Focusing by Contrastive Attention: Enhancing VLMs' Visual Reasoning

Yuyao Ge[1,2,3], Shenghua Liu[1,2,3⋆], Yiwei Wang[4], Lingrui Mei[1,2,3], Baolong Bi[1,2,3], Xuanshan Zhou[1,2,3], Jiayu Yao[1,2,3], Jiafeng Guo[1,2,3], and Xueqi Cheng[1,2,3]

[1]State Key Laboratory of AI Safety, China
[2]Institute of Computing Technology, Chinese Academy of Sciences, China
[3]University of Chinese Academy of Sciences, China
[4]University of California, Merced, USA
{geyuyao24z, liushenghua}@ict.ac.cn

Abstract. Vision-Language Models (VLMs) have demonstrated remarkable success across diverse visual tasks, yet their performance degrades in complex visual environments. Existing enhancement approaches require additional training, rely on external segmentation tools, or operate at coarse-grained levels, overlooking VLMs' innate attention capabilities. To bridge this gap, we investigate VLMs' attention patterns and discover that: (1) visual complexity strongly correlates with attention entropy, negatively impacting reasoning performance; (2) attention progressively refines from global scanning in shallow layers to focused convergence in deeper layers, with the degree of convergence determined by visual complexity; (3) theoretically, under a multiplicative decomposition assumption, we show that contrasting attention maps between general and task-specific queries approximately separates visual signal into semantic and visual noise components. Building on these insights, we propose **C**ontrastive **A**ttention **R**efinement for **V**isual **E**nhancement (**CARVE**), a training-free method that extracts task-relevant visual signals through attention contrasting at the pixel level. Experiments on seven benchmarks show that CARVE improves performance on visual perception tasks, with notable gains on both recent and earlier-generation models. Our analysis reveals how visual complexity affects attention mechanisms and demonstrates an effective strategy for improving visual reasoning through attention contrasting.

Keywords: Vision-Language Models · Attention Mechanism · Visual Reasoning

1 Introduction

Vision-Language Models (VLMs) have shown impressive capabilities across diverse tasks [2, 10, 22, 28]. However, in human vision, complex visual features

⋆ Corresponding author.

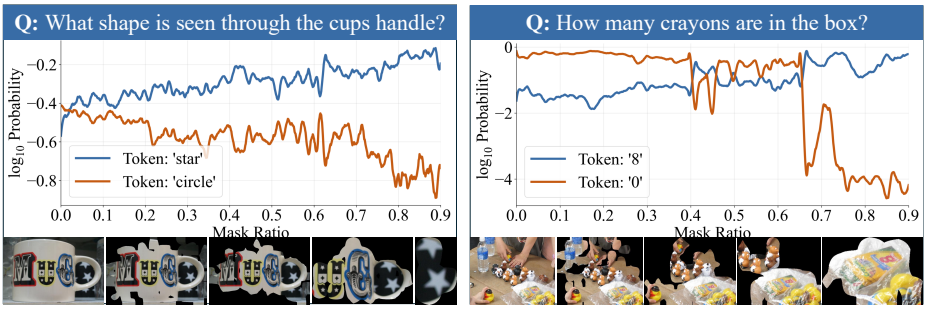


Fig. 1: The effect of manually progressive masking on candidate tokens probabilities predicted by QWEN2.5-VL-3B. The x-axis represents mask ratio and y-axis shows $\log_{10}$ probability.

frequently divert attention from task-relevant regions [29]. Given this cognitive parallel, a question arises: *Similarly, do complex images interfere with VLMs' attention mechanisms, making it difficult for them to focus on task-relevant regions?*

To answer this question, we investigate the relationship between visual complexity and attention patterns via quantitative experiments. Specifically, we define visual complexity along texture and color dimensions, revealing a significant positive correlation between both factors and attention entropy. Our analysis also shows that attention entropy negatively correlates with accuracy on visual reasoning tasks. Through this two-stage analysis, we find that **complex visual information impairs VLMs' reasoning performance via attention dispersion** (Sect. 3).

Based on these findings, and inspired by ViCrop [34], we conduct a preliminary experiment on TextVQA [27] by progressively masking background regions, then cropping and magnifying task-relevant areas. Fig. 1 presents two representative samples where cluttered environments initially cause incorrect predictions. As more background is masked, correct token probability surpasses incorrect probability, validating that **masking visual noise can improve correct token probability**.

To automate the visual noise masking process, we exploit contrasting attention maps between general instructions and task-specific questions to distinguish semantic signal from visual noise. To this end, we propose **C**ontrastive **A**ttention **R**efinement for **V**isual **E**nhancement (**CARVE**), a training-free contrastive method for visual extraction. By masking with contrastive attention maps, CARVE crops and magnifies semantic regions to focus on essential semantic signal (Sect. 4).

2 Related Work

Contrastive Learning in LLMs. Liu *et al.* [15] pioneered contrastive objectives for distilling supervision from LLMs. To address hallucination, Jiang *et al.* [11] employ hallucinated text as hard negatives while Zhang *et al.* [35] apply contrastive learning in hidden representations to suppress hallucinations. Zhai *et al.* [33] trace critical transmission paths across all layers, treating less important pathways as negatives, which Pan *et al.* [20] further adapted to multimodal LLMs through UniKE's semantic-truthfulness space disentanglement. Departing from embedding-space modifications, DeCK [4] shifts contrastive logic to the decoding stage by comparing token probabilities with and without injected knowledge.

Attention-based LLM Optimization. Training-free approaches prune redundant visual tokens [1, 6, 7, 17] or broadcast vision-focused attention heads to sibling heads [36] to accelerate inference and reduce hallucinations. Training-based methods ground and zoom into task-relevant regions: CogCoM [21] trains a chain-of-manipulation pipeline for step-by-step visual reasoning with grounding and zooming, CogAgent [9] employs dual encoders for high-resolution understanding, Ferret-UI [32] splits screens into sub-images for fine-grained grounding, and Gkartzonika *et al.* [8] train attention mechanisms on frozen classifiers to produce visual explanation maps. Complementary analyses reveal that models know where to look even when answering incorrectly [34] and that position bias affects multimodal retrieval [31]. Unlike these methods, our approach contrasts attention maps to distinguish semantic pixels from noise at the pixel level, requiring no training.

3 Failure to Focus: Phenomenon, Mechanism, and Consequence

3.1 Phenomenon: Under What Conditions Visual Focus Fails

We visualize attention maps of Qwen2.5-VL-3B-Instruct on TextVQA (Fig. 2). Attention progressively refines from global scanning in shallow layers to focused convergence in deep layers, with convergence varying across images.

Visual complexity also influences attention convergence. In simple scenes with clear targets and minimal distractors, attention successfully narrows to task-relevant regions as layers deepen (consistent with ViCrop [34]). Conversely, in complex scenes with rich textures and colors, attention remains more dispersed even in deep layers. As indicated by the annotation "Confused where to look" in Fig. 2, this attention dispersion leads to reasoning failures.

These observations motivate two quantitative investigations: Sect. 3.2 examines whether visual complexity disperses attention, and Sect. 3.3 links attention dispersion to performance degradation.

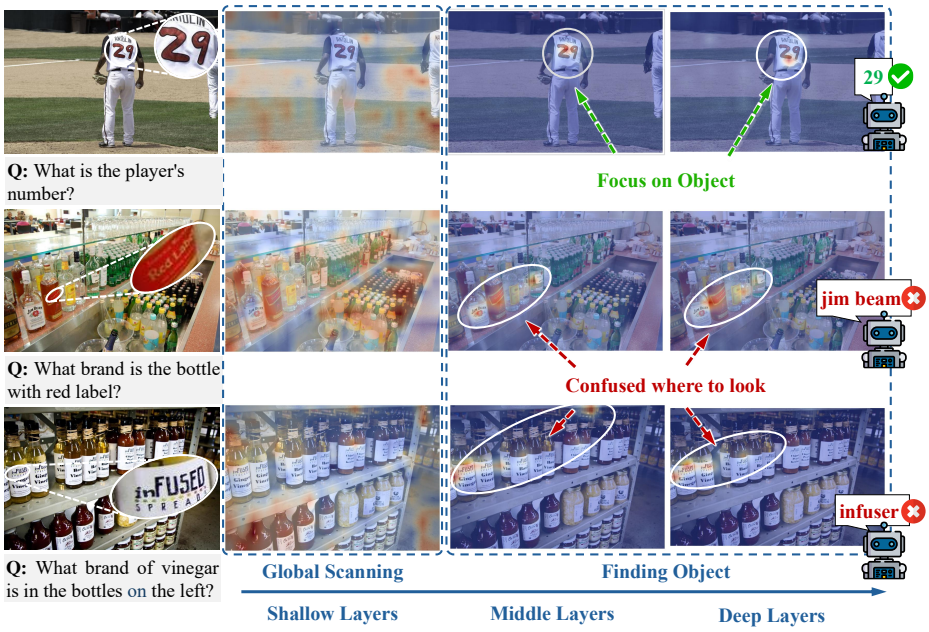


Fig. 2: Attention maps across different layers during inference. Each row represents a visual question-answering task: row 1 shows a simple scene with clear targets, while rows 2-3 depict complex scenes with dense textures and multiple similar objects. From left to right, the columns display: input images, shallow layer attention, middle layer attention, and deep layer attention.

3.2 Mechanism: Visual Complexity and Attention Entropy

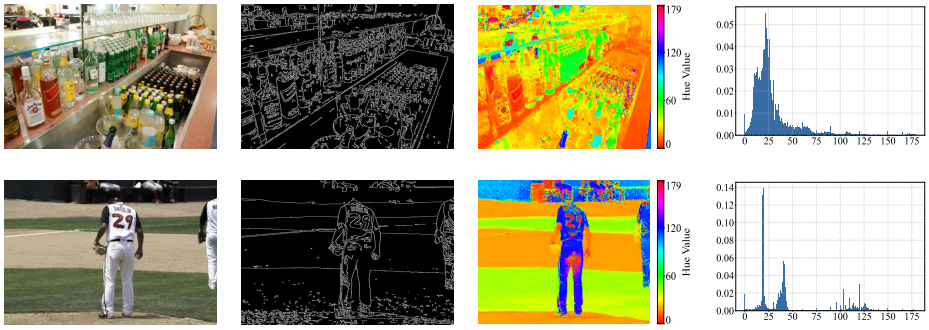
In Fig. 2, images in rows 2-3 differ from row 1 by displaying numerous colorful bottles, containing significantly more textures and colors. Therefore, we decompose visual complexity into two dimensions: **texture** and **color**, and investigate their respective impacts on attention.

We define the texture complexity and the color complexity as follows:

Texture Complexity. Let $\mathcal{I} \in \mathbb{R}^{H\times W\times 3}$ denote an input image. We define the texture complexity $\mathcal{T}_c(\mathcal{I})$ using Canny edge detection [5], where $\mathcal{E}(\mathcal{I}) \in \{0,1\}^{H\times W}$ represents the resulting binary edge map. The texture complexity is then defined as:

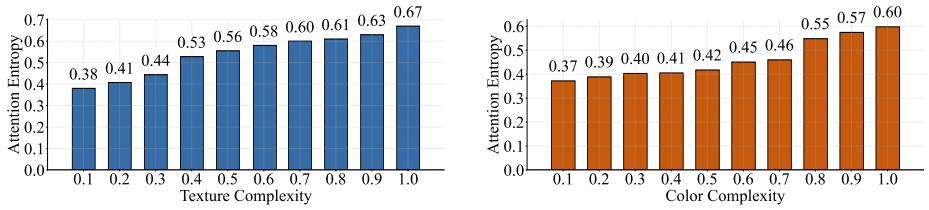
$$\mathcal{T}_c(\mathcal{I}) = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\mathcal{E}(\mathcal{I})_{ij} = \frac{\|\mathcal{E}(\mathcal{I})\|_1}{HW} \in [0,1] \tag{1}$$

Color Complexity. Let $\zeta_{ij} = \text{Hue}(\Psi_{RGB\to HSV}(\mathcal{I}_{ij}))$ denote the hue value at pixel (i,j) after applying the RGB to HSV transformation operator Ψ. The



(a) Original **(b)** Edge detection **(c)** Hue distribution **(d)** Hue histogram

Fig. 3: Visualization of texture and color complexity analysis. Each row represents a sample image: (a) Original image $\mathcal{I}$, (b) Canny edge map $\mathcal{E}(\mathcal{I})$ for texture complexity $\mathcal{T}_c$, (c) Spatial hue distribution ζ in HSV space, and (d) Hue statistic (x-axis: hue value, y-axis: ratio).



(a) Texture-entropy correlation **(b)** Color-entropy correlation

Fig. 4: Correlation analysis between visual complexity and attention entropy. Both attention entropy and complexity are normalized to the $[0, 1]$, divided into intervals of 0.1, and the average attention entropy is calculated within each interval.

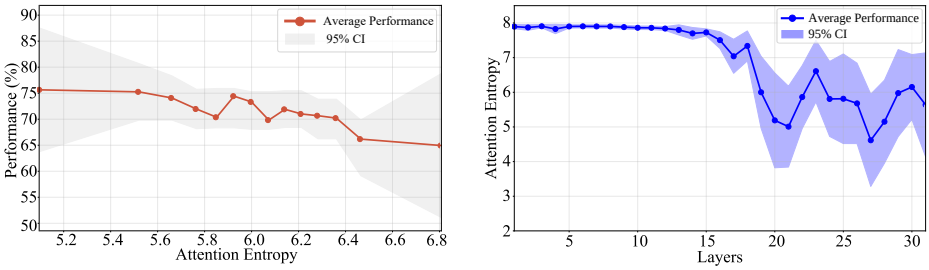
color complexity is then defined as:

$$\mathcal{C}_c(\mathcal{I}) = -\frac{1}{\ln B} \sum_{b=0}^{B-1} \rho_b \ln \rho_b, \quad \text{where } \rho_b = \frac{n_b}{HW}, \quad n_b = |\{(i,j) : \zeta_{ij} = b\}| \quad (2)$$

with $B = 180$ hue bins, yielding $\mathcal{C}_c \in [0, 1]$ where higher values indicate greater color diversity.

Fig. 3 validates our complexity metrics. The first row shows high texture complexity with dense edge networks in $\mathcal{E}(\mathcal{I})$ and diverse color distribution across the hue spectrum. In contrast, the second row exhibits minimal edge density and concentrated hue values, indicating lower complexity scores. We next quantitatively investigate the correlation between complexity metrics and attention distribution.

Inspired by Yao *et al.* [31], we employ Shannon entropy [25] to quantify attention distribution across visual tokens. Let N_v denote the number of visual tokens in the model's representation. We denote the attention map as $A_{l,t}^{(Q)} \in \mathbb{R}^{N_v}$, where l indicates the layer index, t the generation time step, and Q the



(a) Entropy-accuracy correlation **(b)** Layer-wise entropy correlation

Fig. 5: Attention entropy's correlation with accuracy and its evolution across layers. Shaded regions indicate 95% confidence intervals computed as $\bar{x} \pm t_{0.975,n-1} \cdot s/\sqrt{n}$. (a) Shows accuracy for samples grouped by overall attention entropy $\overline{\mathcal{H}}$. (b) Displays mean entropy across N samples per layer as $\frac{1}{N}\sum_{i=1}^{N} \mathcal{H}(A_{l,t_{\text{end}}}^{(Q,i)})$, where $A_{l,t_{\text{end}}}^{(Q,i)}$ is sample i's attention at the final generation step.

input question. For entropy analysis, we focus on the final generation step t_{end} and define the overall attention entropy $\overline{\mathcal{H}}$ as:

$$\overline{\mathcal{H}} = \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} \mathcal{H}(A_{l,t_{\text{end}}}^{(Q)}) = \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} \left(-\sum_{i=1}^{N_v} a_{l,t_{\text{end}},i} \ln a_{l,t_{\text{end}},i} \right) \tag{3}$$

where $\mathcal{L} = [L_{\text{start}}, L_{\text{end}}]$ represents the layer range under consideration, and $a_{l,t,i}$ denotes the attention weight for the i-th visual token. Higher entropy indicates more dispersed attention, while lower entropy indicates more concentrated focus.

Fig. 4 presents the correlation between visual complexity and attention entropy. Both texture complexity (Fig. 4a) and color complexity (Fig. 4b) exhibit strong positive linear relationships with attention entropy. This monotonic trend indicates that **complex visual features lead to dispersed attention patterns in VLMs**.

3.3 Consequence: How Dispersed Attention Impairs Performance

Fig. 5(a) reveals a strong negative correlation: as attention entropy increases from 5.1 to 6.8, accuracy drops from approximately 76% to 65%, suggesting that higher attention dispersion is associated with lower visual reasoning accuracy.

To investigate the hierarchical evolution of attention entropy, we present mean entropy and its distribution across layers in Fig. 5(b). The results show two characteristics: (1) attention entropy monotonically decreases with layer depth, consistent with Fig. 2. (2) The 95% confidence intervals progressively widen with increasing depth, indicating enhanced inter-sample variability. For samples with clear visual targets, deep layers achieve highly concentrated attention. In contrast, for noisy samples, the model maintains dispersed attention patterns even in deep layers.

4 Contrastive Attention Refinement for Visual Enhancement

4.1 Theoretical Foundation: Attention Decomposition

Given that visual complexity causes attention dispersion and performance degradation (Sect. 3), we introduce an attention decomposition to extract pure semantic signal.

Assumption 1 (Attention Decomposition): Attention distributions are influenced by inherent visual noise of the image and task-related semantic signal. We model the attention map $A_{l,t}^{(Q)}(\mathcal{I})$ as:

$$A_{l,t}^{(Q)}(\mathcal{I}) = \mathcal{F}_{\text{vis}}(\mathcal{I}) \otimes \mathcal{F}_{\text{sem}}(Q, \mathcal{I}) \tag{4}$$

where $\mathcal{F}_{\text{vis}}(\mathcal{I}) \in \mathbb{R}^{N_v}$ captures image-inherent visual noise, $\mathcal{F}_{\text{sem}}(Q, \mathcal{I}) \in \mathbb{R}^{N_v}$ captures task-related semantic signal, and $\otimes$ denotes the Hadamard product. We adopt multiplicative rather than additive decomposition because visual noise acts as a spatially varying gain on attention weights; element-wise division normalizes heterogeneous noise magnitudes uniformly, whereas subtraction leaves scale-dependent residuals. Empirically, replacing the multiplicative contrast with an additive alternative $\hat{A}_i = A_i^{(Q)} - A_i^{(G)}$ reduces the relative gain on TextVQA from +9.33% to +4.53%, supporting the benefit of scale-invariant normalization.

The general instruction G is a deliberate design choice rather than an arbitrary one: it must elicit global image scanning without focusing on any specific semantic region. Among five candidate prompts, we adopt "Write a general description of the image." as it attains the highest accuracy and the lowest variance while keeping generation short. When using such general instructions G, lacking task-specific semantic cues, the semantic signal function reduces to an approximately uniform distribution ($\mathcal{F}_{\text{sem}}(G, \mathcal{I}) \approx \mathbf{1}_{N_v}$), as illustrated by the diffuse attention pattern in Stage 1 of Fig. 6. Quantitatively, the KL divergence from uniform is $\text{KL}(A^{(G)} \| \mathcal{U}) = 0.147$, far below the task-specific counterpart $\text{KL}(A^{(Q)} \| \mathcal{U}) = 0.812$, confirming that general-instruction attention closely approximates uniformity. General instruction attention therefore predominantly captures visual noise:

$$A_{l,t}^{(G)}(\mathcal{I}) \approx \mathcal{F}_{\text{vis}}(\mathcal{I}) \otimes \mathbf{1}_{N_v} = \mathcal{F}_{\text{vis}}(\mathcal{I}) \tag{5}$$

Definition 2 (Semantic Extraction Based on Attention Decomposition): We estimate $\mathcal{F}_{\text{sem}}(Q, \mathcal{I})$ by defining $\hat{A} \in \mathbb{R}_+^{N_v}$ as the solution to:

$$\hat{A} = \arg\min_{\tilde{A} \in \mathcal{A}} \mathcal{J}(\tilde{A}; A^{(Q)}, A^{(G)}) \tag{6}$$

where the objective function is constructed based on Assumption 1's decomposition:

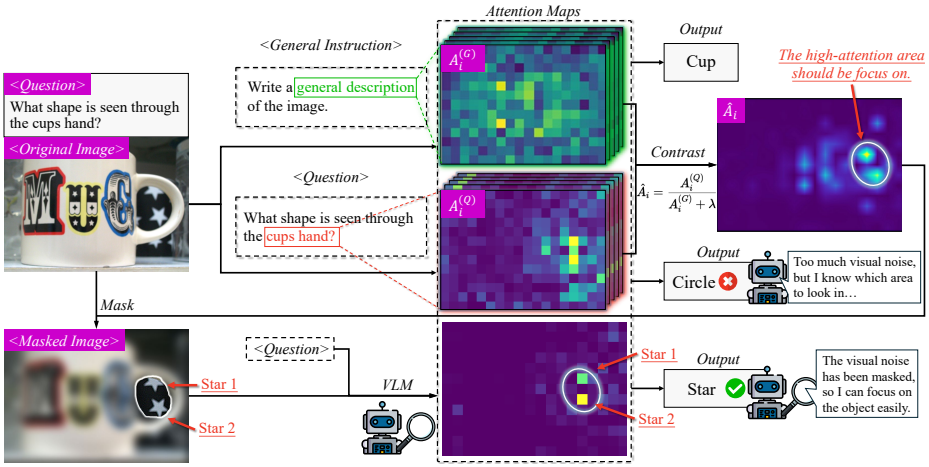


Fig. 6: CARVE comprises three stages: Stage 1 generates general attention $A_i^{(G)}$ via task-agnostic instructions; Stage 2 extracts task-specific attention $A_i^{(Q)}$; Stage 3 applies contrasted attention $\hat{A}_i$ to generate enhanced masked images for noise suppression.

$$\mathcal{J}(\tilde{A}) = \underbrace{\sum_{i=1}^{N_v} \left(\tilde{A}_i \cdot \mathcal{F}_{\text{vis},i}(\mathcal{I}) - [\mathcal{F}_{\text{vis},i}(\mathcal{I}) \cdot \mathcal{F}_{\text{sem},i}(Q,\mathcal{I})] \right)^2}_{\text{Semantic reconstruction error}} + \underbrace{\lambda \sum_{i=1}^{N_v} \tilde{A}_i^2 \cdot \mathcal{F}_{\text{vis},i}(\mathcal{I})}_{\text{Visual suppression regularization}} \tag{7}$$

Proposition 1 (Closed-form Solution for Semantic Extraction): Substituting Assumption 1's relationships $A_i^{(Q)} \approx \mathcal{F}_{\text{vis},i} \cdot \mathcal{F}_{\text{sem},i}$ and $A_i^{(G)} \approx \mathcal{F}_{\text{vis},i}$ into the optimization objective yields:

$$\mathcal{J}(\tilde{A}) = \sum_{i=1}^{N_v} \left(\tilde{A}_i \cdot A_i^{(G)} - A_i^{(Q)} \right)^2 + \lambda \sum_{i=1}^{N_v} \tilde{A}_i^2 \cdot A_i^{(G)} \tag{8}$$

where $\lambda > 0$ is a regularization parameter that controls the strength of visual noise suppression.

Solving the first-order optimality conditions yields the closed-form solution:

$$\hat{A}_i = \frac{A_i^{(Q)}}{A_i^{(G)} + \lambda} = \frac{\mathcal{F}_{\text{vis},i} \cdot \mathcal{F}_{\text{sem},i}}{\mathcal{F}_{\text{vis},i} + \lambda} \approx \mathcal{F}_{\text{sem},i} \quad \text{when } \mathcal{F}_{\text{vis},i} \gg \lambda \tag{9}$$

Eq. 9 demonstrates that normalization suppresses the influence of $\mathcal{F}_{\text{vis},i}$ when it dominates (*i.e.*, $\mathcal{F}_{\text{vis},i} \gg \lambda$), approximating the semantic signal $\mathcal{F}_{\text{sem},i}$.

4.2 Contrastive Attention-based Visual Enhancement

Given the refined attention maps $\{\hat{A}\}$, as shown in Algorithm 1, we generate attention masks that physically remove visual noise from the input image.

Algorithm 1 CARVE: Contrastive Attention Refinement for Visual Enhancement

Notation: $\mathcal{M}$: VLM model; Ξ: attention extraction; $\pi_{H\times W}$: spatial reshape;$\mathcal{Q}_p$: top-p threshold; Φ: visual extraction (mask, crop, resize); G: general instruction; τ: threshold; $\mathcal{R}$: connected regions; K: max regions to keep

Require: $\mathcal{I} \in \mathbb{R}^{H\times W\times 3}$, Q, $\mathcal{M}$, $\Theta = \{\mathcal{L}, \mathcal{T}, p, \lambda, K\}$

Inference: $\mathcal{A}^Q \leftarrow \{A^{(Q)}_{l,t}\}_{l\in\mathcal{L},t\in\mathcal{T}} = \Xi(\mathcal{M}, \mathcal{I}, Q)$ ▷ Question-specific attention

Inference: $\mathcal{A}^G \leftarrow \{A^{(G)}_{l,t}\}_{l\in\mathcal{L},t\in\mathcal{T}} = \Xi(\mathcal{M}, \mathcal{I}, G)$ ▷ General attention

Contrast: $\hat{A}_{l,t} \leftarrow \frac{A^{(Q)}_{l,t}}{A^{(G)}_{l,t}+\lambda}$ for all $l \in \mathcal{L}, t \in \mathcal{T}$ ▷ following Eq. 9

Fuse: $S \leftarrow \sum_{t\in\mathcal{T}} w_t \sum_{l\in\mathcal{L}} \pi_{H\times W}(\hat{A}_{l,t})$, $w_t = t - t_{\text{start}} + 1$ ▷ Weighted attention fusion

Threshold: $\tau \leftarrow \mathcal{Q}_p(S)$ ▷ Compute threshold to retain top p percentile

Mask: $M^* = \bigcup_{k=1}^{K} R^*_k$ where $R^*_k = \arg\max_{R\in\mathcal{R}} \sum_{(i,j)\in R} S(i,j)$ with $\mathcal{R}$ from $S \geq \tau$

Extract: $\mathcal{I}_{\text{refined}} \leftarrow \Phi(\mathcal{I}, M^*)$ ▷ Visual extraction

Inference: return $\mathcal{M}(\mathcal{I}_{\text{refined}}, Q)$ ▷ Final inference

Attention Maps Fusion. Since different layers and time steps capture complementary information, we fuse attention maps across the layer range $\mathcal{L}$ and generation time steps $\mathcal{T} = [t_{\text{start}}, t_{\text{end}}]$ through weighted aggregation. Later tokens receive higher fusion weights, as they encode richer context from longer preceding sequences.

Mask Generation and Visual Extraction. Task-relevant regions are identified by applying the top-p percentile threshold $\tau = \mathcal{Q}_p(S)$, which retains the top $p \in (0, 1]$ proportion of pixels from attention map S. Connected component analysis extracts coherent regions from the thresholded map. We select the top-K regions ranked by cumulative attention scores and generate the enhanced image through $\mathcal{I}_{\text{refined}} = \Phi(\mathcal{I}, M^*)$, where Φ applies masking, cropping, and resizing, and K controls the maximum number of regions to preserve.

As shown in Fig. 6, we propose **C**ontrastive **A**ttention **R**efinement for **V**isual **E**nhancement (**CARVE**), a method that contrasts attention maps to distinguish semantic pixels from noise, preserving only task-relevant regions for enhanced model focus.

5 Experiments

5.1 Experimental Setup

Datasets. We conduct experiments on seven datasets spanning diverse task dimensions. For visual reasoning and understanding, we use A-OKVQA [24], POPE [13], V* [30], and TextVQA [27]. To further evaluate generalizability, we extend evaluation to DocVQA [19] for document understanding (Exact Match and ANLS), ChartQA [18] for chart comprehension (Exact Match and Relaxed Accuracy), and ScienceQA [16] for knowledge-based reasoning (multiple-choice

Table 1: Accuracy of CARVE across VLMs on four datasets. We evaluate three temporal configurations: t_{start} uses attention from initial generated tokens, t_{end} from final tokens, and $\mathcal{T}_{\text{full}}$ applies weighted fusion across all tokens. We use layer range $\mathcal{L} = [20, 25]$ for attention fusion. **Bold**: best; underline: second-best.

Model	Step: $\mathcal{T}$	A-OKVQA	POPE	V*	TextVQA
Qwen2.5-VL-3B	Vanilla	73.0	86.9	50.3	72.8
	t_{start}	76.5 ▲ 4.79	87.1 ▲ 0.23	56.0 ▲ 11.33	76.1 ▲ 4.53
	t_{end}	**79.2** ▲ 8.49	**88.4** ▲ 1.73	**57.1** ▲ 13.52	**76.4** ▲ 4.95
	$\mathcal{T}_{\text{full}}$	78.3 ▲ 7.26	87.9 ▲ 1.15	56.5 ▲ 12.33	76.3 ▲ 4.81
Qwen2.5-VL-7B	Vanilla	75.0	87.0	50.8	75.0
	t_{start}	77.0 ▲ 2.67	87.9 ▲ 1.03	58.6 ▲ 15.35	80.7 ▲ 7.60
	t_{end}	**78.3** ▲ 4.40	**89.7** ▲ 3.10	**59.7** ▲ 17.52	**81.9** ▲ 9.20
	$\mathcal{T}_{\text{full}}$	78.0 ▲ 4.00	88.6 ▲ 1.84	58.1 ▲ 14.37	81.7 ▲ 8.93
LLaVA-1.5-7B	Vanilla	71.5	83.6	38.7	47.8
	t_{start}	73.9 ▲ 3.36	86.8 ▲ 3.83	57.1 ▲ 47.55	57.9 ▲ 21.13
	t_{end}	**78.2** ▲ 9.37	**89.0** ▲ 6.46	**66.5** ▲ 71.83	**58.2** ▲ 21.76
	$\mathcal{T}_{\text{full}}$	75.4 ▲ 5.45	**89.0** ▲ 6.46	**66.5** ▲ 71.83	57.9 ▲ 21.13
LLaVA-1.5-13B	Vanilla	75.7	84.6	42.4	57.1
	t_{start}	76.2 ▲ 0.66	90.0 ▲ 6.38	65.4 ▲ 54.25	59.2 ▲ 3.68
	t_{end}	**76.9** ▲ 1.59	**90.7** ▲ 7.21	**74.3** ▲ 75.24	**61.2** ▲ 7.18
	$\mathcal{T}_{\text{full}}$	**76.9** ▲ 1.59	90.1 ▲ 6.50	70.2 ▲ 65.57	**61.2** ▲ 7.18

accuracy). For TextVQA, we provide only images and questions without external OCR augmentation, following the OCR-free protocol of ViCrop [34].

Models. We conduct experiments on four VLMs: Qwen2.5-VL-3B-Instruct, Qwen2.5-VL-7B-Instruct [3], LLaVA-1.5-7B, and LLaVA-1.5-13B [14]. The Qwen family processes images at 448×448 resolution, while the LLaVA-1.5 family operates at 336×336 resolution. All models employ greedy decoding with regularization parameter $\lambda = 0.05$.

Baselines. We compare CARVE against two concurrent methods that also enhance VLM visual perception: ViCrop [34] and SD-RPN [26].

5.2 Results

Main Results. Tables 1 and 2 show that CARVE improves performance across all evaluated models and datasets, with larger gains on perceptually demanding tasks. Earlier-generation models show larger gains than their more recent counterparts. For instance, LLaVA-1.5-7B gains +27.8 absolute points on V*, whereas Qwen2.5-VL-7B gains +8.9 points. This pattern indicates that less capable models suffer more from visual complexity interference and benefit more

Table 2: We investigate CARVE's accuracy using both single-layer and multi-layer intervention strategies at shallow, middle, and deep model depths, where single-layer interventions use attention maps $\hat{A}_i$ from individual layers, while multi-layer interventions fuse maps across multiple layers to guide masking decisions. We employ $\mathcal{T}_{\text{full}}$ as the time step configuration.

Model	Layer(s): $\mathcal{L}$		A-OKVQA	POPE	V*	TextVQA
Qwen2.5-VL-3B	Vanilla		73.0	86.9	50.3	72.8
	Single Layer	14	74.3 ▲ 1.78	87.1 ▲ 0.23	53.9 ▲ 7.16	73.6 ▲ 1.10
		20	76.5 ▲ 4.79	87.4 ▲ 0.58	56.0 ▲ 11.33	74.7 ▲ 2.61
		25	76.7 ▲ 5.07	87.5 ▲ 0.69	56.0 ▲ 11.33	75.9 ▲ 4.26
	Multi-Layers	[10, 15]	74.0 ▲ 1.37	86.9 0.00	53.4 ▲ 6.16	73.0 ▲ 0.27
		[15, 20]	76.8 ▲ 5.21	87.7 ▲ 0.92	56.0 ▲ 11.33	76.0 ▲ 4.40
		[20, 25]	**78.3** ▲ 7.26	**87.9** ▲ 1.15	**56.5** ▲ 12.33	**76.3** ▲ 4.81
Qwen2.5-VL-7B	Vanilla		75.0	87.0	50.8	75.0
	Single Layer	14	75.2 ▲ 0.27	87.5 ▲ 0.57	54.5 ▲ 7.28	75.2 ▲ 0.27
		20	76.9 ▲ 2.53	87.9 ▲ 1.03	56.5 ▲ 11.22	77.9 ▲ 3.87
		25	77.0 ▲ 2.67	88.2 ▲ 1.38	57.1 ▲ 12.40	78.4 ▲ 4.53
	Multi-Layers	[10, 15]	75.0 0.00	87.0 0.00	51.3 ▲ 0.98	75.0 0.00
		[15, 20]	77.1 ▲ 2.80	88.4 ▲ 1.61	57.6 ▲ 13.39	79.5 ▲ 6.00
		[20, 25]	**78.0** ▲ 4.00	**88.6** ▲ 1.84	**58.1** ▲ 14.37	**81.7** ▲ 8.93
LLaVA-1.5-7B	Vanilla		71.5	83.6	38.7	47.8
	Single Layer	14	71.7 ▲ 0.28	85.1 ▲ 1.79	63.4 ▲ 63.82	54.0 ▲ 12.97
		20	74.0 ▲ 3.50	87.2 ▲ 4.31	65.4 ▲ 68.99	56.1 ▲ 17.36
		25	74.1 ▲ 3.64	87.1 ▲ 4.19	65.4 ▲ 68.99	56.2 ▲ 17.57
	Multi-Layers	[10, 15]	71.5 0.00	84.5 ▲ 1.08	48.2 ▲ 24.55	49.2 ▲ 2.93
		[15, 20]	74.2 ▲ 3.78	87.5 ▲ 4.67	65.4 ▲ 68.99	56.4 ▲ 17.99
		[20, 25]	**75.4** ▲ 5.45	**89.0** ▲ 6.46	**66.5** ▲ 71.83	**57.9** ▲ 21.13
LLaVA-1.5-13B	Vanilla		75.7	84.6	42.4	57.1
	Single Layer	14	75.8 ▲ 0.13	86.1 ▲ 1.77	66.5 ▲ 56.84	58.2 ▲ 1.93
		20	76.2 ▲ 0.66	88.2 ▲ 4.26	68.6 ▲ 61.79	59.1 ▲ 3.50
		25	76.2 ▲ 0.66	88.1 ▲ 4.14	69.1 ▲ 62.97	59.2 ▲ 3.68
	Multi-Layers	[10, 15]	75.7 0.00	85.0 ▲ 0.47	52.9 ▲ 24.76	57.4 ▲ 0.53
		[15, 20]	76.8 ▲ 1.45	88.6 ▲ 4.73	69.1 ▲ 62.97	59.4 ▲ 4.03
		[20, 25]	**76.9** ▲ 1.59	**90.1** ▲ 6.50	**70.2** ▲ 65.57	**61.2** ▲ 7.18

from contrastive attention focusing. The particularly large V* gains are attributable to its benchmark design: every sample requires localizing a fine-grained target within a cluttered scene, precisely the scenario CARVE addresses. Despite the modest sample size, the improvement on V* for LLaVA-1.5-7B is statistically significant: the 95% bootstrap confidence intervals do not overlap, spanning 31.9–46.1% for the vanilla model versus 59.7–72.8% for CARVE, and a paired McNemar's test further confirms significance at $p < 0.001$.

Cross-Dataset Comparison. Table 3 compares CARVE with ViCrop and SD-RPN on three additional benchmarks, where SD-RPN additionally requires training. CARVE, despite being entirely training-free, achieves the best results in the majority of settings. Improvements scale with visual perception demands:

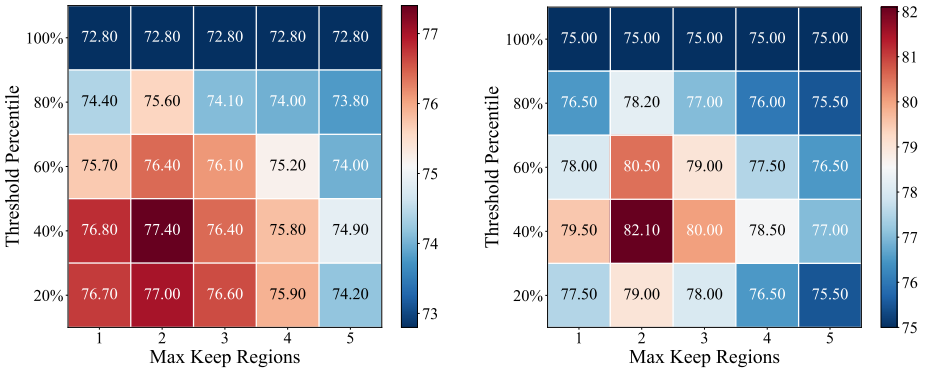
Table 3: Extended cross-dataset evaluation on three additional benchmarks. DocVQA and ChartQA primarily involve local visual perception, while ScienceQA requires knowledge-based reasoning.

Model		DocVQA		ChartQA		ScienceQA
		EM	ANLS	EM	Relaxed	MC Acc
Qwen2.5-VL-3B	Vanilla	49.70	66.07	67.64	77.92	75.11
	ViCrop	55.80 ▲ 12.27	72.27 ▲ 9.38	68.92 ▲ 1.89	79.64 ▲ 2.21	75.21 ▲ 0.13
	SD-RPN	60.90 ▲ 22.54	73.50 ▲ 11.25	69.88 ▲ 3.31	80.56 ▲ 3.39	75.36 ▲ 0.33
	CARVE	**61.70** ▲ 24.14	**74.94** ▲ 13.43	**70.80** ▲ 4.67	**81.76** ▲ 4.93	**75.56** ▲ 0.60
Qwen2.5-VL-7B	Vanilla	54.30	69.97	69.32	79.12	81.21
	ViCrop	57.60 ▲ 6.08	70.84 ▲ 1.24	69.76 ▲ 0.63	79.48 ▲ 0.46	81.56 ▲ 0.43
	SD-RPN	55.90 ▲ 2.95	72.23 ▲ 3.23	70.36 ▲ 1.50	**80.12** ▲ 1.26	**83.84** ▲ 3.24
	CARVE	**63.00** ▲ 16.02	**75.95** ▲ 8.55	**72.48** ▲ 4.56	79.96 ▲ 1.06	82.50 ▲ 1.59
LLaVA-1.5-7B	Vanilla	11.20	19.78	8.96	17.88	59.10
	ViCrop	11.90 ▲ 6.25	27.00 ▲ 36.50	9.12 ▲ 1.79	18.12 ▲ 1.34	59.49 ▲ 0.66
	SD-RPN	12.20 ▲ 8.93	**29.20** ▲ 47.62	10.72 ▲ 19.64	20.12 ▲ 12.53	59.84 ▲ 1.25
	CARVE	**12.40** ▲ 10.71	28.21 ▲ 42.62	**11.20** ▲ 25.00	**20.60** ▲ 15.21	**60.39** ▲ 2.18
LLaVA-1.5-13B	Vanilla	14.00	24.42	9.88	18.20	61.68
	ViCrop	14.40 ▲ 2.86	31.20 ▲ 27.76	12.08 ▲ 22.27	20.12 ▲ 10.55	62.92 ▲ 2.01
	SD-RPN	15.90 ▲ 13.57	33.32 ▲ 36.45	11.64 ▲ 17.81	19.76 ▲ 8.57	63.46 ▲ 2.89
	CARVE	**16.20** ▲ 15.71	**33.48** ▲ 37.10	**13.32** ▲ 34.82	**21.20** ▲ 16.48	**65.34** ▲ 5.93

DocVQA, requiring fine-grained text recognition in dense layouts, yields the largest relative gains of up to +24.14% EM for Qwen2.5-VL-3B; ChartQA shows moderate improvements of +4.67% EM for Qwen2.5-VL-3B; and ScienceQA, relying on knowledge reasoning, yields consistently positive yet smaller gains ranging from +0.60% to +5.93%. This gradient across tasks suggests that suppressing visual noise benefits tasks roughly in proportion to their perceptual demands, with no observed degradation of reasoning performance.

Time Step Ablation. Table 1 shows a clear ranking among time step selection strategies: t_{end} generally outperforms $\mathcal{T}_{\text{full}}$, which in turn surpasses t_{start}. For example, on TextVQA with Qwen2.5-VL-7B, t_{end} reaches 81.9%, $\mathcal{T}_{\text{full}}$ 81.7%, and t_{start} 80.7%. This ranking is consistent with autoregressive decoding: later tokens access longer contexts, so the final token's attention maps better localize task-relevant objects for CARVE's masking.

Layer Selection Ablation. Table 2 reports the effect of layer selection on attention extraction. Across all models, [20,25] consistently achieves the best performance while [10,15] ranks last. The general trend from best to worst is: [20,25], [15,20], single layer 25, single layer 20, single layer 14, and [10,15], with occasional ties among adjacent single-layer configurations. For instance, LLaVA-1.5-7B on TextVQA gains +21.13% with [20,25], +17.99% with [15,20], but only +2.93% with [10,15]. Multi-layer fusion captures complementary patterns while reducing noise, consistent with the layer-wise entropy trend in Fig. 5: early layers perform



(a) Accuracy for Qwen2.5-VL-3B on TextVQA **(b)** Accuracy for Qwen2.5-VL-7B on TextVQA

Fig. 7: Mask generation hyperparameter impact on TextVQA accuracy for Qwen family, across top-p threshold and maximum keep regions K.

Table 4: Component ablation (accuracy %, Qwen2.5-VL-7B). Direct $A_i^{(Q)}$ masking omits contrastive normalization; in-attention modification replaces pixel-level masking with direct attention weight manipulation.

Method	TextVQA	ChartQA	ScienceQA
Vanilla	75.0	79.12	81.21
Direct $A_i^{(Q)}$ masking	78.6 ▲ 4.80	79.92 ▲ 1.01	82.30 ▲ 1.34
In-attention modification	77.1 ▲ 2.80	79.56 ▲ 0.56	82.10 ▲ 1.10
CARVE (full)	**82.0** ▲ 9.33	**79.96** ▲ 1.06	**82.50** ▲ 1.59

global scanning with high entropy, while deeper layers focus on task-relevant patterns.

Sensitivity Analysis of Mask Generation. Fig. 7 reports accuracy on a 1,000-instance TextVQA subset across varying p and K. Setting $p \in [0.2, 0.6]$ with $K \in \{2, 3\}$ yields optimal performance by balancing noise suppression with object preservation. Fig. 8 qualitatively confirms this: as τ tightens, CARVE progressively filters background noise, shifting focus to task-relevant cues and correcting hallucinations.

Component Ablation Study. Table 4 evaluates CARVE's core design choices using Qwen2.5-VL-7B. On TextVQA, direct $A_i^{(Q)}$ masking yields +4.80% while CARVE achieves +9.33%, confirming that contrastive normalization effectively suppresses $\mathcal{F}_{\text{vis}}(\mathcal{I})$. In-attention modification produces the weakest gains, indicating that pixel-level masking outperforms internal representation manipulation. Even on high-baseline tasks such as ChartQA and ScienceQA, all variants yield positive gains ranging from +0.56% to +1.59%, indicating that both contrastive normalization and pixel-level masking contribute.

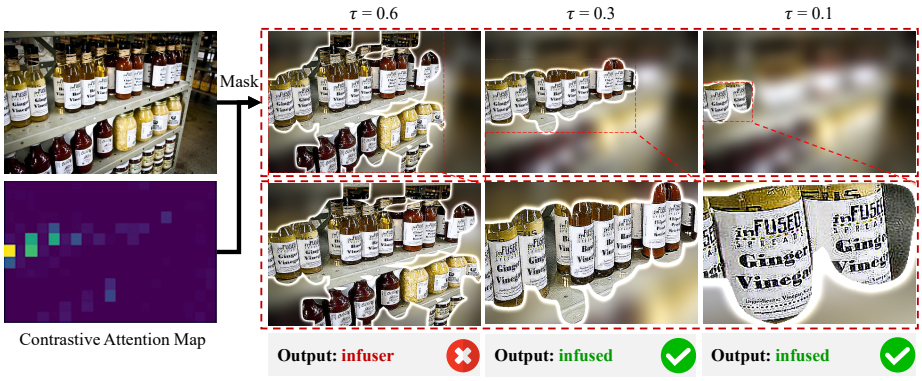


Fig. 8: Visualization of contrastive attention refinement in VQA. As the threshold τ decreases from 0.6 to 0.1, CARVE progressively suppresses visual noise to isolate task-relevant regions. Precise grounding corrects initial hallucinations (Red ×) and yields accurate answers (Green ✓).

Table 5: CARVE vs. external tools and ViCrop on TextVQA with LLaVA-1.5-7B: accuracy (%) and inference latency per sample (s).

Method	Original	SAM	YOLO	CLIP	ViCrop			CARVE
					rel-att	grad-att	pure-grad	
Accuracy	47.80	49.42	48.84	48.55	55.17	<u>56.06</u>	51.67	**58.2**
Inf. Latency	0.17	3.33	0.35	1.07	1.16	0.89	2.36	1.34

Comparison with External Tools. As shown in Table 5, CARVE outperforms all external tool-based approaches: SAM [12], YOLO [23], CLIP [22], and the recent ViCrop [34] variants. External tools rely on generic segmentation algorithms that are not conditioned on the input question. While ViCrop effectively reduces visual noise through strategic cropping, it does not perform pixel-level noise masking. In terms of latency, CARVE requires an inference latency[1] of 1.34 seconds, exceeding simpler approaches such as YOLO at 0.35 seconds but remaining within practical deployment constraints.

Inference Overhead Analysis. Table 6 reports GPU time[2] and memory measurements. CARVE's GPU time ranges from 0.5 to 1.3 seconds per sample. Besides the three autoregressive passes, overhead arises from attention extraction and storage at each decoding step, connected component analysis for mask generation, and image cropping with resizing. The memory overhead is primarily

[1] Inference latency denotes end-to-end time from input to output, including all processing steps such as image preprocessing, GPU computation, and post-processing.

[2] GPU time measures pure GPU computation time, excluding image pre/post-processing and data transfer.

Table 6: GPU time (ms) and peak GPU memory (GB), averaged over 100 ChartQA samples. CARVE GPU time includes general description inference, attention processing, and enhanced inference.

Model	Vanilla	CARVE	Vanilla Mem.	CARVE Mem.
Qwen2.5-VL-3B	358	1,038	7.5	8.1
Qwen2.5-VL-7B	415	1,286	15.9	16.6
LLaVA-1.5-7B	149	506	13.6	15.2
LLaVA-1.5-13B	233	762	25.7	28.9

attributed to attention map storage during the first two passes, ranging from +0.6 GB to +3.2 GB and scaling with model size. Early termination and attention caching strategies can further reduce this overhead.

Limitations. CARVE suppresses background to magnify spatially localized cues, making it most effective for local fine-grained perception such as text recognition and small-target identification. For tasks that hinge on *global* spatial context—such as depth ordering or object-relationship reasoning—masking may remove necessary contextual cues; in such cases CARVE gracefully degrades to the vanilla model as $p \to 1.0$. CARVE also requires three forward passes, trading additional inference cost for accuracy.

6 Conclusion

In this work, we demonstrate that visual complexity correlates with attention entropy, which in turn negatively impacts VLMs' performance. Theoretically, under a multiplicative decomposition assumption, we show that contrasting attention maps between general and specific instructions approximately separates visual signal into semantic and noise components. To this end, we propose **CARVE**, a training-free method that exploits this decomposition to extract task-relevant signal through attention contrasting and pixel-level masking, yielding consistent gains across seven benchmarks.

Acknowledgement. This work is supported in part by the National Key R&D Program of China under Grant Nos. 2025YFC3309700, the National Natural Science Foundation of China under Grant Nos. U25B2076, 62441229, and 62377043, and the Beijing Natural Science Foundation No. 4262033.

References

1. Acharya, S., Jia, F., Ginsburg, B.: Star attention: Efficient llm inference over long sequences. arXiv preprint arXiv:2411.17116 (2024)

2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-vl technical report (2025)
4. Bi, B., Liu, S., Mei, L., Wang, Y., Fang, J., Ji, P., Cheng, X.: Decoding by contrasting knowledge: Enhancing large language model confidence on edited facts. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 17198–17208 (2025)
5. Canny, J.: A computational approach to edge detection. IEEE Transactions on pattern analysis and machine intelligence **PAMI-8**(6), 679–698 (1986)
6. Chen, L., Zhao, H., Liu, T., Bai, S., Lin, J., Zhou, C., Chang, B.: 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. pp. 19–35. Springer (2024)
7. Chen, L., Zhou, B., Ge, Y., Chen, J., Ni, S.: Pis: Linking importance sampling and attention mechanisms for efficient prompt compression. arXiv preprint arXiv:2504.16574 (2025)
8. Gkartzonika, I., Gkalelis, N., Mezaris, V.: Learning visual explanations for dcnn-based image classifiers using an attention mechanism. In: European Conference on Computer Vision. pp. 396–411. Springer (2022)
9. Hong, W., Wang, W., Lv, Q., Xu, J., Yu, W., Ji, J., Wang, Y., Wang, Z., Dong, Y., Ding, M., et al.: Cogagent: A visual language model for gui agents. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14281–14290 (2024)
10. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
11. Jiang, C., Xu, H., Dong, M., Chen, J., Ye, W., Yan, M., Ye, Q., Zhang, J., Huang, F., Zhang, S.: Hallucination augmented contrastive learning for multimodal large language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27036–27046 (2024)
12. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
13. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, X., Wen, J.R.: 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)
14. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
15. Liu, Y., Shi, K., He, K., Ye, L., Fabbri, A.R., Liu, P., Radev, D., Cohan, A.: On learning to summarize with large language models as references. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 8647–8664 (2024)
16. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. Advances in neural information processing systems **35**, 2507–2521 (2022)

17. Ma, F., Zhou, Y., Zhang, Z., Yan, S., Li, H., He, Z., Wu, S., Rao, F., Zhang, Y., Sun, X.: Ee-mllm: A data-efficient and compute-efficient multimodal large language model. arXiv preprint arXiv:2408.11795 (2024)
18. Masry, A., Tan, J.Q., Joty, S., Hoque, E., et al.: Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In: Findings of the association for computational linguistics: ACL 2022. pp. 2263–2279 (2022)
19. Mathew, M., Karatzas, D., Jawahar, C.: Docvqa: A dataset for vqa on document images. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2200–2209 (2021)
20. Pan, K., Fan, Z., Li, J., Yu, Q., Fei, H., Tang, S., Hong, R., Zhang, H., Sun, Q.: Towards unified multimodal editing with enhanced knowledge collaboration. Advances in Neural Information Processing Systems **37**, 110290–110314 (2024)
21. Qi, J., Ding, M., Wang, W., Bai, Y., Lv, Q., Hong, W., Xu, B., Hou, L., Li, J., Dong, Y., et al.: Cogcom: Train large vision-language models diving into details through chain of manipulations. arXiv preprint arXiv:2402.04236 (2024)
22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
23. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
24. Schwenk, D., Khandelwal, A., Clark, C., Marino, K., Mottaghi, R.: A-okvqa: A benchmark for visual question answering using world knowledge. In: European conference on computer vision. pp. 146–162. Springer (2022)
25. Shannon, C.E.: A mathematical theory of communication. The Bell system technical journal **27**(3), 379–423 (1948)
26. Shi, Y., Pei, X., Dong, M., Xu, C.: Catching the details: Self-distilled roi predictors for fine-grained mllm perception. arXiv preprint arXiv:2509.16944 (2025)
27. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8317–8326 (2019)
28. Team, K., Bai, T., Bai, Y., Bao, Y., Cai, S.H., Cao, Y., Charles, Y., Che, H.S., Chen, C., Chen, G., et al.: Kimi k2.5: Visual agentic intelligence (2026), arXiv:2602.02276
29. Treisman, A.M., Gelade, G.: A feature-integration theory of attention. Cognitive psychology **12**(1), 97–136 (1980)
30. Wu, P., Xie, S.: V*: Guided visual search as a core mechanism in multimodal llms. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13084–13094 (2024)
31. Yao, J., Liu, S., Wang, Y., Mei, L., Bi, B., Ge, Y., Li, Z., Cheng, X.: Who is in the spotlight: The hidden bias undermining multimodal retrieval-augmented generation. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 15194–15204 (2025)
32. You, K., Zhang, H., Schoop, E., Weers, F., Swearngin, A., Nichols, J., Yang, Y., Gan, Z.: Ferret-ui: Grounded mobile ui understanding with multimodal llms. In: European Conference on Computer Vision. pp. 240–255. Springer (2024)
33. Zhai, S., Meng, Y., Zhang, Y., Qi, G.: Parameter-aware contrastive knowledge editing: Tracing and rectifying based on critical transmission paths. In: Proceed-

ings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 28189–28200 (2025)
34. Zhang, J., Khayatkhoei, M., Chhikara, P., Ilievski, F.: MLLMs know where to look: Training-free perception of small visual details with multimodal LLMs. In: The Thirteenth International Conference on Learning Representations (2025)
35. Zhang, S., Yu, T., Feng, Y.: Truthx: Alleviating hallucinations by editing large language models in truthful space. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 8908–8949 (2024)
36. Zhang, X., Quan, Y., Gu, C., Shen, C., Yuan, X., Yan, S., Cheng, H., Wu, K., Ye, J.: Seeing clearly by layer two: Enhancing attention heads to alleviate hallucination in lvlms. arXiv preprint arXiv:2411.09968 (2024)