REVISITING MULTIMODAL PRE-TRAINING FOR VISION LANGUAGE ALIGNMENT IN MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language alignment is a key prerequisite for MLLMs to perceive and understand the visual world. Although existing MLLMs typically employ a multimodal pre-training stage for building alignment, we find that the fine-grained, patch-level cross-modal correspondence remains insufficiently established. As text-to-image (T2I) attention plays a critical role in building cross-modal correspondence, we examine it on a pre-trained model and find that T2I attention rapidly shifts from broadly distributing over the image to concentrating on a small subset of textually irrelevant visual tokens, termed *visual attention sinks*, rather than attending to text-relevant visual evidence. Through further analyses of visual sinks' attention flow and training dynamics, we observe that they emerge in the early loss-reduction phase and persistently hinder the projector from learning fine-grained cross-modal correspondence. To address this issue, we propose **Vi**sual **S**ink-aware **Align**ment (**ViSAlign**), a simple and effective method that temporarily blocks the visual sinks' attention flow during multimodal pre-training. By preventing visual sinks from dominating the attention, it encourages the model to capture text-relevant visual evidence, thus learning a finer-grained cross-modal correspondence. Extensive experiments across diverse multimodal settings demonstrate that ViSAlign improves projector learning, strengthens vision-language alignment, and yields broad improvements over a variety of downstream tasks.

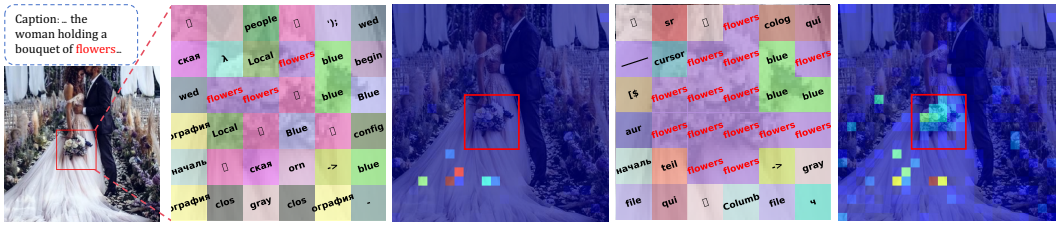
1 INTRODUCTION

Vision-language alignment determines whether multimodal large language models (MLLMs) (Liu et al., 2023; 2024; Bai et al., 2025; Wang et al., 2025) can effectively integrate and use information from diverse modalities (Yin et al., 2024). Because vision encoders and large language models (LLMs) are optimized with different objectives and operate in distinct representation spaces, directly integrating visual representations into an LLM remains challenging. Existing MLLMs therefore commonly introduce a lightweight cross-modal projector to bridge the two modalities and a multimodal pre-training stage for training this randomly initialized projector. At this stage, the projector learns to map visual representations into the LLM space, establishing both dimensional and semantic vision-language alignment for the following instruction tuning stage.

Although existing MLLMs demonstrate strong visual understanding across multimodal tasks, whether they establish fine-grained vision-language alignment remains underexplored. By probing the cross-modal correlation within the pre-trained model of a representative MLLM, LLaVA-1.5 (Liu et al., 2024), our observations in Figure 1(a) indicate that the patch-level vision-language alignment is not fully established, where only a few visual tokens exhibit clear correspondence with the relevant text concept "flowers". Because cross-modal attention is the critical mechanism for building alignment in MLLMs (Shukor & Cord, 2024), we examine the attention flow in the pre-trained model to investigate the underlying cause of this suboptimal alignment. Our observations show that the model broadly aggregates global visual information within the shallow layers; afterwards, as shown in Figure 1(b), its attention rapidly concentrates on a few background visual tokens irrelevant to the token "flowers", usually termed *visual attention sinks* (Kang et al., 2025), rather than attending to text-relevant visual evidence.

Our further investigations focus on the properties of these visual sinks, which are textually irrelevant but absorb excessive cross-modal attention, revealing how they influence the projector learning for

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



(a) cross-modal correlation (b) T2I attention visualization (c) correlation of ViSAlign (d) T2I attention of ViSAlign

Figure 1: (a) Illustration of cross-modal correlation, probing via Embed Lens (Fan et al., 2026), for a specific region of "flowers" in the image. (b) Text-to-image (T2I) attention concentrates on visual sinks while ignoring text-relevant visual evidence when predicting the target caption "flowers". (c) Cross-modal correlation when ViSAlign is applied. (d) T2I attention of "flowers" when ViSAlign is applied, where attention is relocated to the flowers region in the image.

vision-language alignment. First, we confirm that visual sinks in the pre-trained model exhibit patterns similar to those observed in LLMs and ViTs, referred to as LLM-emerged sinks (L-sinks) and ViT-propagated sinks (V-sinks), respectively (Luo et al., 2026; Choi et al., 2026). Based on attention analysis, intervention study, and tracking training dynamics, we make three key findings: **(1)** L-sinks dominate the cross-modal attention and cause the attention pattern transition, while V-sinks become increasingly prominent in deeper layers; **(2)** Both L-sinks and V-sinks weaken the model's ability to exploit text-relevant fine-grained visual information; **(3)** The rapid formation of visual sinks closely coincides with the sharp decrease in training loss. Together, these observations suggest that *visual sinks emerge rapidly during the early loss-reduction phase, persistently hindering the projector from learning fine-grained cross-modal correspondences.*

To facilitate the learning of the projector for enhanced visual-language alignment, we propose a simple and effective method, **ViSAlign** (**Vi**sual **S**ink-aware **Align**ment). It temporarily blocks the attention flow attending to visual sinks during multimodal pre-training, thereby suppressing the overdominance of visual sinks and encouraging the model to acquire information from the text-relevant visual regions. With modified cross-modal attention, the projector learns richer, more fine-grained cross-modal correspondences, establishing better vision-language alignment. As shown in Figure 1(c) and Figure 1(d), ViSAlign enhances the patch-level cross-modal correlation and reallocates T2I attention toward the regions of flowers. We then validate the effectiveness of ViSAlign through comprehensive experiments and provide detailed analyses of its positive impact on the vision-language alignment of the pre-trained model and downstream task performance.

Our contributions are summarized as follows. **Firstly**, we investigate the cross-modal attention flow and investigate visual attention sinks of the representative pre-trained MLLM, revealing that visual sinks emerge during the early loss-reduction phase and consistently undermine the learning of fine-grained cross-modal correspondence. **Secondly**, we propose ViSAlign, a simple and effective method that blocks the text captions from attending to visual sinks, forcing the model to learn fine-grained cross-modal correlations. **Finally**, through comprehensive experiments, we demonstrate that ViSAlign enhances alignment and improves model performance on downstream tasks.

2 Related Work

Multimodal large language models (MLLMs) (Team, 2026; Bai et al., 2025; An et al., 2025; Liu et al., 2023; Wang et al., 2025) typically consist of a pre-trained LLM backbone, a vision encoder, and a cross-modal connector. Contrastively trained models such as CLIP (Radford et al., 2021) and SigLIP (Zhai et al., 2023; Tschannen et al., 2025) are widely used as vision encoders, while connectors range from MLPs (Liu et al., 2023) and cross-attention modules (Li et al., 2023a; Alayrac et al., 2022; Cha et al., 2024) to pre-trained language models (Yu et al., 2026). Although visual and textual tokens remain separated in the representation space (Shukor & Cord, 2024), their cross-modal correlations can be observed through Logit Lens (Wang, 2025; Neo et al., 2025; Wang et al., 2026b) or Embed Lens (Tang et al., 2025; Fan et al., 2026). Prior works have investigated cross-modal information flow within MLLMs, revealing that the model first captures global information, then focuses on text-related visual evidence, and finally processes information within text tokens (Zhang et al., 2025b), inspiring several works to enhance model performance (Tang et al.,

2025; Wang et al., 2026a; Jiang et al., 2026a). However, these studies mainly focus on VQA task performance, whereas we investigate how the multimodal pre-trained model processes visual information and achieves vision-language alignment.

Attention sinks (Su et al., 2026) refer to semantically irrelevant tokens that receive disproportionately high attention in transformer-based models. In LLMs, the beginning-of-sequence (BOS) token often acts as an attention sink (Xiao et al., 2024; Barbero et al., 2025; Guo et al., 2024), accompanied by massive activations in specific dimensions and unusually large token norms (Sun et al., 2024; Gu et al., 2025). In ViTs, sinks often correspond to background patches with high L_2 norms (Darcet et al., 2024), and their emergence has been attributed to coarse-grained supervision that encourages such patches to serve as global semantic shortcuts (Shi et al., 2026). In MLLMs, Kang et al. (2025) first identify the visual attention sinks and Luo et al. (2026); Choi et al. (2026) further distinguish ViT-propagated sinks (V-sinks) from LLM-emerged sinks (L-sinks). Existing methods mitigate attention sinks by modifying attention structures via gated attention or softmax functions in LLMs (Qiu et al., 2026; OpenAI, 2025), introducing register tokens in ViTs (Darcet et al., 2024; Jiang et al., 2026b), or redistributing attention toward relevant visual regions at inference in MLLMs (Shu et al., 2026; Lyu et al., 2026; Zhang et al., 2025a). However, recent studies on MLLMs mainly analyze instruction-tuned models and inference-time behavior. In this work, we focus on the multimodal pre-training stage where both the vision encoder and LLM backbone remain frozen, and investigate the emergence and effects of visual sinks in projector learning.

3 Analysis

3.1 Preliminary

The input sequence of MLLMs consists of system prompts, visual tokens, and text tokens, whose index sets are denoted by $\mathcal{I}_{sys}$, $\mathcal{I}_v$, and $\mathcal{I}_t$, respectively. Following widely used input templates (Liu et al., 2023), sequence indices are organized as $\mathcal{I} = [\mathcal{I}_{sys}, \mathcal{I}_v, \mathcal{I}_t]$, where $N = |\mathcal{I}|$ is the sequence length. Let $X^l \in \mathbb{R}^{N\times d}$ denote the input to the l-th transformer layer, and let $x_i^l \in \mathbb{R}^d$ denote the representation of the i-th token. Each transformer layer consists of a multi-head attention $MHA^l(\cdot)$ and a feed-forward network $FFN^l(\cdot)$. Based on these definitions, the layer update is formulated as

$$\hat{x}_i^l = MHA^l(x_i^l, X^l, M_i) + x_i^l, x_i^{l+1} = FFN^l(\hat{x}_i^l) + \hat{x}_i^l, \tag{1}$$

where $\hat{x}_i^l$ denotes the intermediate feature produced by $MHA^l()$ and $M_i \in \{0, -\infty\}^N$ is the causal attention mask.

For attention head h, the attention distribution of the i-th token over the complete sequence is computed using the projection matrices $W_Q^{l,h}, W_K^{l,h} \in \mathbb{R}^{d\times d_k}$. The attention weights are defined as

$$\Omega_i^{l,h} = \frac{(x_i^l W_Q^{l,h})(X^l W_K^{l,h})^T}{\sqrt{d_k}}. \tag{2}$$

Then, the attention score could be computed with the causal attention mask M_i,

$$A_i^{l,h} = softmax\left(\Omega_i^{l,h} + M_i\right), \tag{3}$$

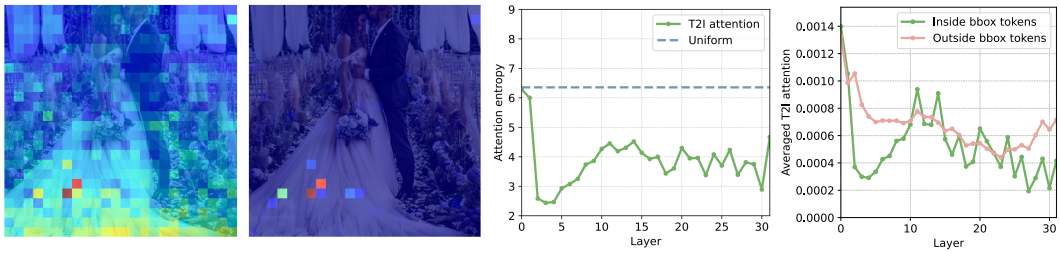
which could be further expanded as $A_i^{l,h} = [a_{(i,1)}^{l,h}, \ldots, a_{(i,N)}^{l,h}]$, where $a_{(i,j)}^{l,h}$ denotes the attention score assigned when the i-th token attends to the j-th token. For subsequent analysis, we average the attention scores over all H heads:

$$\hat{A}_i^l = \frac{1}{H}\sum_{h=1}^{H} A_i^{l,h} = [\hat{a}_{(i,1)}^l, \ldots, \hat{a}_{(i,N)}^l], \tag{4}$$

where $\hat{a}_{(i,j)}^l$ denotes the corresponding head-averaged attention score.

3.2 Cross-modal Information Flow of Pre-trained Model

In this section, we investigate *how Text-to-Image (T2I) attention develops across layers and whether the model learns to capture text-relevant visual information in intermediate layers*, as observed



(a) T2I attention, layer 0 (b) T2I attention, layer 16 (c) T2I attention entropy (d) T2I attention in bbox

Figure 2: (a,b) Cross-modal attention heatmaps projected onto the original image. (c) Evolution of T2I attention entropy, plotted against a uniform distribution baseline. (d) Average T2I attention over tokens within and outside the target bounding box, calculated based on the local caption.

in instruction-tuned models (Kang et al., 2025; Zhang et al., 2025b). Hence, we use the trained projector of LLaVA-1.5 (Liu et al., 2024) to connect the LLM backbone and vision encoder without instruction tuning and simulate pre-training forward passes with image-caption data.

Cross-modal attention heatmaps exhibit distinct attention patterns. Firstly, we project the cross-modal attention back to the original image to obtain layer-wise attention heatmaps. As shown in Figure 2(a) and (b), we visualize the attention of the predicted target text token "flowers", where an ideal T2I attention would localize the visual region of flowers. However, the cross-modal attention map exhibits two distinct patterns. At the shallow layer 0, attention is broadly distributed over the image, suggesting a *uniform pattern* to achieve global visual information aggregation. As the layer depth increases, attention converges to *visual attention sinks* (Kang et al., 2025; Luo et al., 2026) that concentrate on a small set of image tokens with extremely high attention scores.

Attention patterns shift at shallow layers. Then, we compute the entropy of the normalized T2I attention distribution for each text token and average it across the caption sequence $\mathcal{I}_t$ to obtain a more precise layer-wise evolution. Attention entropy enables us to distinguish between two characteristic patterns: a near-uniform attention pattern yields high entropy close to that of a uniform distribution, whereas visual attention sinks produce substantially lower entropy. As shown in Figure 2(c), the T2I attention entropy remains close to the uniform baseline in layers 0 and 1, indicating broadly distributed visual information aggregation. Afterwards, it drops sharply at layer 2, which rapidly concentrates on visual sinks and remains sink-dominated throughout the deeper layers.

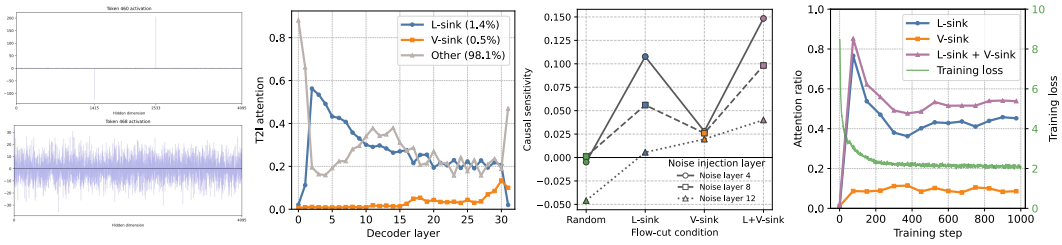
Multimodal pre-trained models cannot capture fine-grained text-relevant visual evidence. The observations in Figure 1(b) show that visual sinks absorb a large portion of the attention. We further quantitatively examine whether the model learns to capture text-relevant visual information in middle layers as discussed in instruction-tuned models (Kan et al., 2026). To this end, we construct a curated dataset of $1,000$ samples based on RefCOCO (Kazemzadeh et al., 2014). Each sample contains a caption that ends with a local caption referring to the objects circled by a bounding box. We compute the average T2I attention of the local caption that falls inside and outside the bounding box. As shown in Figure 2(d), we observe that the text-relevant visual region does not receive significantly more attention than textually irrelevant patches. More layer-wise attention visualizations supporting the inadequate fine-grained visual information acquisition can be found in Appendix A.

In conclusion, the model exhibits a transition from a uniform attention pattern for merely two layers to visual attention sinks. Unlike fully trained models, the pre-trained model acquires inadequate fine-grained visual information from text-relevant visual evidence.

3.3 Visual Sinks Undermine Projector Learning

As Section 3.2 establishes the attention pattern that emerges after multimodal pre-training, we next investigate *what causes the pattern transition and how these visual sinks influence the learning of vision-language alignment*.

Both LLM-emerged sinks (L-sinks) and ViT-propagated sinks (V-sinks) exist in the multimodal pre-trained model. Following previous studies on instruction-tuned models (Luo et al., 2026; Choi et al., 2026), we first distinguish L-sinks and V-sinks among the top-k visual tokens receiving the highest T2I attention in pre-trained model. As shown in Figure 3(a), token 440 exhibits



(a) activation dimensions (b) T2S attention (c) causal sensitivity (d) training dynamics

Figure 3: (a) Illustration of activation dimensions of token 440 and 468, representing L-sink and V-sink, respectively. (b) Text-to-sink (T2S) attention across layers. (c) Sensitivity of the model to perturbations within the bounding box region when the Attention Knockout is applied to different information pathways. (d) Training loss and the T2S attention evolution as training dynamics.

massive activations in specific dimensions $\{1415, 2533\}$, coinciding with that of the LLM (Sun et al., 2024). By contrast, visual sink 468 exhibits a relatively uniform activation pattern across dimensions, consistent with the characteristics of attention sinks in ViTs. These findings confirm that both L-sinks and V-sinks exist in the multimodal pre-trained model, and details of identification are shown in Appendix B.

L-sinks dominate visual attention sinks in the pre-trained model. Then, we analyze the cross-modal attention assigned to L-sinks, V-sinks, and the remaining visual tokens. As shown in Figure 3(b), the attention assigned to L-sinks increases sharply from layer 1 to layer 2 and exceeds the total attention assigned to the remaining visual tokens, resulting in the pattern transition in Section 3.2. In subsequent layers, L-sinks persistently receive attention comparable to that of all remaining visual tokens, dominating the cross-modal attention, while V-sinks receive much lower attention until layer 15 and become increasingly prominent in deeper layers.

Visual sinks undermine fine-grained visual information acquisition. To examine the relation between visual sinks and fine-grained acquisition of text-relevant visual evidence, we conduct a sensitivity analysis on the curated dataset in Section 3.2. As illustrated in Figure 4(a), we first perturb the visual tokens within the bounding box using Gaussian noise and compute the resulting captioning loss $\mathcal{L}'$. We define the change in loss compared to the original loss $\mathcal{L}$ as

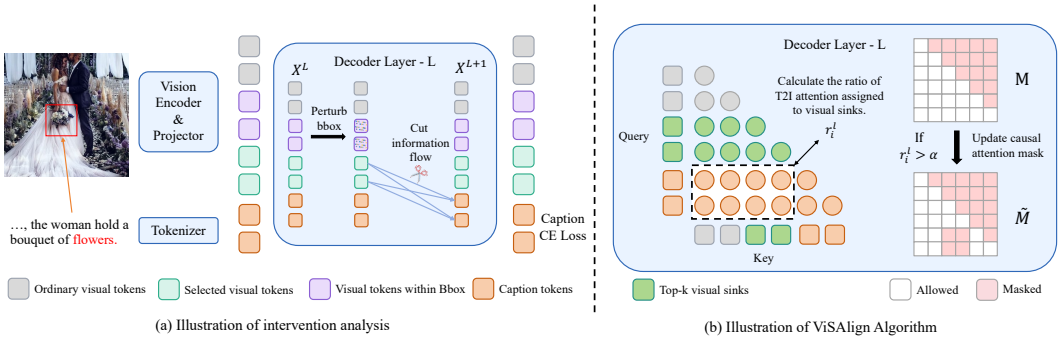
$$D = \mathcal{L}' - \mathcal{L}, \tag{5}$$

which indicates the model's reliance on the text-relevant visual region for caption prediction. Then, we consider an information flow $\mathcal{F}$ associated with a set of visual tokens $\mathcal{I}_s$. After applying Attention Knockout (Geva et al., 2023; Zhang et al., 2025b; Neo et al., 2025), which sets the pre-softmax cross-modal attention weight from caption tokens to the visual tokens in $\mathcal{I}_s$ to $-\infty$, a new loss $\mathcal{L}_\mathcal{F}$ is obtained. Under the same intervention, we again perturb bounding box visual tokens and obtain $\mathcal{L}'_\mathcal{F}$. The corresponding divergence is defined as $D_\mathcal{F} = \mathcal{L}'_\mathcal{F} - \mathcal{L}_\mathcal{F}$. Finally, the causal sensitivity of pathway $\mathcal{F}$ is defined as:

$$C_\mathcal{F} = D_\mathcal{F} - D. \tag{6}$$

This quantity measures how intervening on information flow $\mathcal{F}$ changes the model's dependence on the evidential visual region. A larger $C_\mathcal{F}$ indicates that cutting the pathway causes the model to rely more strongly on region-specific visual evidence, suggesting that the original pathway may suppress such dependence. For comparison, we intervene on information flows associated with randomly selected visual tokens, L-sinks, V-sinks, and the union of L-sinks and V-sinks, while varying the layer at which the visual tokens are perturbed. As shown in Figure 3(c), both L-sinks and V-sinks harm the dependence of text prediction on region-specific visual evidence. L-sinks dominates this harmness, while the union of L-sinks and V-sinks yields the strongest effect.

Visual sinks emerge as the early training loss converges. To investigate when visual attention sinks emerge, we save the intermediate checkpoints and compute the T2I attention assigned to L-sinks and V-sinks with 1,000 samples in the pre-training dataset, alongside recording the training loss curve. We dynamically track the activations to determine visual sinks for each checkpoint. As shown in Figure 3(d), we observed that: (1) visual sinks emerge at the early steps in pre-training, where the attention score assigned to L-sinks and V-sinks rises sharply as training loss decreases



(a) Illustration of intervention analysis

(b) Illustration of ViSAlign Algorithm

Figure 4: (a) Illustration of how the intervention study is conducted, where the bounding box hidden state is perturbed by Gaussian noise within decoder layers and the information flow from selected visual tokens to caption tokens is adaptively cut down. (b) Illustration of the ViSAlign algorithm, which identifies top-k visual sinks and adaptively updates the causal attention mask.

rapidly within early steps. (2) After this initial increase, L-sinks attention gradually decreases but remains considerable as training progresses, and V-sinks maintain a smaller but stable proportion.

In summary, visual sinks emerge alongside the rapid convergence of the training loss and persist throughout the subsequent training process. This attention pattern reduces the dependence of the model on region-specific visual evidence, thereby hindering the projector from learning the fine-grained cross-modal correspondence for vision-language alignment.

4 Methodology

To boost vision-language alignment during multimodal pre-training, we propose **Vi**sual **S**ink-aware **Align**ment (**ViSAlign**), a simple and effective method that temporarily blocks the visual information flow to mitigate the visual attention sinks, encouraging the model to learn fine-grained cross-modal correspondence. Notably, ViSAlign is applied exclusively during the multimodal pre-training without modifications during inference. Algorithm 1 summarizes the procedure.

Specifically, ViSAlign first locates the indices of visual sinks and then suppresses the text-to-sink (T2S) attention flow when the visual attention sinks become overdominant. This design inherently forces the model to predict caption tokens based on visual evidence rather than semantically irrelevant visual sinks. Since visual sinks receive disproportionately high attention from caption tokens, we define the visual sink set $\mathcal{S}^l_{i,k}$ as the top-k visual tokens with the largest T2I attention scores of the i-th token:

$$\mathcal{S}^l_{i,k} = \mathrm{TopK}_{v \in \mathcal{I}_v} \left(\hat{a}^l_{i,v}, k \right). \tag{7}$$

Given the indices of visual sinks, we seek to adaptively remove the influence of visual sinks during the multimodal pre-training stage when they receive excessively high attention that hinders the projector's learning. To this end, we introduce a threshold α on the attention sink ratio r^l_i, which measures the fraction of attention absorbed by visual sinks:

$$r^l_i = \frac{\sum_{s \in \mathcal{S}^l_{i,k}} \hat{a}^l_{i,s}}{\sum_{v \in \mathcal{I}_v} \hat{a}^l_{i,v}}. \tag{8}$$

When $r^l_i > \alpha$, we regard the T2I attention as being dominated by the visual attention sinks, which suppress the learning of fine-grained cross-modal correspondence. In such cases, we block the T2S attention flow to prevent the model from attending to visual sinks, which could fundamentally

Algorithm 1: Visual Sink-aware Alignment

Require: Input sequence X^0, number of sinks set k, threshold α

for each transformer layer l **do**
  Calculate the $\Omega^{l,h}_i$ using Eq. 2
  `with torch.no_grad():`
    Calculate the $\hat{A}^l_i$ using Eq. 3 and Eq. 4
    Define the visual sink set using Eq. 7
    Calculate the r^l_i using Eq. 8
  if $r^l_i > \alpha$ **then**
    Update the M to $\widetilde{M}$ using Eq. 9
  end if
  Calculate X^{l+1} from X^l, with $\Omega^{l,h}_i$ and $\widetilde{M}$ using Eq. 1
end for

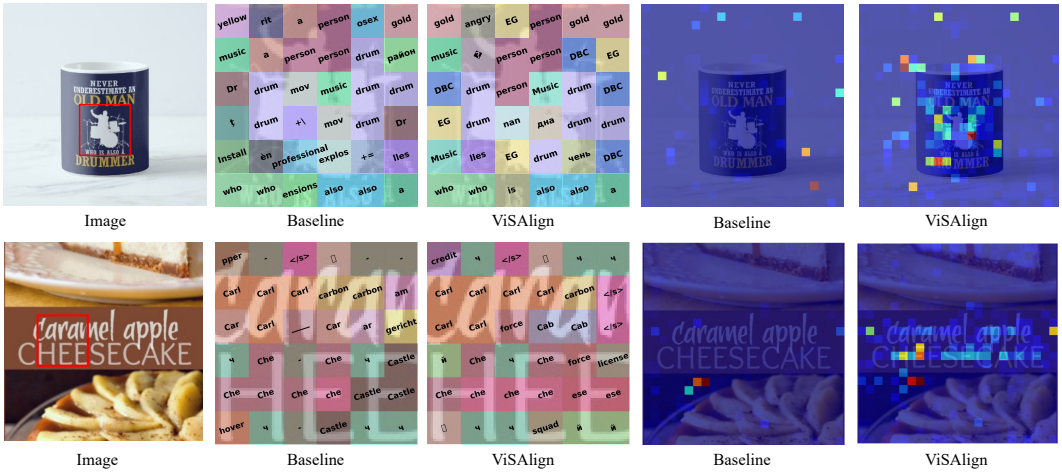


Figure 5: Comparison between the baseline and ViSAlign on cross-modal correspondence within a specific bounding box, probed by Embed Lens (Fan et al., 2026), and cross-modal attention.

Table 1: Captioning performance comparison between the baseline and ViSAlign on COCO2014 (Lin et al., 2014) validation set. The best results are highlighted in bold.

LLM	ViSAlign	BLEU				ROUGE-L	CIDEr
		$BLEU_1$	$BLEU_2$	$BLEU_3$	$BLEU_4$		
Vicuna-v1.5-7B	-	0.519	0.360	0.243	0.163	0.436	0.493
	✓	**0.630**	**0.443**	**0.305**	**0.208**	**0.463**	**0.736**
Vicuna-v1.5-13B	-	0.650	0.474	0.333	0.231	0.472	0.803
	✓	**0.665**	**0.485**	**0.341**	**0.237**	**0.476**	**0.831**
Qwen2.5-7B-Instruct	-	0.546	0.372	0.248	0.163	0.443	0.464
	✓	**0.577**	**0.396**	**0.263**	**0.172**	**0.460**	**0.528**
Phi3-3.8B	-	0.428	0.281	0.188	0.120	0.320	0.310
	✓	**0.594**	**0.418**	**0.288**	**0.195**	**0.415**	**0.583**

reshape the T2I attention distribution. Therefore, we modify the causal attention mask for all text queries as follows:

$$\widetilde{M}^l_{i,j} = \begin{cases} -\infty, & M_{i,j} = -\infty, \\ -\infty, & i \in \mathcal{I}_t,\ j \in \mathcal{S}^l_{i,k}, r^l_i > \alpha, \\ 0, & \text{otherwise}, \end{cases} \tag{9}$$

where $M_{i,j}$ denotes the original mask value when the i-th token attends to the j-th token, and $\widetilde{M}^l_{i,j}$ denotes the attention knockout mask value at layer l.

Overall, by identifying visual sinks and adaptively applying sink-aware visual attention masks, ViSAlign temporarily prevents the model from developing defective attention patterns during training, and thus encourages the model to learn a finer-grained cross-modal correspondence.

5 Experiments

Experimental setup. We follow the LLaVA-1.5 setup (Liu et al., 2024) to ensure a fair comparison. Specifically, we use CLIP ViT-L/14-336px (Radford et al., 2021) as the vision encoder, and a two-layer MLP with GELU activation as the projector. The LLM backbones with different architectures and parameter scales range from Vicuna-v1.5-7B (Chiang et al., 2023), denoted as Baseline, Vicuna-v1.5-13B (Chiang et al., 2023), Qwen2.5-7B-Instruct (Team, 2024), and Phi-3-mini-4k-instruct-3.8B (Abdin et al., 2024). As for the training strategy, ViSAlign is exclusively applied in the multimodal pre-training stage. The implementation details can be found in Appendix C.1.

Table 2: Pointing-in-box accuracy comparison between the baseline and ViSAlign on referring expression grounding benchmarks. The best results are highlighted in bold.

Model	RefCOCO			RefCOCO+			RefCOCOg	
	testA	testB	val	testA	testB	val	test	val
Baseline	60.83	49.54	55.75	61.44	48.93	53.45	55.94	55.69
ViSAlign	**64.82**	**56.57**	**62.75**	**65.13**	**57.34**	**60.27**	**65.07**	**63.48**

Table 3: Controlled comparison after instruction tuning with four LLM backbones. Higher values are better for every metric. Green and red annotations show increases and decreases relative to the corresponding reproduced baseline. The best results are highlighted in bold.

LLM	Method	General			Vision-Centric			Hallu
		MME-P	GQA	MMMU	MMVP	RWQA	MM*	POPE
Vicuna-v1.5-7B	-	1487.2	62.4	35.7	27.3	54.9	35.5	86.6
	+VAR	1427.2	62.6	34.3	28.0	54.8	36.4	87.0
	+SinkF	1484.6	62.5	35.9	26.0	**56.0**	35.4	**87.3**
	+Ours	**1490.7**+3.5	**62.7**+0.3	**36.8**+1.1	**28.7**+1.4	**56.0**+1.1	**36.7**+1.2	**87.3**+0.7
Vicuna-v1.5-13B	-	1536.2	**63.9**	35.1	33.3	54.6	37.0	**87.5**
	+VAR	1498.8	63.6	34.8	32.7	54.5	36.9	87.0
	+SinkF	1533.6	**63.9**	35.0	**35.3**	55.0	36.7	**87.5**
	+Ours	**1598.2**+62.0	63.7-0.2	**35.2**+0.1	32.7-0.6	**56.6**+2.0	**37.4**+0.4	**87.5**+0.0
Qwen2.5-7B-Instruct	-	1513.5	62.2	44.3	**36.7**	54.1	44.4	87.5
	+VAR	1513.4	62.2	44.4	**36.7**	**55.6**	**45.0**	87.5
	+SinkF	1515.0	62.4	44.0	**36.7**	55.2	44.9	87.4
	+Ours	**1533.4**+19.9	**62.5**+0.3	**44.8**+0.5	**36.7**+0.0	55.4+1.3	42.9-1.5	**87.6**+0.1
Phi3-3.8B	-	1448.6	60.7	40.0	30.7	**56.0**	37.8	86.7
	+VAR	1442.9	60.7	39.9	30.7	55.8	37.7	**86.8**
	+SinkF	1457.1	60.5	39.3	29.3	55.2	37.4	86.5
	+Ours	**1481.6**+33.0	**60.9**+0.2	**40.6**+0.6	**31.3**+0.6	55.7-0.3	**39.3**+1.5	86.3-0.4

5.1 Multimodal Pre-trained Model Evaluation

Enhanced fine-grained alignment and caption quality. We probe the visual tokens to obtain their corresponding text token. As shown in Figure 5, ViSAlign builds better patch-level correlations, enabling more visual patches to align with their corresponding textual semantics. Beyond qualitative analysis, we evaluate the captioning capability of pre-trained models on the COCO2014 validation set (Lin et al., 2014) using BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), and CIDEr (Vedantam et al., 2015) metrics. As shown in Table 1, ViSAlign consistently outperforms the baseline across all metrics, indicating that our method improves the projector's learning and enhances vision-language alignment. Implementation details are provided in Appendix C.2.

Modified cross-modal attention flow. To assess whether ViSAlign encourages the projector to learn fine-grained visual information acquisition, we repeat the forward pass in Section 3.2. As shown in Figure 5, ViSAlign redirects attention from background visual sinks to text-relevant visual evidence. We further adopt pointing-in-box (PiB) accuracy (Zhang et al., 2018; Xiao et al., 2017; Shi et al., 2026; Li et al., 2025) to quantify whether cross-modal attention falls within the evidential visual regions on RefCOCO (Kazemzadeh et al., 2014), RefCOCO+ (Kazemzadeh et al., 2014), and RefCOCOg (Mao et al., 2016). As shown in Table 2, ViSAlign shows superior performance to the baseline across all sets, demonstrating a stronger ability to allocate attention to the target regions. Implementation details are provided in Appendix C.3, and more visualizations are provided in Appendix D.2.

5.2 Supervised Fine-Tuning Model Evaluation

Evaluations and compared methods. We conduct comprehensive evaluation across diverse multimodal understanding benchmarks, consisting of (1) general VQA: MME-P (Fu et al., 2026), GQA (Hudson & Manning, 2019), and MMMU (Yue et al., 2024); (2) vision-centric benchmarks: MMVP (Tong et al., 2024), RealWorldQA (RWQA) (xAI, 2024), and MMStar (MM*) (Chen et al., 2024); (3) hallucination benchmark: POPE (Li et al., 2023b). Besides, we select other visual sink-aware modulation methods for comparison, including inference-time methods VAR (Kang et al.,

Table 4: Comparison with the trainable sink modulation method LSG. "-" indicates not supported by the official implementation. The best results are highlighted in bold.

Method	General			Vision-Centric			Hallu
	MME-P	GQA	MMMU	MMVP	RWQA	MM*	POPE
Baseline	1487.2	62.4	35.7	27.3	54.9	35.5	86.6
LSG(L10)	1420.9	61.5	-	-	49.5	35.9	86.9
ViSAlign	**1490.7**	**62.7**	**36.8**	**28.7**	**56.0**	**36.7**	**87.3**

2025) and Sink-to-the-front (SinkF) (Luo et al., 2026), as well as training-based methods LSG (Choi et al., 2026). Details can be found in Appendix C.4.

Main results. As shown in Table 3, ViSAlign, *which only enhances the learning of approximately 20M trainable parameters for a 7B model*, yields broad performance improvements across diverse LLM backbones and benchmarks. **Firstly**, ViSAlign consistently improves the performance of the widely adopted Vicuna-v1.5-7B backbone. Similar positive trends on both general and vision-centric tasks are also observed with other LLM backbones. ViSAlign improves the MME score of Vicuna-v1.5-13B from 1536.2 to 1598.2, RealWorldQA of Qwen2.5-7B-Instruct from 54.1 to 55.4, and MMStar of Phi3-3.8B from 37.8 to 39.3. These performance gains confirm that boosting vision-language alignment in pre-training leads to superior downstream task performance. **Secondly**, the performance gains on POPE fluctuate, which may be attributed to ViSAlign not engaging in parameter updates of the LLM backbone, whereas hallucinations largely stem from intrinsic statistical biases of LLM (Wang et al., 2026b). **Thirdly**, ViSAlign achieves superior performance across most settings among the compared methods (Table 3 and 4). As these methods mainly modify the visual information processing in MLLMs, ViSAlign focuses on aligning the representation by guiding projector learning, thereby providing a stronger representation foundation for downstream tasks.

5.3 Ablation Study

To verify the design of ViSAlign, we compare different masking strategies, including random masking and targeted masking of L-sinks and V-sinks. As shown in Table 5, random masking degrades performance by indiscriminately disrupting useful visual information, while mask-L performs close to ViSAlign and consistently outperforms mask-V. This result is consistent with the observation in Section 3.3 that L-sinks dominate the visual sink behavior. However, explicitly identifying visual sinks requires prior knowledge of activation dimensions and manually defined activation thresholds, both of which may vary across LLM backbones and training processes. ViSAlign instead identifies visual sinks directly from the T2I attention distribution, providing a simpler and more model-agnostic intervention without relying on backbone-specific activation patterns. Furthermore, we provide ablation studies on model architectures by changing the vision encoder to SigLIP2-ViT-SO400M/14@384 (Tschannen et al., 2025) or changing the cross-modal projector to C-Abstractor (Cha et al., 2024) in Appendix D.1

Table 5: Ablation study over mask strategies.

Ablation	MME-P	MMMU	MMVP	MM*	POPE
Baseline	1487.2	35.7	27.3	35.5	86.6
Random	1477.6	36.7	25.3	34.2	86.3
mask-L	1489.9	35.7	**29.3**	35.1	**87.5**
mask-V	1449.2	36.0	28.0	34.3	87.2
ViSAlign	**1490.7**	**36.8**	28.7	**36.7**	87.3

6 Discussion and Limitations

In this work, we investigate cross-modal information flow during multimodal pre-training of MLLMs, revealing the layer-wise evolution of attention and the emergence and effects of visual attention sinks. However, the mechanisms underlying their formation remain unclear. While V-sinks may arise from their higher feature norms and global visual information, the origin of L-sinks is less understood. Notably, their emergence closely coincides with the rapid decrease in training loss, suggesting a possible connection to model optimization. One possible explanation is that L-sinks serve as an optimization shortcut when visual information is not yet effectively processed at the early stage of training, allowing the model to preserve textual representations while suppressing fine-grained visual information in deeper layers. Future work could further investigate the role of visual sinks throughout both multimodal pre-training and post-training.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, clean and reformat datasets, support qualitative and thematic data analysis, interpret results, cleaning and reformatting the dataset. We have reviewed all AI-assisted work. We manually check and polish the translation from AI tools for better clarity. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL https://arxiv.org/abs/2404.14219.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

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

Federico Barbero, Alvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Michael Bronstein, Petar Veličković, and Razvan Pascanu. Why do llms attend to the first token? *arXiv preprint arXiv:2504.02732*, 2025.

Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. Honeybee: Locality-enhanced projector for multimodal llm. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13817–13827. IEEE, 2024.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language

models? In *Advances in Neural Information Processing Systems*, volume 37, pp. 27056–27087, 2024.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Jiho Choi, Jaemin Kim, Sanghwan Kim, Seunghoon Hong, and Jin-Hwi Park. When sinks help or hurt: Unified framework for attention sink in large vision-language models. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2026.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International conference on learning representations*, volume 2024, pp. 2632–2652, 2024.

Yingqi Fan, Junlong Tong, Anhao Zhao, and Xiaoyu Shen. What do visual tokens really encode? uncovering sparsity and redundancy in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11987–11997, 2026.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38, 2026.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235, 2023.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. In *International Conference on Learning Representations*, volume 2025, pp. 97114–97144, 2025.

Tianyu Guo, Druv Pai, Yu Bai, Jiantao Jiao, Michael I Jordan, and Song Mei. Active-dormant attention heads: Mechanistically demystifying extreme-token phenomena in llms. *arXiv preprint arXiv:2410.13835*, 2024.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6700–6709, 2019.

Jiachen Jiang, Jinxin Zhou, Bo Peng, Xia Ning, and Zhihui Zhu. Analyzing fine-grained alignment and enhancing vision understanding in multimodal language models. *Advances in Neural Information Processing Systems*, 38:874–899, 2026a.

Nicholas Jiang, Amil Dravid, Alexei Efros, and Yossi Gandelsman. Vision transformers don't need trained registers. *Advances in neural information processing systems*, 38:56557–56595, 2026b.

Zhehan Kan, Xin Li, Yanlin Liu, Xiaochen Yang, Xinghua Jiang, Yinsong Liu, Deqiang Jiang, Xing Sun, Qingmin Liao, and Wenming Yang. Rar: Reversing visual attention re-sinking for unlocking potential in multimodal large language models. In *The Fourteenth International Conference on Learning Representations*, 2026.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *International Conference on Learning Representations*, volume 2025, pp. 87676–87703, 2025.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In Alessandro Moschitti, Bo Pang, and Walter Daelemans (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 787–798, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1086. URL https://aclanthology.org/D14-1086.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023a.

Weiming Li, Yan Shao, Jing Yang, Yujing Lu, Ling Zhong, Yuhan Wang, Min Yu, Tongxiao Ruan, and Manni Duan. How auxiliary reasoning unleashes gui grounding in vlms. *arXiv preprint arXiv:2509.11548*, 2025.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, 2023b.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

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

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

Jiayun Luo, Wan-Cyuan Chris Fan, Lyuyang Wang, Xiangteng He, Tanzila Rahman, Purang Abolmaesumi, and Leonid Sigal. To sink or not to sink: Visual information pathways in large vision-language models. In *International Conference on Learning Representations*, volume 2026, pp. 93626–93660, 2026.

Guangtao Lyu, Qi Liu, Chenghao Xu, Jiexi Yan, Muli Yang, Xueting Li, Fen Fang, and Cheng Deng. Revealing and enhancing core visual regions: Harnessing internal attention dynamics for hallucination mitigation in lvlms. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 15214–15230, 2026.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11–20, 2016.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. In *International Conference on Learning Representations*, volume 2025, pp. 57172–57189, 2025.

OpenAI. gpt-oss-120b gpt-oss-20b model card, 2025. URL `https://arxiv.org/abs/2508.10925`.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Cheng Shi, Yizhou Yu, and Sibei Yang. Vision transformers need more than registers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26328–26337, 2026.

Yan Shu, Hangui Lin, Yexin Liu, Yan Zhang, Gangyan Zeng, Yan Li, Yu Zhou, Ser Nam Lim, Harry Yang, and Nicu Sebe. When semantics mislead vision: Mitigating large multimodal models hallucinations in scene text spotting and understanding. *Advances in Neural Information Processing Systems*, 38:68952–68980, 2026.

Mustafa Shukor and Matthieu Cord. Implicit multimodal alignment: On the generalization of frozen llms to multimodal inputs. *Advances in Neural Information Processing Systems*, 37:130848–130886, 2024.

Zunhai Su, Hengyuan Zhang, Wei Wu, Yifan Zhang, Yaxiu Liu, He Xiao, Qingyao Yang, Yuxuan Sun, Rui Yang, Chao Zhang, et al. Attention sink in transformers: A survey on utilization, interpretation, and mitigation. *arXiv preprint arXiv:2604.10098*, 2026.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*, 2024.

Jianting Tang, Yubo Wang, Haoyu Cao, and Linli Xu. Basic: Boosting visual alignment with intrinsic refined embeddings in multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20582–20592, 2025.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL `https://qwenlm.github.io/blog/qwen2.5/`.

Qwen Team. Qwen3.5: Accelerating productivity with native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

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

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

Enguang Wang, Qiang Wang, Yuanchen Wu, Ke Yan, Xinbin Yuan, Shouhong Ding, Xialei Liu, and Ming-Ming Cheng. Predictive regularization against visual representation degradation in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8813–8824, 2026a.

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

Zhenyu Wang. Logitlens4llms: Extending logit lens analysis to modern large language models. *arXiv preprint arXiv:2503.11667*, 2025.

Zichuan Wang, Songlin Yang, Bo Peng, Zhenchen Tang, Yang Li, Beibei Dong, and Jing Dong. Same attention, different truths: Put logit-lens over visual attention to detect and mitigate lvlm object hallucination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25315–25325, 2026b.

xAI. Grok-1.5 vision preview. `https://x.ai/news/grok-1.5v`, April 2024. Accessed: 2026-05-20.

Fanyi Xiao, Leonid Sigal, and Yong Jae Lee. Weakly-supervised visual grounding of phrases with linguistic structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5945–5954, 2017.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024.

Tianyu Yu, Kechen Fang, Zihao Wan, Kaidong Zhang, Yicheng Zhang, Jun Song, Bo Zheng, and Yuan Yao. Deep pre-alignment for vlms. *arXiv preprint arXiv:2605.15300*, 2026.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126 (10):1084–1102, 2018.

Xiaofeng Zhang, Yihao Quan, Chen Shen, Chaochen Gu, Xiaosong Yuan, Shaotian Yan, Jiawei Cao, Hao Cheng, Kaijie Wu, and Jieping Ye. Shallow focus, deep fixes: Enhancing shallow layers vision attention sinks to alleviate hallucination in lvlms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3512–3534, 2025a.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19781–19791, 2025b.

A Visualization of T2I Attention on Fine-grained Text-related Visual Evidence

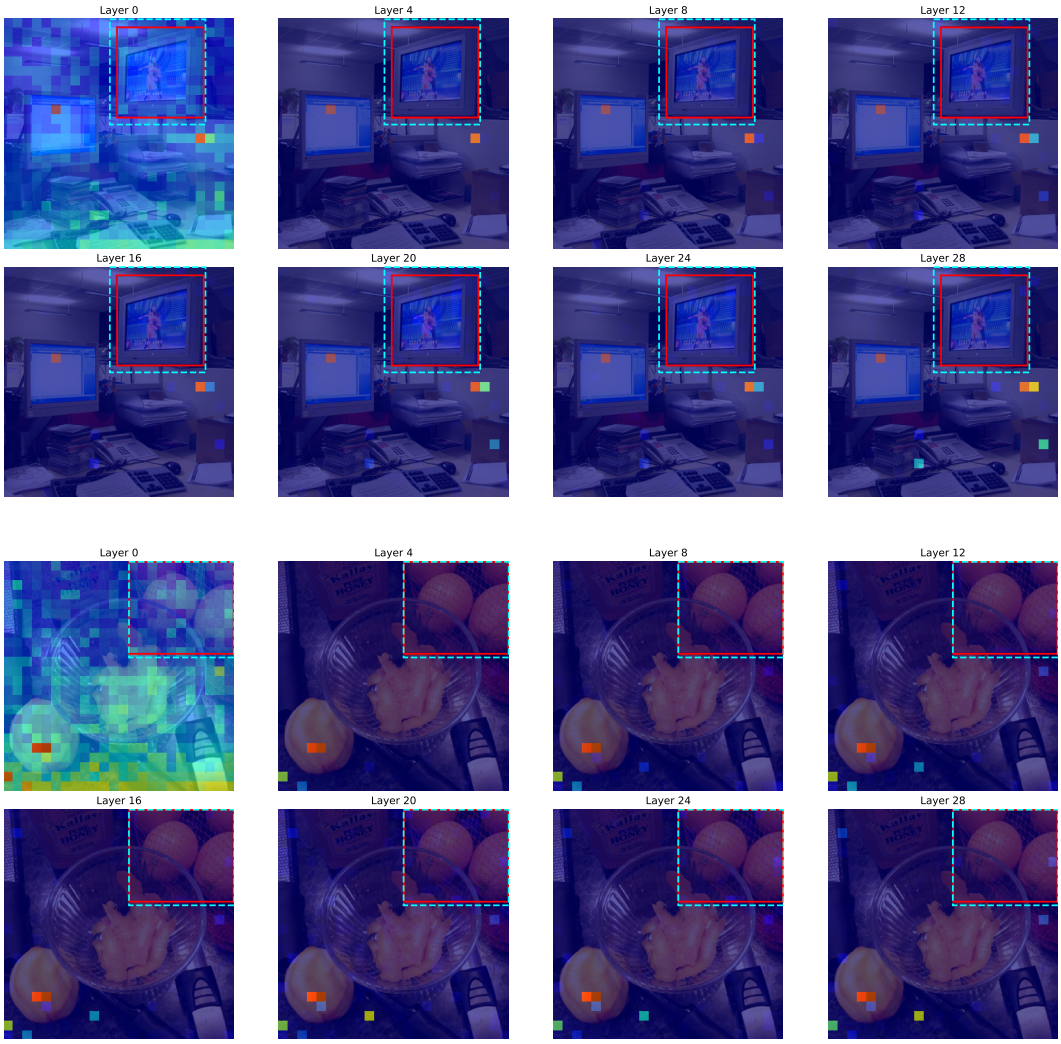


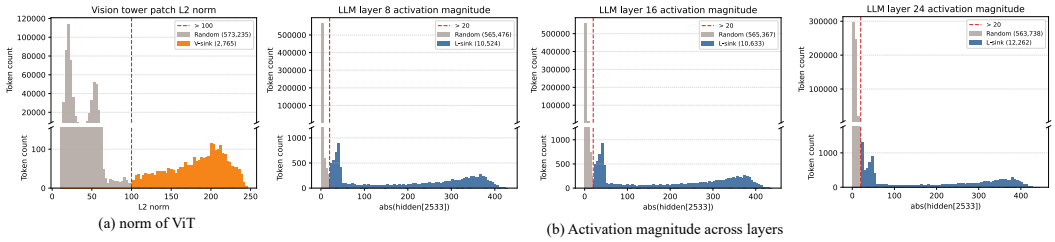
Figure 6: Visualizations of T2I attention across layers on the curated dataset based on RefCOCO (Kazemzadeh et al., 2014). Only the localized caption for the bbox region is used for visualization. The red lines represent the original pixel-level bbox, and the blue lines represent the bbox aligned to each patch.

We visualize the T2I attention on our curated dataset with global and local descriptions of images. Specifically, we focus on the local description and overlay the T2I attention and bounding box on the original image. As shown in Figure 6, the T2I attention is overwhelmingly absorbed by visual sinks while ignoring the bounding box region across all model layers. In contrast with previous conclusions on instruction-tuned models, which observed the T2I attention concentrating on the task-specific text-relevant visual evidence within the middle layers of MLLMs.

B Implementation Details for Identifying Visual Sinks

B.1 Attention Sink Definition

Recent studies (Kang et al., 2025; Luo et al., 2026; Choi et al., 2026) have demonstrated that tokens in Transformer-based models can exhibit abnormally high activation values in specific dimensions



(a) norm of ViT

(b) Activation magnitude across layers

Figure 7: (a) Distribution of the L_2 norms of the output features from the vision encoder. (b) Distribution of dimension-wise activation magnitudes across LLM layers.

or unusually large L_2 feature norms. Based on these characteristics, the sink tokens at layer l are identified as

$$\mathcal{I}^l_{\text{sink}} = \{i \in \mathcal{I} \mid \phi(x^l_i) \geq \tau\}, \tag{10}$$

where $\phi(\cdot)$ denotes the sink characteristic function and τ is a predefined threshold. According to previous works, the sink characteristic function could vary from activation magnitude on a specific dimension (Sun et al., 2024), the L_2 norm (Darcet et al., 2024), and the attention score (Kang et al., 2025). We conduct the following analyses using the pre-trained model of LLaVA-1.5-7B (Liu et al., 2024) on 1,000 randomly sampled image–text pairs from the LLaVA-Pretrain-558K dataset. For each sample, the image is represented by $24 \times 24 = 576$ visual tokens.

B.2 Identifying V-Sinks via the L_2 Norm of ViT Hidden States

Following previous analyses of ViTs (Darcet et al., 2024), we use the L_2 feature norm as the sink characteristic function for ViT hidden states:

$$\phi(x^l_i) = \|x^l_i\|_2. \tag{11}$$

We analyze the feature norm distribution of the output hidden states of the ViT, specifically the 2nd-to-last layer. As shown in Figure 7(a), the distribution exhibits a clear separation between V-sinks and ordinary visual tokens. We therefore set $\tau = 100$ as a robust threshold for identifying V-sinks. Each image contains approximately $2 \sim 4$ V-sinks, which is consistent with the observations of previous studies (Luo et al., 2026; Choi et al., 2026).

B.3 Identifying L-Sinks via Activation Dimensions

Following previous analyses of MLLMs (Luo et al., 2026), we use the activation magnitude in dimension 2533 as the sink characteristic function for visual tokens in the LLM backbone:

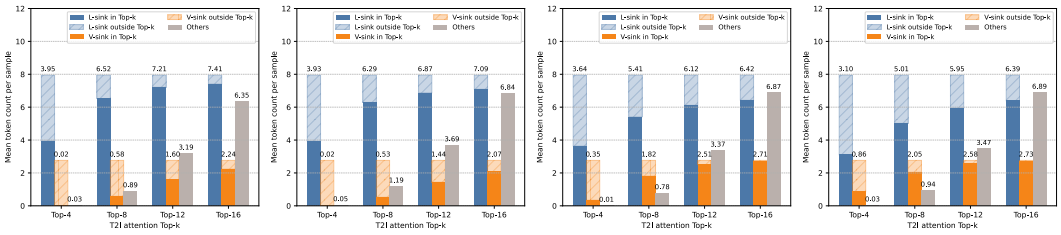
$$\phi(x^l_i) = \|x^l_i[d]\|_1, \tag{12}$$

where $d = 2533$.

We analyze the distribution of activation values across different layers of the LLM. As shown in Figure 7(b), a clear separation between L-sinks and ordinary visual tokens emerges from layer 8 to layer 24. We therefore set $\tau = 20$ as the threshold for identifying L-sinks within these layers. On average, approximately $7 \sim 10$ L-sinks emerge in the LLM for each image.

B.4 Identifying Visual Attention Sinks via Top-K Text-to-image Attention

Previous studies (Kang et al., 2025) also identify visual sinks simply by selecting the top-k visual tokens with the highest T2I attention. Therefore, we conduct an additional study to examine the consistency between this attention-based criterion and the activation-based identification of L-sinks and V-sinks discussed above. Using the same pre-trained model and 1,000 samples, we reproduce the above identification procedure and identify L-sinks based on their characteristic activation dimensions at layer 16. We then measure, on average, how many identified L-sinks and V-sinks are



(a) layer 4 (b) layer 12 (c) layer 20 (d) layer 28

Figure 8: Coverage of visual sinks identified by activation patterns and feature norms among the top-k visual tokens ranked by T2I attention across layers.

included among the top-k T2I-attended visual tokens at selected layers, with $k \in \{4, 8, 12, 16\}$ and layers $\{4, 12, 20, 28\}$.

As shown in Figure 8, although the top-k T2I attention indices do not strictly coincide with the identified L-sinks and V-sinks, they achieve substantial coverage. Specifically, given approximately 8 L-sinks and 3 V-sinks per sample on average, the top-k attention indices cover up to 7 out of 8 L-sinks and 2 out of 3 V-sinks on average. Across different values of k, L-sinks consistently dominate the top-k selections for $k \in \{4, 8\}$, indicating that most L-sinks receive the highest T2I attention scores. As k increases to $\{12, 16\}$, the number of covered V-sinks and other visual tokens gradually increases, suggesting that V-sinks are more frequently ranked between the top-8 and top-16 positions. We further observe a clear layer-wise evolution: L-sinks dominate the highest-attention positions in relatively shallow layers $\{4, 12\}$, whereas the coverage of V-sinks increases in deeper layers $\{20, 28\}$ and can eventually reach full coverage. This trend is consistent with our observations in Section 3.3. In summary, although T2I attention-based identification does not strictly coincide with L-sinks and V-sinks, it provides a reliable and simpler approximation for identifying visual sinks.

C Implementation Details of Experiments

C.1 Training Details

Adopting the standard two-stage training paradigm of LLaVA-1.5 (Liu et al., 2024), our pipeline comprises a pre-training phase followed by instruction tuning. During pre-training, we freeze both the LLM backbone and the vision encoder, keeping only the randomly initialized projector weights trainable. The model is trained on the 558K subset of LAION-CC-SBU with a batch size of 256 and a learning rate of 1e-3, following LLaVA's configuration. ViSAlign is applied exclusively in this stage. For instruction tuning, we perform supervised fine-tuning with both the LLM backbone and cross-modal projector unfrozen, using the LLaVA-1.5 instruction dataset with a batch size of 128 and a learning rate of 2e-5, consistent with the original LLaVA-1.5 settings.

C.2 Captioning Capability Evaluations

In this experiment, we adopt the COCO2014 (Lin et al., 2014) validation set as our evaluation dataset, which comprises 40,504 image–caption pairs. For each sample, we feed the original image into the model and use the plain conversation mode from the pre-training stage of LLaVA-1.5 to prompt it to generate a corresponding caption. After recording the generated captions for all images, we employ the pycocoevalcap toolkit to compute BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), and CIDEr (Vedantam et al., 2015) scores between the ground-truth captions and the generated ones.

C.3 Pointing-in-box Accuracy Implementation

Pointing-in-box accuracy, also referred to as pointing game accuracy, is a widely adopted metric for directly assessing the grounding capability of models based on attention distributions (Zhang et al., 2018; Xiao et al., 2017; Shi et al., 2026; Li et al., 2025). In our experiments, we evaluate on three

standard referring expression grounding benchmarks, namely RefCOCO (Kazemzadeh et al., 2014), RefCOCO+ (Kazemzadeh et al., 2014), and RefCOCOg (Mao et al., 2016). We use the processed dataset provided by lmms-lab on huggingface. Each sample in these datasets consists of a question, an answer, and a bounding box (bbox) location. Typically, the question is a prompt such as "please describe the bbox," and the answer provides the corresponding descriptive annotation for the bbox.

To compute the metric, we construct an image-caption pair using the input image and the answer, and feed it into the model for a forward pass. We then monitor the cross-modal attention and compute the averaged text-to-image (T2I) attention over the caption tokens. From the resulting attention map, we select the top-k indices with the highest attention scores, where $k = 5$. A prediction is considered a hit if any of these selected indices falls within the ground-truth bounding box. Finally, we report the overall accuracy across the entire benchmark.

C.4 Reproduction of Compared Methods

VAR (Kang et al., 2025). We run the official implementation and change the L-sinks dimension according to the massive activation dimension of Qwen2.5-7B-Instruct and Phi3-mini-4k-instruct.

Sink-to-the-front (Luo et al., 2026). We implement it as described in the paper, keeping the positional encoding unchanged and reordering only the visual tokens. We use the L_2-norm to identify the V-sinks according to Appendix B.

LSG (Choi et al., 2026). We run the official training and evaluation implementation on the reproduced LLaVA-1.5 checkpoint, using the official scripts and configs to train the two-layer MLP gate module at layer 10.

D Further Results

D.1 Ablation study on vision encoder and projector

Table 6: Performance comparison with and without ViSAlign when adopting MLP/SigLIP2 or C-Abstractor/CLIP as the vision encoder and the cross-modal projector.

Projector/Encoder	ViSAlign	General			Vision-Centric			Hallu
		MME	GQA	MMMU	MMVP	RWQA	MM*	POPE
MLP/SigLIP2	-	1505.5	64.6	36.3	38.7	58.2	38.1	87.3
MLP/SigLIP2	✓	1535.0$_{+29.5}$	64.5$_{-0.1}$	36.9$_{+0.6}$	40.7$_{+2.0}$	58.3$_{+0.1}$	37.9$_{-0.2}$	87.6$_{+0.3}$
C-Abs/CLIP	-	1466.3	60.2	37.2	24.7	53.2	33.9	84.8
C-Abs/CLIP	✓	1472.5$_{+6.2}$	60.8$_{+0.6}$	36.9$_{-0.3}$	22.7$_{-2.0}$	54.0$_{+0.8}$	34.3$_{+0.4}$	85.8$_{+1.0}$

To further assess whether the effectiveness of ViSAlign generalizes across different vision encoders and cross-modal projectors, we conduct additional ablation studies under two architectural variants. First, we replace the commonly used CLIP vision encoder with SigLIP2-ViT-SO400M/14@384 (Tschannen et al., 2025), while retaining Vicuna-v1.5-7B as the LLM backbone and an MLP projector. Second, we replace the MLP projector with a C-Abstractor (Cha et al., 2024) while keeping the Vicuna-v1.5-7B backbone and CLIP vision encoder unchanged. As shown in Table 6, ViSAlign generally maintains or improves performance under both architectural variants. With SigLIP2, ViSAlign improves MME by 29.5 points and MMVP by 2.0%. With C-Abstractor, it improves MME by 6.2 points and POPE by 1.0%, despite minor fluctuations on several benchmarks. This suggests that the benefits of ViSAlign are not tied to a specific model architecture, but reflect a general improvement in cross-modal alignment.

D.2 Further Visualizations of Modified Attention Flow

We observe the cross-modal attention following Section 3.2 on more samples. We select layer 16 for visualization, and results are shown in Figure 9.

The image shows a dish of barbecue ribs in a rectangular baking dish. The ribs are generously coated with a glossy, reddish-brown barbecue sauce. The dish is placed on a table with a pink napkin underneath and a small pitcher of sauce or condiment to the side. The overall presentation suggests a home-cooked



Image Baseline ViSAlign

The image is a book cover for \"Willows in the Wind\" by John Stevens. The cover features a serene landscape with a moonlit sky, a calm body of water, and trees in the background. The title and author's name are prominently displayed in the center of the cover.



Image Baseline ViSAlign

The image shows a person walking on a red carpet, wearing a white dress with blue floral patterns and black ankle-strap heels. The background is filled with photographers taking pictures.



Image Baseline ViSAlign

Figure 9: Further visualizations of how ViSAlign modifies the T2I attention flow.