of the main configuration while keeping the others unchanged. The modified component is highlighted in bold.

serving competitive performance on broader multimodal reasoning tasks.

The comparisons with GATED-ATTN and VAR further clarify the role of the proposed intervention. Under the same multimodal adaptation protocol, GATED-ATTN underperforms RAVE and even falls below the *Standard Attention* baseline, suggesting that adding a learned gate alone is insufficient to explain the gains. This highlights the importance of where and how the gate is applied within the attention computation. In contrast to VAR’s hand-crafted inference-time redistribution, RAVE employs a learned, content-dependent recalibration mechanism that offers greater flexibility in regulating both the overall attention allocated to visual evidence and its distribution across visual tokens.

6.3 Ablation studies

Table 3 summarizes the ablation studies of RAVE on Mistral-v0.3-7B. We examine four aspects of the design: the location and form of attention recalibration, the GQA-aware head ratio, and the application stage. Each variant changes one component of the default configuration while keeping the remaining settings unchanged.

Recalibration location and form. Pre-softmax recalibration provides a more effective design than post-softmax modulation. By adjusting visual-key logits before normalization, it allows the correction to affect both intra-visual token allocation and cross-modal competition between visual and textual evidence. Additive recalibration is also more stable than its multiplicative counterpart. Although

both forms derive recalibration scores from query-key gating, multiplicative modulation introduces a sign-dependent interaction with the original logits, making negative-logit modulation less predictable.

GQA-aware head ratio. The head-ratio variants reveal a trade-off between corrective capacity and preservation of the pretrained attention prior. The default setting with $p = 0.25$ achieves the best overall performance, while larger gated-head ratios do not yield further gains. Although gating more heads increases the capacity to modify visual attention, it may also disturb the attention patterns inherited from the pretrained language backbone. By leaving part of the query heads ungated within each GQA group, RAVE preserves standard attention pathways for language-side reasoning and autoregressive generation, while using the gated heads to correct visual evidence allocation.

Application stage. Applying RAVE during both prefill and decoding outperforms decoding-only application. This result suggests that visual attention recalibration is useful not only during answer generation but also during the initial construction of image-conditioned representations. By recalibrating question-to-visual attention, the model can form a better visual context before autoregressive decoding begins, which leads to more effective use of visual evidence in subsequent generation.

6.4 Cross-modal attention allocation

To examine whether RAVE changes the cross-modal allocation pattern identified in Section 4, we compare the segment-wise image attention mass of the

Mistral-based *Standard Attention* and RAVE models on the same perception and reasoning probes. As shown in Table 4, RAVE consistently assigns more attention mass to the image segment at both the first and final decoding positions across the two task groups. In particular, it maintains substantially greater visual attention at the final decoding position, indicating that RAVE mitigates cross-modal visual under-allocation late in generation. Full sampling details and statistics are provided in Appendix A.

6.5 Backbone generalization

To examine whether RAVE transfers beyond the controlled LLaVA-style setting, we additionally evaluate it on Qwen2.5-VL-7B-Instruct (Bai et al., 2025b), a multimodal model that has already undergone multimodal pretraining and instruction tuning. Starting from the same released checkpoint, the *Standard Attention* and RAVE variants undergo matched continued supervised fine-tuning. Full settings are provided in Appendix A. As shown in Table 5, RAVE achieves a higher average score than *Standard Attention* across the nine benchmarks, supporting the transferability of the proposed recalibration to an already instruction-tuned LMM. The improvement is more modest than in the Mistral-based setting, possibly because Qwen2.5-VL already has well-established multimodal alignment and attention patterns before RAVE is introduced. Continued fine-tuning may therefore provide less opportunity for the newly inserted gate to co-adapt with and substantially reshape the decoder attention, while the stronger starting model also leaves less headroom for improvement.

7 Conclusion

We presented RAVE, a lightweight attention recalibration mechanism for decoder-only large multimodal models. Motivated by the observation that standard attention can misallocate visual evidence across both modality-level and intra-visual dimensions, RAVE introduces a query-dependent pre-softmax correction to attention logits over visual keys. This enables context-aware visual attention reallocation without modifying the backbone architecture or changing the standard training pipeline. Under matched model and training configurations, RAVE achieves clear improvements on multimodal benchmarks, with the strongest gains on perception-intensive tasks that require fine-grained visual grounding, while maintaining com-

petitive performance on reasoning- and knowledge-heavy benchmarks. Ablation studies show that pre-softmax additive recalibration, GQA-aware partial-head intervention, and prefill+decoding application are all important to the final design. These results suggest that visual attention recalibration is an effective mechanism-level correction for improving visual evidence utilization in decoder-only LMMs.

Limitations

Our experiments focus on instruction-tuned LMMs built from pretrained language backbones. Due to the substantial computational cost of training LMMs from scratch or pretraining native multimodal backbones, we do not evaluate RAVE under such settings. Following the controlled protocol of D-ATTN, we restrict our main controlled comparison to a matched decoder-only LMM setup, which helps isolate the effect of the attention mechanism. Although we additionally validate RAVE on Qwen2.5-VL-7B-Instruct, evaluation across broader model scales and multimodal architectures remains future work. In addition, our current implementation is based on SDPA and has not yet been integrated with FlashAttention-style kernels, which may affect training and inference efficiency.

Ethical Considerations

This work studies attention recalibration for large multimodal models and does not collect, create, or release new datasets containing private, sensitive, or personally identifiable information. It also does not involve human participants or annotators. We use publicly available models, datasets, and benchmarks, with citations to their original creators. The potential risks are similar to those of general LMM systems, including misuse, over-reliance, and insufficiently supervised deployment. Although RAVE improves visual evidence utilization, it does not eliminate hallucination or guarantee reliability in high-stakes settings. We recommend careful evaluation, appropriate safeguards, and human oversight for real-world use.

Acknowledgments

This work was supported by Qwen Business Unit through Alibaba Research Intern Program.

References

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai.

2023. GQA: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhi-fang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. [Qwen3-vl technical report](#). *Preprint*, arXiv:2511.21631.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. [Qwen2.5-vl technical report](#). *Preprint*, arXiv:2502.13923.
- Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024a. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision (ECCV)*.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024b. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer.
- Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 11198–11201.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *Preprint*, arXiv:2310.06825.
- Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. 2025. See what you are told: Visual attention sink in large multimodal models. *arXiv preprint arXiv:2503.03321*.
- Dongwan Kim, Viresh Ranjan, Takashi Nagata, Arnab Dhua, and Amit Kumar K C. 2025. [Rethinking visual information processing in multimodal llms](#). *Preprint*, arXiv:2511.10301.
- Chia-Wen Kuo, Sijie Zhu, Fan Chen, Xiaohui Shen, and Longyin Wen. 2025. Rethinking homogeneity of vision and text tokens in large vision-and-language models (D-Attn). *arXiv preprint arXiv:2502.01906*.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. In *Advances in Neural Information Processing Systems*.
- Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xucheng Yin, Chenglin Liu, Lianwen Jin, and Xiang Bai. 2023b. [OCRBench: On the hidden mystery of OCR in large multimodal models](#). *arXiv preprint arXiv:2305.07895*.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*.
- Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics (ACL)*.
- Minesh Mathew, Viraj Bagal, Ruben Perez Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. 2022. InfographicVQA. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1697–1706.
- Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. DocVQA: A dataset for VQA on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2200–2209.
- Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24.
- Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. In *Advances in Neural Information Processing Systems (NeurIPS) – Best Paper*.
- Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. [Laion-5b: An open large-scale dataset for training next generation image-text models](#). *Preprint*, arXiv:2210.08402.
- Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned,

- hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.
- Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. 2024. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.
- Hai-Long Sun, Zhun Sun, Houwen Peng, and Han-Jia Ye. 2025. Mitigating visual forgetting via take-along visual conditioning for multi-modal long CoT reasoning. *arXiv preprint arXiv:2503.13360*.
- Feilong Tang, Chengzhi Liu, Zhongxing Xu, Ming Hu, Zile Huang, Haochen Xue, Ziyang Chen, Zelin Peng, Zhiwei Yang, Sijin Zhou, Wenxue Li, Yulong Li, Wenxuan Song, Shiyan Su, Wei Feng, Jionglong Su, Minquan Lin, Yifan Peng, Xuelian Cheng, and 2 others. 2025a. Seeing far and clearly: Mitigating hallucinations in MLLMs with attention causal decoding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, An-Lan Wang, Chunhui Lin, Hao Feng, Zhen Zhao, Yanjie Wang, Yuliang Liu, Hao Liu, Xiang Bai, and Can Huang. 2025b. **MTVQA: Benchmarking multilingual text-centric visual question answering**. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7748–7763, Vienna, Austria. Association for Computational Linguistics.
- Chongjun Tu, Peng Ye, Dongzhan Zhou, Lei Bai, Gang Yu, Tao Chen, and Wanli Ouyang. 2025. Attention reallocation: Towards zero-cost and controllable hallucination mitigation of MLLMs. *arXiv preprint arXiv:2503.08342*.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-VL: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Wei-Yao Wang, Zhao Wang, Helen Suzuki, and Yoshiyuki Kobayashi. 2025. Seeing is understanding: Unlocking causal attention into modality-mutual attention for multimodal LLMs. *arXiv preprint arXiv:2503.02597*.
- Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations (ICLR)*.
- Shuo Yang, Yuwei Niu, Yuyang Liu, Yang Ye, Bin Lin, and Li Yuan. 2025. Look-back: Implicit visual re-focusing in MLLM reasoning. *arXiv preprint arXiv:2507.03019*.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. **Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi**. *Preprint*, arXiv:2311.16502.
- Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhao Chen, and Graham Neubig. 2025. **MMMUPRO: A more robust multi-discipline multimodal understanding benchmark**. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15134–15186, Vienna, Austria. Association for Computational Linguistics.
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.
- Xuange Zhang, Dengjie Li, Bo Liu, Zenghao Bao, Yao Zhou, Baisong Yang, Zhongying Liu, Yujie Zhong, and Tongtong Yuan. 2025a. Layer-wise vision injection with disentangled attention for efficient vlms. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7004–7013. IEEE.
- Zhicheng Zhang, Wuyou Xia, Chenxi Zhao, Yan Zhou, Xiaoqiang Liu, Yongjie Zhu, Wenyu Qin, Pengfei Wan, Di Zhang, and Jufeng Yang. 2025b. **MODA: Modular duplex attention for multimodal perception, cognition, and emotion understanding**. *arXiv preprint arXiv:2507.04635*.
- Kai Zhao, Wubang Yuan, Yuchen Lin, Liting Ruan, Xiaofeng Lu, Deng-Ping Fan, Ming-Ming Cheng, and Dan Zeng. 2025. Attention debiasing for token pruning in vision language models. *arXiv preprint arXiv:2508.17807*.
- Guanyu Zhou, Yibo Yan, Xin Zou, Kun Wang, Aiwei Liu, and Xuming Hu. 2025. Mitigating modality prior-induced hallucinations in multimodal large language models via deciphering attention causality. In *International Conference on Learning Representations (ICLR)*.

A Experimental Details

Controlled evaluation protocol. Our experiments comprise two complementary settings: a primary controlled comparison built on Mistral-v0.3-7B (Jiang et al., 2023), and a separate backbone-generalization study on Qwen2.5-VL-7B-Instruct

(Bai et al., 2025b). The primary comparison follows the controlled evaluation protocol of D-ATTN (Kuo et al., 2025). The goal is not to maximize absolute performance through larger-scale data, stronger backbones, or additional alignment stages, but to isolate the effect of the attention mechanism. Accordingly, the trainable Mistral-based variants—*Standard Attention*, D-ATTN, GATED-ATTN, and RAVE—use the same pretrained language backbone, visual encoder, projector, training data, and optimization schedule. The only intended difference among these variants is the attention intervention used inside the decoder. Because VAR is a training-free inference-time method, we apply it post hoc to the same trained *Standard Attention* checkpoint rather than training a separate model.

Scope of comparison. Table 1 organizes prior approaches to visual-attention allocation into three broad families: training-free inference-time intervention, architectural redesign, and training-based visual revisiting. Our numerical comparison includes representative methods from the first two families that can be evaluated under matched or maximally comparable conditions. For training-free intervention, we select VAR because it directly targets visual attention sinks, a central form of intra-visual misallocation considered in this work. For architectural redesign, we select D-ATTN because it provides the closest architectural counterpart: it modifies visual-text interaction by decomposing attention into modality-specific paths, whereas RAVE retains native joint attention and recalibrates the pre-softmax logits over visual keys. We additionally include GATED-ATTN as a mechanism-level comparator because, like RAVE, it introduces learned gating within the attention operator, while applying the gate to the attention output rather than to the attention allocation. We do not include training-based revisiting methods such as TVC and LOOK-BACK as numerical baselines because their original pipelines rely on curated long-chain reasoning data and specialized visual-revisiting generation procedures. LOOK-BACK further uses supervised fine-tuning, reinforcement learning, and a dedicated visual-refocusing token. Including these methods would therefore alter both the training data and the decoding protocol, preventing a controlled comparison focused on the attention mechanism.

Mistral-based model construction and baselines. Following D-ATTN, we adopt a LLaVA-

style decoder-only LMM architecture consisting of a pretrained SigLIP (Zhai et al., 2023) visual encoder, a two-layer MLP projector with RMSNorm for visual-language alignment, and a pretrained Mistral-v0.3-7B language backbone. Under this architecture, *Standard Attention* retains the original causal self-attention mechanism. D-ATTN decomposes causal attention into visual-to-visual, textual-to-visual, and textual-to-textual paths. For GATED-ATTN (Qiu et al., 2025), we follow the original paper’s best-performing *O*-gating configuration, which applies a head-specific sigmoid gate to the attention output. We use its reported hyperparameters where applicable, without method-specific tuning on our evaluation benchmarks. RAVE retains the native joint-attention formulation while adding the proposed query-key pair gate to the pre-softmax logits associated with visual keys. For the training-free comparison, we apply VAR (Kang et al., 2025) to the trained *Standard Attention* model. Specifically, we re-identify the sink dimensions on our Mistral-based checkpoint following the procedure in the original work, and then apply its attention-redistribution rule during inference using the default hyperparameters.

Training recipe. The trainable Mistral-based variants follow the three-stage training recipe of D-ATTN where applicable. In the first stage, the visual-language projector is trained for modality alignment using the LLaVA LAION/CC/SBU (Liu et al., 2023a; Ordonez et al., 2011; Schuhmann et al., 2022; Sharma et al., 2018) 558K subset. In the second stage, the model is fine-tuned on 1.25M dense captions from ShareGPT4V-PT (Chen et al., 2024b). In the third stage, instruction tuning is performed using a combination of 665K LLaVA-1.5 instruction-tuning samples (Liu et al., 2023a) and 102K dense-caption samples from ShareGPT4V (Chen et al., 2024b). All trainable variants use the standard next-token cross-entropy objective. For GATED-ATTN and RAVE, the added gate parameters are optimized jointly with the LMM during the stages in which the decoder is updated. RAVE requires neither auxiliary supervision nor an additional gating-specific loss. VAR requires no additional training. For the Qwen2.5-VL experiment, both matched variants undergo continued supervised fine-tuning using the datasets from all three stages described above, starting from the same publicly released instruction-tuned checkpoint.

Group	Position	Image Mass (%)			Answer Mass (%)	
		Std.	RAVE	Δ	Std.	RAVE
Perception	First	24.8	39.6	+14.8	4.8	4.1
	Final	20.9	34.7	+13.8	16.0	15.5
Reasoning	First	12.8	26.1	+13.3	5.0	4.3
	Final	12.3	25.3	+13.0	7.2	6.1

Table 4: Segment-wise attention mass (%) for *Standard Attention* and RAVE at the first and final decoding positions. Values are averaged over all decoder layers, query heads, and 20 probes per task group. Δ denotes the image-segment attention mass of RAVE minus that of *Standard Attention*.

Evaluation protocol. The original D-ATTN study evaluates ten image benchmarks under the LLaVA evaluation protocol. In this work, we instead use VLMEvalKit (Duan et al., 2024) to evaluate a benchmark suite covering both perception-intensive and reasoning-oriented capabilities. The perception-intensive group includes MTVQA-TEST, ChartQA-TEST, DocVQA-VAL, OCRBench, InfoVQA-VAL, and TextVQA-VAL. The reasoning-oriented group includes MathVista-MINI, MMMU-DEV-VAL, and MMMU-Pro-10c. For each benchmark, we report the score returned by VLMEvalKit, together with *Perception Avg.*, *Reasoning Avg.*, and *Overall Avg.* over the corresponding benchmark groups. The same benchmark definitions and evaluation settings are used for both the Mistral-based comparison and the Qwen2.5-VL backbone study.

Cross-modal attention-mass analysis. To examine whether RAVE changes the cross-modal allocation pattern identified in Section 4, we compare the Mistral-based *Standard Attention* and RAVE models on the same perception and reasoning probes. For the perception group, we randomly sample 10 examples each from OCRBench and DocVQA; for the reasoning group, we randomly sample 10 examples each from MathVista and MMMU, yielding 20 probes per group. Both models use greedy decoding with identical images, image preprocessing, and prompt templates. Following Eq. 3, we compute the image- and answer-segment attention masses at each generated position and average them over decoder layers, query heads, and probe examples. Table 4 reports the first and final decoding positions, for which all 20 probes in each group contribute. Across both task groups, RAVE assigns substantially more attention to the image segment,

including at the final decoding position. These results indicate that RAVE mitigates cross-modal visual under-allocation late in generation.

Backbone-generalization setting. To assess whether RAVE transfers beyond the controlled LLaVA-style architecture, we additionally use the publicly released Qwen2.5-VL-7B-Instruct checkpoint (Bai et al., 2025b). Unlike the Mistral-based models constructed from a pretrained language backbone through our multimodal training pipeline, Qwen2.5-VL-7B-Instruct is already an instruction-tuned multimodal model with its own vision tower and M-RoPE-based positional treatment. We therefore treat this experiment as continued supervised fine-tuning from an existing instruction-tuned checkpoint, rather than as a reproduction of Qwen2.5-VL’s original multimodal pre-training and instruction-tuning pipeline. Starting from the same released checkpoint, the *Standard Attention* and RAVE variants undergo continued supervised fine-tuning using the data from all three stages of our controlled recipe: the 558K LLaVA LAION/CC/SBU subset used in Stage 1, the 1.25M ShareGPT4V-PT dense-caption set used in Stage 2, and the instruction-tuning mixture used in Stage 3. The two variants use the same data sequence, training schedule, and optimization settings; the only difference is whether RAVE is inserted into the language decoder.

Computational overhead. In the matched Qwen2.5-VL continued-SFT experiment under the same compute configuration, *Standard Attention* and RAVE require approximately 20 and 22 hours, respectively, for the same training schedule, corresponding to approximately 10% additional training time. Under the same compute configuration, batch settings, and maximum generation lengths in VLMEvalKit, evaluation on the six perception-intensive benchmarks takes approximately 12 hours for *Standard Attention* and 14 hours for RAVE, corresponding to approximately 17% end-to-end inference overhead. These measurements reflect our current non-fused SDPA implementation, which explicitly materializes the pairwise gate tensor, and therefore provide a conservative estimate of the deployment overhead. Because each gate entry is generated from two per-token scalar vectors, a fused implementation could compute the correction on the fly without materializing an additional $N \times N$ gate tensor.

Attention Method	Perception-intensive						Reasoning / Knowledge-heavy			Perc. Avg.	Reason. Avg.	Overall Avg.
	MTVQA	ChartQA	DocVQA	OCRBench	InfoVQA	TextVQA	MathVista	MMMU	MMMU-Pro			
<i>Standard Attention</i>	22.03	80.04	92.33	82.50	74.76	82.38	58.20	48.11	35.18	72.34	47.16	63.95
RAVE	22.19	82.44	92.85	84.40	76.10	82.96	58.50	49.22	36.06	73.49	47.93	64.97

Table 5: Backbone-generalization results on Qwen2.5-VL-7B-Instruct. Both variants start from the same publicly released instruction-tuned checkpoint and undergo matched continued supervised fine-tuning using the datasets from all three stages of our controlled recipe; they differ only in whether RAVE is inserted into the language decoder.

Baseline implementation and reproducibility.

For D-ATTN, we use the official implementation and reported hyperparameter settings without additional tuning. For GATED-ATTN, we reproduce the original paper’s best-performing output-side O -gating configuration. We note that the original GATED-ATTN study introduces its gate during language-model pretraining, whereas our controlled setup introduces it during multimodal adaptation, consistent with the other trainable Mistral-based variants. Thus, this comparison evaluates GATED-ATTN as a drop-in intervention under a matched multimodal training budget rather than reproducing its full language-model pretraining set-

ting. For VAR, we follow the original inference-time procedure and default hyperparameters, while identifying the sink dimensions on our own trained backbone. For the Qwen2.5-VL study, the two variants share the same initial checkpoint, training data, and optimization schedule, and differ only in whether RAVE is inserted into the language decoder. We use publicly released models, datasets, and benchmarks for research purposes and follow their released licenses or terms of use. Our implementation is based on PyTorch SDPA, DeepSpeed, and VLMEvalKit. Package versions and scripts will be released with the code.