Why Is Spatial Reasoning Hard for VLMs? An Attention Mechanism Perspective on Focus Areas

Shiqi Chen [1] **Tongyao Zhu** [2] **Ruochen Zhou** [1] **Jinghan Zhang** [3] **Siyang Gao** [1] **Juan Carlos Niebles** [4 5] **Mor Geva** [6] **Junxian He** [3] **Jiajun Wu** [4] **Manling Li** [4 7]

Abstract

Large Vision Language Models (VLMs) have long struggled with spatial reasoning tasks. Surprisingly, even simple spatial reasoning tasks, such as recognizing "under" or "behind" relationships between only two objects, pose significant challenges for current VLMs. In this work, we study the spatial reasoning challenge from the lens of **mechanistic interpretability**, diving into the model's internal states to examine the interactions between image and text tokens. By tracing attention distribution over the image throughout intermediate layers, we observe that successful spatial reasoning correlates strongly with the model's ability to align its attention distribution with actual object locations, particularly differing between familiar and unfamiliar spatial relationships. Motivated by these findings, we propose ADAPTVIS based on inference-time confidence scores to sharpen the attention on highly relevant regions when confident, while smoothing and broadening the attention window to consider a wider context when confidence is lower. This training-free decoding method shows significant improvement (e.g., up to a 50 absolute point improvement) on spatial reasoning benchmarks such as WhatsUp and VSR with negligible cost. We make code and data publicly available for research purposes at https://github.com/shiqichen17/AdaptVis.

[1]City University of Hong Kong [2]National University of Singapore [3]Hong Kong University of Science and Technology [4]Stanford University [5]Salesforce Research [6]Tel Aviv University [7]Northwestern University. Correspondence to: Shiqi Chen <schen438-c@my.cityu.edu.hk>, Manling Li <manling.li@northwestern.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

1. Introduction

Despite rapid advancements in Large Vision-Language Models (VLMs), a significant deficiency persists, i.e., their struggle with vision-centric abilities (Gao et al., 2023; Kamath et al., 2023; Tong et al., 2024a; Chen et al., 2024a). This limitation is particularly notable in spatial reasoning given the simplicity of the task. Spatial reasoning involves inferring basic relationships between just two objects, such as "*left*", "*right*", "*above*", "*below*", "*behind*", or "*front*", as shown in Figure 1. For example, given the image with a book "*behind*" the candle in Figure 1, VLMs describe the book as being "*left*" of the candle. This error is not an isolated incident but a frequent, recurring pattern, highlighting a fundamental bottleneck in how VLMs process visual-centric information.

Recent studies have probed the limitations of vision encoders like CLIP (Radford et al., 2021) in VLMs' vision processing (Tong et al., 2024b;a), yet a critical aspect remains underexplored: how vision and text tokens interact within the model's internal states to construct geometric understanding. Specifically, the model must simultaneously identify objects while maintaining awareness of the broader geometric context, grasping the complex geometric relationships between them. This geometric understanding fundamentally manifests in how models distribute their attention across the visual tokens. This unique challenge makes spatial reasoning an ideal lens for studying how VLMs internally process vision-centric information.

Therefore, we open up the black box of VLMs and examine their internal mechanisms through a suite of carefully designed spatial reasoning tasks. By looking into the attention distributions, we can systematically investigate how vision and text tokens interact to construct, or fail to construct, accurate spatial understanding. Our investigation begins with a key observation: despite image tokens comprising around 90% of the input sequence, they receive only about 10% of the model's attention. This significant imbalance suggests that textual priors often overshadow visual evidence, explaining VLMs' struggles with vision-centric tasks.

While a straightforward solution might seem to be simply increasing attention to visual tokens, our deeper analysis re-

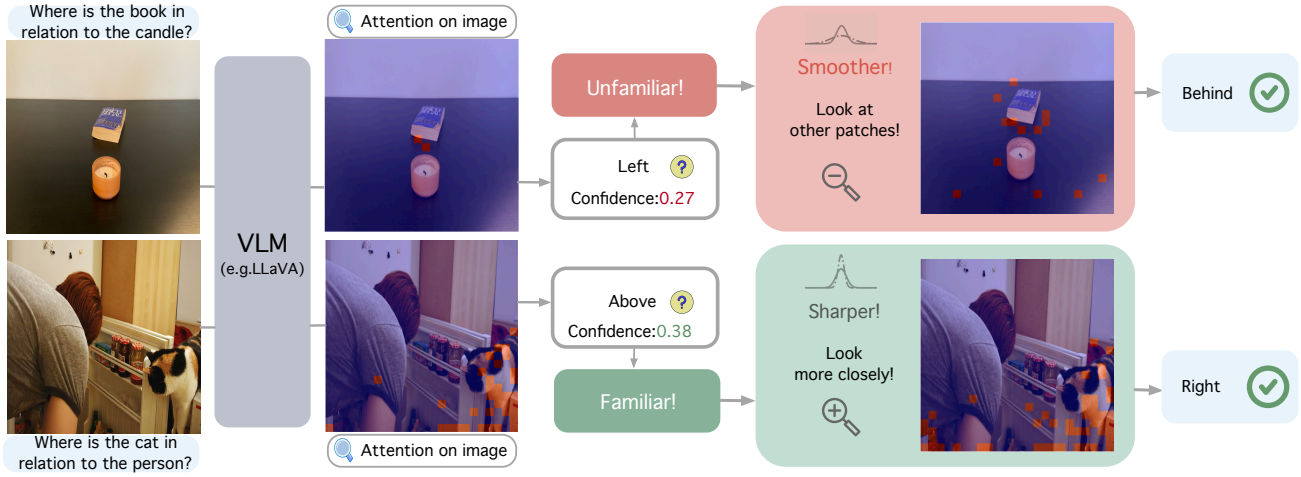

Figure 1. The framework of AdaptVis. We adaptively intervene in the temperature of the attention logits of the image tokens. Top: For generations with **low** confidence, we **smoothen** the attention distribution to broaden the context window for better concentration on the correct objects. Bottom: For generations with **high** confidence, we trust the attention pattern and **sharpen** the attention distribution.

veals that the challenge lies not just in the quantity but in the geometric distribution of visual attention. By examining attention patterns across model layers, we observe a pattern in VLMs' attention behavior: their initial attention distribution often reflects learned priors that may or may not align with the actual geometric distribution over the image. This led us to a key insight: rather than accepting this initial distribution, we can dynamically intervene based on the model's self-belief, which is measured using its confidence score. As shown in Figure 1, when the model exhibits high confidence in its spatial reasoning (as measured by generation probability), we sharpen its attention distribution to strengthen the focus on its current beliefs. Conversely, when confidence is low, we smooth the attention distribution to encourage the exploration of alternative spatial relationships.

We call this confidence-guided attention intervention AdaptVis, which proves remarkably effective while remaining computationally efficient. In experiments across diverse spatial reasoning benchmarks, including WhatsUp (Kamath et al., 2023) and VSR (Liu et al., 2023), our approach achieves substantial improvements of up to 50% points. These gains are observed across both synthetic datasets with clean backgrounds and real-world images with complex scenes, demonstrating the robustness of our method.

Our visualization of this intervention strategy reveals its underlying mechanism: by dynamically adjusting attention patterns, we effectively guide the model's focus to better align with actual object locations and their spatial relationships. This suggests that successful spatial reasoning isn't just about having the right attention mechanism, but about having the right confidence to know when to trust or question one's initial spatial understanding.

2. Preliminary on VLMs

We center our analysis on investigating how VLMs distribute their attention over image tokens, aiming to gain deeper insights into spatial reasoning errors.

Notation Large VLMs like LLaVA (Liu et al., 2024a) consist of three components: a visual encoder like CLIP (Radford et al., 2021), a pretrained language model, and a projector to connect them. The visual encoder functions as a perception tool to "see" the image, while the image information is processed through a projector to be mapped into the token space. The LLM is often based on the transformer architecture, consisting of L layers stacked together. Each layer consists of two major components: a Multi-Head Attention (MHA) module, followed by a feed-forward network. For each layer l, given the input $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ (where n is the number of tokens and d is the embedding dimension), MHA performs self-attention in each head N_h. The output is a concatenation of all heads' outputs: $\text{MHA}^{(l)}(\boldsymbol{X}) = \text{Concat}\left(\boldsymbol{N}_h^{(l,1)}, \ldots, \boldsymbol{N}_h^{(l,H)}\right)\boldsymbol{W}_o$. Here H is the number of heads; $\boldsymbol{N}_h$ is the output of head N_h:

$$\boldsymbol{N}_h^{(l,h)} = \text{Softmax}(\boldsymbol{A}^{(l,h)})\boldsymbol{V} = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d_h}} + \boldsymbol{M}\right)\boldsymbol{V}, \tag{1}$$

where attention logits $\boldsymbol{A}^{(l,h)}$ are computed via $\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}_{q_h}, \boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}_{k_h}, \boldsymbol{V} = \boldsymbol{X}\boldsymbol{W}_{v_h}$, and $\boldsymbol{W}_{q_h}, \boldsymbol{W}_{k_h}, \boldsymbol{W}_{v_h} \in \mathbb{R}^{d\times d_h}$ are learnable projection matrix of the head N_h. A causal mask $\boldsymbol{M}_{ij} = 0$ if $i \geq j$, and $-\infty$ otherwise, prevents tokens from attending to future tokens.

Dataset	Num of Options		Num of Objects		Source	
	4 options	6 options	1 object	2 objects	Syn	Real
Cont_A	✓			✓	✓	
Cont_B	✓			✓	✓	
VG_one		✓	✓			✓
VG_two		✓		✓		✓
COCO_one	✓		✓			✓
COCO_two	✓			✓		✓

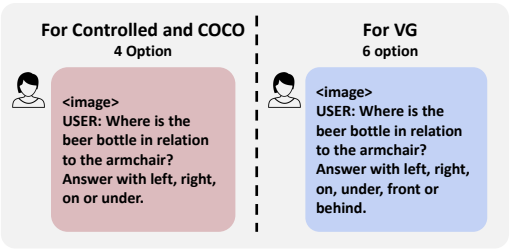

Figure 2. Left: Choice counts, object counts and data source by subset in WhatsUp ("Syn" = Synthetic, "Real" = Real). Right: Evaluation prompts we use in evaluation.

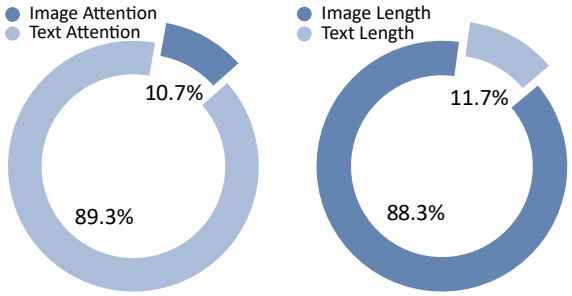

Figure 3. A striking imbalance between visual and textual attention: while image tokens take approximately 90% of the sequence length, they receive only about 10% of the model's total attention on WhatsUp. This severe disparity in attention allocation suggests that VLMs fundamentally underutilize visual information.

3. Text-Vision Attention Interactions

We hypothesize that the failure on spatial reasoning stems from issues in the text-vision interactions, specifically due to insufficient attention being allocated to the image tokens. In this section, we systematically investigate the impact of the absolute values of attention logits on the image tokens during spatial reasoning.

Experiment settings. We select a widely-used spatial reasoning benchmark WhatsUp (Kamath et al., 2023) since it contains both synthetic data and realistic data. The synthetic data (**Controlled_Image**) features clean backgrounds with two objects, as shown in the upper example of Figure 1. It comprises two subsets: **Controlled_A**, with one large object (e.g., table) and one small object (e.g., cup), and **Controlled_B** with two small objects (e.g., book and plate). We use **Cont_A** and **Cont_B** as abbreviations in the paper. The realistic data, as shown in the lower image of Figure 1, contains complex backgrounds with multiple objects, sourced from MS COCO (Lin et al., 2014) and Visual Genome (Krishna et al., 2017) (referred to as **COCO** and **VG** later, both datasets include captions involving either one object or two objects, referred to as **COCO_one**, **COCO_two**, **VG_one**, and **VG_two**, respectively). Compared with realistic images, the synthetic images enables clearer observation of VLMs's inner workings with just two objects. Each image is paired with a ground truth caption describing the spatial relationship of two objects.

We reformat the original $\langle image, caption \rangle$ setting into a generative question-answering setting $\langle image, question, spatial\_label \rangle$, enabling evaluation of generative models like VLMs and tracing of internal states. Questions are generated using GPT-4 (OpenAI, 2024). Details in each subset with prompts are shown at Figure 2. This QA task asks about positions in a multiple-choice format, with possible answers like "left", "right", "on" and so on.

For evaluation, we use LLaVA-1.5 (Liu et al., 2024a) throughout the analysis part (Section 3 and 4). For metric, we apply **accuracy** of exact match as the primary metric. To maintain consistency in the label space across datasets, we use a four-option setting $\langle left, right, on, under \rangle$ for the Controlled_Image and COCO subsets, and a six-option setting $\langle left, right, on, under, behind, front \rangle$ for the VG as it contains additional spatial annotations.

To better understand model performance, we analyze the **label distribution**, which provides insight into the representation of different spatial relationships within the dataset. In **Controlled_Image**, labels are uniformly distributed across categories (e.g., equal number of samples for "*left*", "*right*", "*on*", "*under*"). Another interesting feature of this dataset, which contributes to our choice to use it, is its contrastive setting. Controlled_Image includes **pairs** of images with same objects in both "*left*" and "*right*" positions, and **sets** of same objects exhibiting "*left*", "*right*", "*on*", and "*under*" relationships. It enables us further assess model performance using **pair accuracy** and **set accuracy**, requiring correct identification of all relationships within a pair or set, defined by Kamath et al. (2023). It provides a comprehensive evaluation of spatial relationships.

3.1. VLMs allocate sparse attention to the image.

We analyze how output tokens attend to image tokens by extracting the attention logits across layers, and present the following key findings: The sum of attention scores to **the image tokens is significantly lower** than that to all the

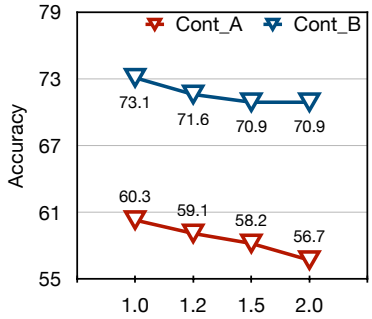

Figure 4. Accuracy of **adding** image attention in **logit space** (which corresponds to the multiplication operation in probability space; the x-axis of the figure represents the multiplication coefficient). ADAPTVIS, on the other hand, utilizes multiplication in logit space.

input text tokens, despite the considerably higher number of image tokens. In Figure 3, we focus on the attention scores from the first generated token and sum the attention allocated to the image tokens in average for all attention heads in all samples in WhatsUp. The results reveal that image tokens receive substantially less attention, with text tokens receiving approximately nine times more. Although the image sequence has a length of 576, compared to the text sequence, which typically ranges from 30 to 40 tokens in our short question-answering setting, the model predominantly focuses on text when generating outputs. That is, image information is sparsely processed by the language model.

3.2. Is the answer more accurate if the model sees the image more? Not really.

Building on earlier observations, a natural question arises: since the attention scores assigned to the image are relatively low compared to those for text, could increasing attention to the image improve the factual accuracy of the model? To investigate this, we conduct an experiment by increasing the attention weights allocated to the entire image by the final answer, intervening by adding positive constants to the image attention logits across all patches , as described by Zhang et al. (2023). In Figure 4, we observe that adding a constant weight uniformly across all image tokens does not improve performance on spatial reasoning tasks. This observation motivates us to explore more intelligent ways to focus on key visual features.

4. Visual Attention Distribution

Our analysis of text-image attention imbalance has established that visual information is underutilized, yet simply increasing attention to image tokens fails to improve spatial reasoning. We hypothesize that the way attention is geometrically distributed across the image is the key factor. To investigate this hypothesis, we examine how attention patterns are distributed spatially by mapping the 576 image tokens in LLaVA 1.5 to their 24×24 corresponding image patches. This mapping enables us to trace and visualize the geometric flow of attention, showing us not just where the model looks, but how its attention aligns with actual spatial relationships in the image.

4.1. The model automatically focuses on the relevant entity when correctly answering questions.

To determine whether the model's factual inaccuracy is linked to incorrect attention behavior, we use YOLO (Redmon, 2016) to annotate the relevant entities in the images. We compare the overlap between these annotations with the model's attention distribution and evaluate whether the overlap can serve as a metric to predict the correctness of the model's answers. We transform the bounding box coordinates obtained from YOLO annotations into image-sized index vectors, where each element is either 0 or 1. We then normalize these index vectors along with the attention logits of the image attended by the final answer of certain layers corresponding to the image patches. We then calculate the cosine similarity of the two tensors for each sample, and analyze this metric on Cont_A. As shown in Figure 6, the AUROC - a metric used to measure how well a model can distinguish between positive and negative classes.) is notably high in the middle layers, with results for all the layers at Figure 7. Additional results on other subsets are presented in the Appendix E.5.

To investigate why the middle layers exhibit the most noticeable patterns, we analyze both attention scores and the overlap AUROC (Figure 7) to track information flow and assess each layer's contribution. In the initial layers, attention to the input image is the highest compared to other layers. However, the AUROC in these layers is very low, suggesting that the model primarily captures a broad, global understanding rather than focusing on local details. In contrast, the middle layers play a more critical role in refining the model's understanding. As shown on the right side of Figure 7, the overlap AUROC peaks in these layers, indicating that they are where the model begins to effectively "process" image information. Additionally, the left side of Figure 7 shows a modest peak in attention logits, further supporting our hypothesis.

We further separately examine the attention patterns for correctly and incorrectly answered questions. Our observations reveal that hallucinations frequently occur due to two types of attention failures: (1) insufficient attention to the correct object, and (2) misplaced attention on irrelevant objects in the image. Figure 5 illustrates these findings. More examples are shown in Appendix E.5. In the two correct examples on the left, the attention scores are well-aligned with the referenced entities, with sufficient focus. On the other hand,

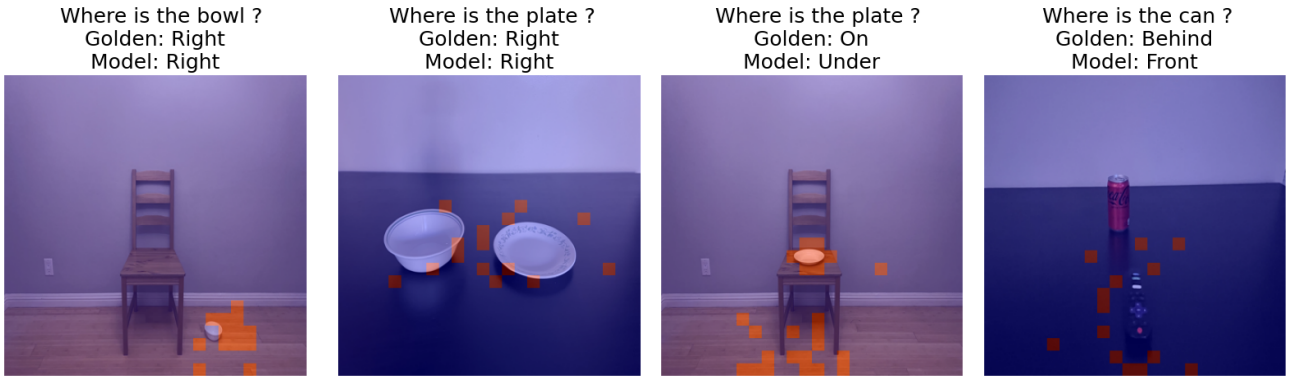

Figure 5. Attention visualization examples from the WhatsUp Dataset. The left two examples are answered correctly, while the right two are incorrect. For correctly answered questions, the attention scores are precisely focused on the core entities mentioned. In contrast, incorrect answers show attention scores distributed to irrelevant image regions. The visualizations use attention from the 17th layer, and the title in each image is an abbreviation of "Where is A in relation to B".

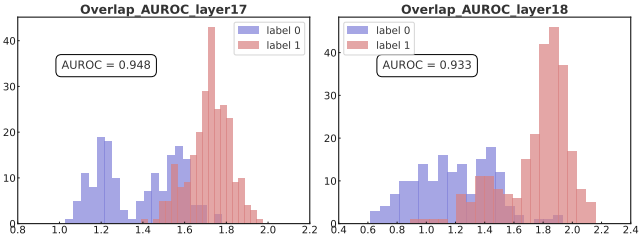

Figure 6. The figure illustrates the AUROC of the overlap between YOLO annotations and the attention patterns for Cont_A at the 17th and 18th layers. This high AUROC suggests that **attention could be an effective metric for detecting answer correctness**. more AUROC results are shown at Appendix D.

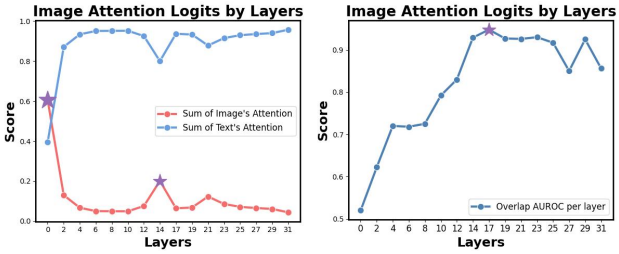

Figure 7. Left: Variance of image token attention across layers in Cont_A, showing sparse attention on image, consistent with other subsets (Appendix E.4). Right: The Overlap's AUROC value across the layer of Cont_A. We can infer from the two figures that the model "sees" the image information in the early layers and "processes" this information in the intermediate layers. So for case study in our paper, we use the intermediate layers.

the two incorrect examples on the right demonstrate how the model incorrectly assigns attention, effectively "seeing" the wrong parts of the image. While these examples highlight the qualitative differences in attention patterns, they do not provide a quantitative metric that aligns with our goal–developing a method to detect the reliability of internal states and enable intervention. Therefore, the primary challenge lies in devising effective strategies to adjust the attention scores intelligently, given that we have no prior information about the scores until a single inference run generates the attention map.

4.2. SCALINGVIS: Temperature scaling to image attention distribution

Our observations from Section 4.1 reveal that the model often misallocates attention logits within images, leading to errors in spatial reasoning. To address this, we aim to enhance the model's ability to focus on key visual features, improving its capacity to accurately ground spatial relationships, especially in complex or ambiguous scenarios. To achieve this, we propose a straightforward yet effective method that dynamically adjusts the image attention by modifying the temperature of the attention distributions. By adjusting the temperature, we can make the distribution either sharper or smoother, effectively altering the attention distribution. For instance, if the attention pattern is mostly correct but lacks precision, we want to sharpens it. Conversely, if the attention pattern is fundamentally incorrect, we want to smooth it out, allowing us to explore alternative regions. This intervention targets the attention of the final input token (at the n-th position) to the image tokens.

$$\boldsymbol{A}_{n,j}^{(l,h)} = \begin{cases} \alpha \boldsymbol{A}_{n,j}^{(l,h)} & \text{if } j \in \mathcal{I} \\ \boldsymbol{A}_{n,j}^{(l,h)} & \text{otherwise} \end{cases} \tag{2}$$

where $\mathcal{I}$ represents the indices of all image tokens. In essence, we change the temperature of image attention distribution by multiplying a coefficient α. From a physical perspective, multiplying by a coefficient greater than 1 encourages the model to focus more on the original distribution, while multiplying by a coefficient less than 1 makes the

distribution more dispersed. In experiments, we uniformly apply this coefficient to all H heads across all L layers to avoid the need for extensive hyperparameter search.

Experiment Setting. We select two widely-used benchmarks on evaluating the model's ability on spatial reasoning WhatsUp (Kamath et al., 2023) (introduced in Section 2), and VSR (Liu et al., 2023), which contains contains 1223 image-caption pairs with boolean labels. The original VSR is designed in $\langle$*image*, *caption*$\rangle$ format to evaluate encoder models without generation capabilities. To adapt it for our purposes, we utilize GPT-4o to generate questions for the VSR dataset. For evaluation, we report both accuracy and F_1 scores. A small validation set is allocated for each subset to optimize the temperature based on validation performance, and the final test is conducted on the test set. For both methods, the hyperparameter α is selected from $[0.5, 0.8, 1.2, 1.5.2.0]$. For baselines, we adopt DoLa (Chuang et al., 2023) as a baseline, which employs the inner knowledge to calibrate the output logits by substracting the logits from intermediate layers. Additionally, we incorporate VCD (Leng et al., 2024), which is a contrastive decoding method to contrast the logits before and after eliminating the image.

Results. Our results for ScalingVis are presented in Table 1 and Table 2. By controlling the distribution of attention weights, spatial reasoning performance improves significantly, with gains of up to 37.2 absolute points. An interesting pattern emerges: a temperature below one tends to enhance performance on synthetic data in most cases (3 out of 4), while a temperature above one benefits real image datasets across all cases. Table 1 indicates that for **synthetic data**, smoothing the image attention logits improves performance. Conversely, for **real image datasets** (COCO and VG), the optimal temperature is consistently above one, demonstrating that a sharper attention distribution helps the language model discern relationships more effectively. Intuitively, we believe that for familiar datasets and spatial relationships, the model requires a sharper attention distribution, as it is generally correct but may not be sufficiently precise. In contrast, for unfamiliar datasets and spatial relationships, a smoother attention distribution is necessary to explore a broader context.

5. Adaptively Intervening the Attention Distribution by Model Confidence

The findings from our study raise a key question: Since sharpening the distribution can sometimes improve performance while smoothing it is preferable in other cases, can we establish a metric to determine when to make these adjustments adaptively? In other words, can we establish a metric on whether we want to strengthen the original distribution or break it? In this section, we explore how to assess when to trust image attention.

5.1. Self-confidence reflects whether we can trust the model's image attention distribution.

We aim to determine whether an internal metric can indicate when the model's attention pattern is trustworthy. Since we found that attention patterns often correlate with the correctness of the model's generation as shown in § 4.1, we reframe our goal as identifying a metric that helps assess the reliability of its outputs. Following previous findings of using generation logits as a measure of self-confidence to assess generation reliability (Kadavath et al., 2022; Chen et al., 2024b), we also adopt the logits of generated outputs to represent the model's self-confidence. Specifically, we adjust the coefficient for all four spatial relationships in Cont_A and Cont_B, as these subsets provide a uniform distribution of different spatial relationships. We then examine how confidence and accuracy change in response and present the results in Figure 8. First, when the coefficient is 1.0 (baseline), the model exhibits significantly higher confidence for "left" and "right" relationships while showing notably lower confidence for other relationship types. This suggests that the model is more familiar with "left" and "right" relationships and less familiar with others. Furthermore, the accuracy of predicting "left" and "right" relationships is also higher, indicating the positive correlation between the model's self-confidence and its correctness in predicting them. Second, we observe distinct response patterns across different coefficients. Increasing the coefficient improves performance for "left" and "right" relationships, whereas decreasing it enhances performance for other relationships. This suggests that the model responds differently based on the relationship type, necessitating tailored intervention methods. These insights motivate us to propose an adaptive intervention strategy.

5.2. AdaptVis

Motivated by our observation that attention misallocation leads to spatial reasoning errors and that the model exhibits varying self-confidence and correctness across different spatial relationships, we propose AdaptVis: Confidence-aware temperature scaling, an extension of ScalingVis:

Confidence-Based Attention Intervention.

Recall from Section 5.1 that we observe two distinct patterns in the model's factuality behavior: (1) synthetic data presents more unfamiliar cases than the real data, and (2) VLMs could express uncertainty through confidence scores. These insights motivate using confidence scores as a metric for adaptive intervention in the model's internal states. Our intuition is straightforward: when confidence is low, suggesting that the attention pattern may be unreliable, we smooth

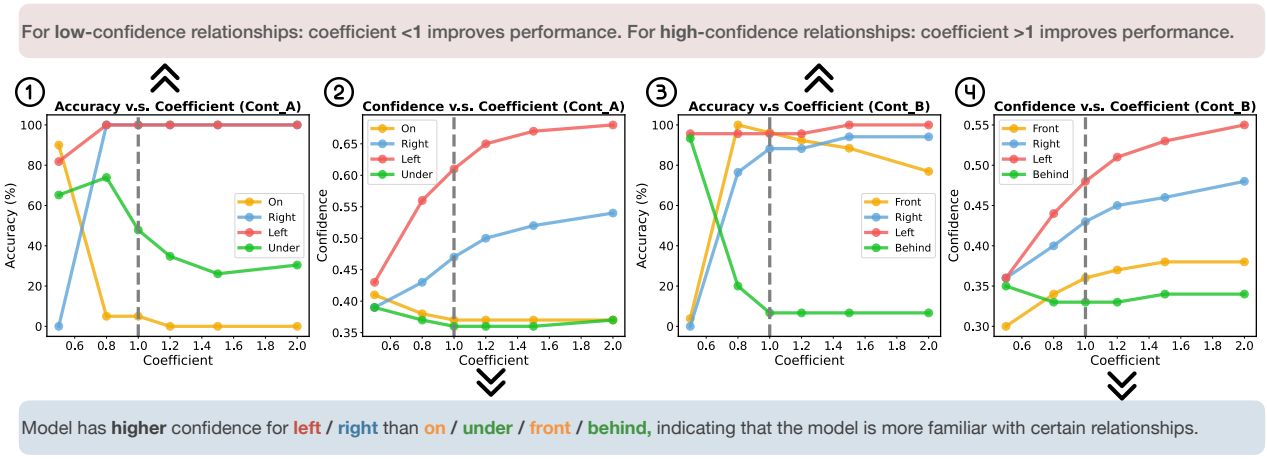

Figure 8. ①: Change in accuracy with the coefficient on Cont_A. ②: Change in confidence with the coefficient on Cont_A. ③: Change in accuracy with the coefficient on Cont_B. ④: Change in confidence with the coefficient on Cont_B. The baseline (greedy decode) corresponds to a coefficient of 1.

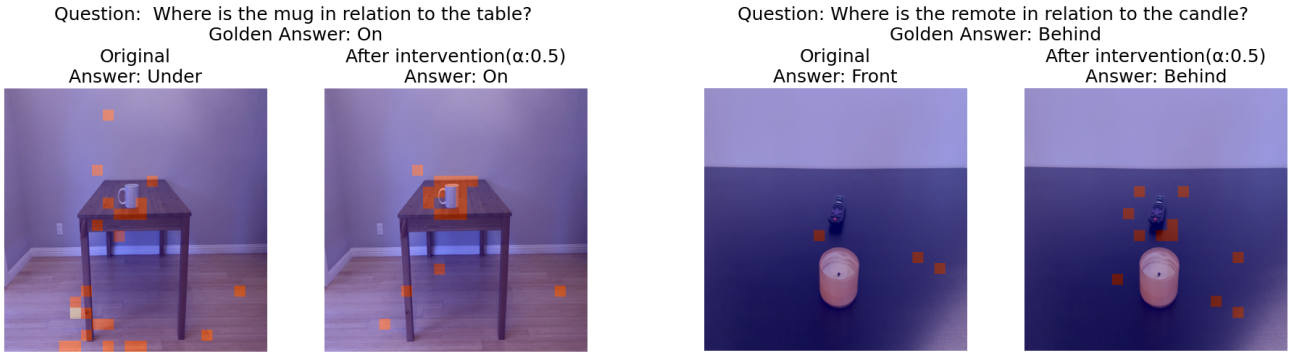

Figure 9. Attention scores on the patches before and after our intervention in the 17th layer for the images in **synthetic** datasets Cont_A and Cont_B. We employ a "**smoothing**" intervention method to **expand the context length** of the model's focused area. From the figure, it is evident that the model's focused position undergoes significant changes after our intervention.

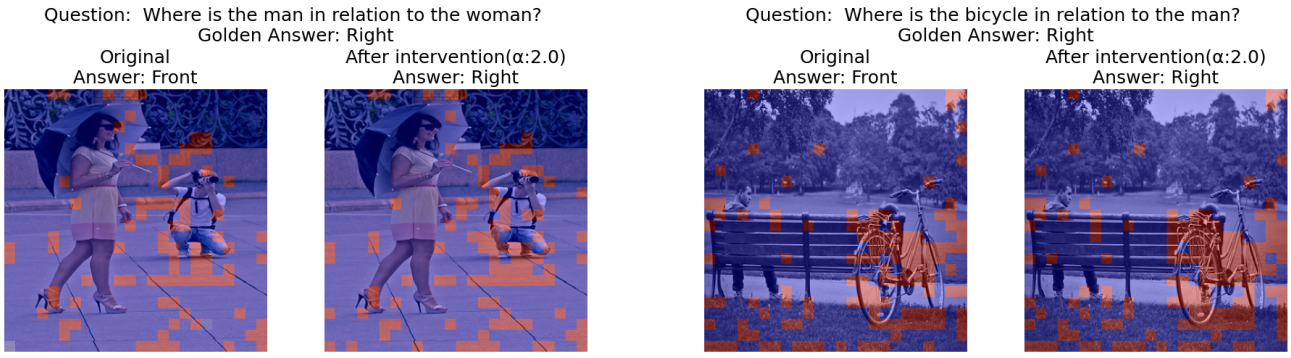

Figure 10. Attention scores on the patches before and after our intervention in the 17th layer for the images in **real** datasets COCO and VG. We utilize a "**sharpening**" intervention method to enhance the original attention pattern. The highlighted areas remain largely consistent, with our method serving to reinforce the focus rather than significantly altering it.

the attention distribution. This encourages the model to explore a broader range of image regions, increasing the likelihood of focusing on the correct patches. Conversely, when confidence is high and attention is dispersed across

$$\mathbf{A}_{n,j}^{(l,h)} = \begin{cases} \alpha_1 \mathbf{A}_{n,j}^{(l,h)} & \text{if } j \in \mathcal{I} \\ \mathbf{A}_{n,j}^{(l,h)} & \text{otherwise} \end{cases}, \text{if } \mathcal{C} < \beta \quad (3a)$$

$$\mathbf{A}_{n,j}^{(l,h)} = \begin{cases} \alpha_2 \mathbf{A}_{n,j}^{(l,h)} & \text{if } j \in \mathcal{I} \\ \mathbf{A}_{n,j}^{(l,h)} & \text{otherwise} \end{cases}, \text{if } \mathcal{C} > \beta \quad (3b)$$

Model	Controlled_A			Controlled_B			COCO_one	COCO_two	VG_one	VG_two
	Acc	P Acc	S Acc	Acc	P Acc	S Acc	Acc	Acc	Acc	Acc
LLaVA-1.5	60.3	40.6	0.0	73.1	41.6	3.7	53.0	58.2	35.9	40.8
+VCD	61.5 ↑1.2	39.4 ↓1.2	0.0	73.4 ↑0.3	42.2 ↑0.6	3.7	53.3 ↑0.3	58.2	35.8 ↓0.1	42.5 ↑1.7
+DoLa	61.2 ↑0.9	41.6 ↑1.0	0.0	73.4 ↑0.3	42.2 ↑0.6	3.7	**53.7** ↑0.7	57.5 ↓0.7	36.2 ↑0.3	42.1 ↑1.3
+ScalingVis	64.5 ↑4.2	40.6	0.0	75.2 ↑2.1	44.6 ↑3.0	9.8 ↑6.1	53.6 ↑0.6	59.4 ↑1.2	42.7 ↑6.8	48.1 ↑7.3
+AdaptVis	**84.9** ↑24.6	**61.2** ↑20.6	**30.3** ↑30.3	**83.8** ↑10.7	**55.7** ↑14.1	**18.3** ↑14.6	53.6 ↑0.6	**59.9** ↑1.7	**42.7** ↑6.8	**48.1** ↑7.3
LLaVA-1.6	48.2	37.6	0.0	63.0	39.1	3.7	59.7	41.8	31.6	7.3
+VCD	61.8 ↑13.6	41.8 ↑4.2	10.9 ↑10.9	65.4 ↑2.4	41.6 ↑2.5	7.3 ↑3.6	60.6 ↑0.9	44.9 ↑3.1	33.8 ↑2.2	11.6 ↑4.3
+DoLa	48.2	37.6	0.0	62.7 ↓0.3	39.1	3.7	59.7	41.5 ↓0.3	31.5 ↓0.1	7.3
+ScalingVis	97.0 ↑48.8	76.4 ↑38.8	54.5 ↑54.5	73.4 ↑10.4	48.9 ↑9.8	15.9 ↑12.2	**63.1** ↑3.4	**47.7** ↑5.9	**38.2** ↑6.6	14.6 ↑7.3
+AdaptVis	**98.2** ↑50.0	**78.8** ↑41.2	**57.0** ↑57.0	**73.4** ↑10.4	**48.9** ↑9.8	**15.9** ↑12.2	**63.1** ↑3.4	**47.7** ↑5.9	35.2 ↑3.6	**17.2** ↑9.9

Table 1. Results on Controlled A, Controlled B, COCO, and VG datasets. (Metrics in $\times 10^{-2}$). Best-performing method per model and dataset are highlighted in bold. P Acc and S Acc represents Pair Acc and Set Acc.

Model	VSR	
	Exact Match	F1 Score
LLaVA-1.5	62.4	51.3
+VCD	62.4	50.6 ↓0.7
+DoLa	62.8 ↑0.4	53.2 ↑1.9
+ScalingVis	64.9 ↑2.5	62.5 ↑11.2
+AdaptVis	**65.0** ↑2.6	**62.5** ↑11.2
LLaVA-1.6	58.8	29.4
+VCD	58.8	29.4
+DoLa	59.3 ↑0.5	31.2 ↑1.8
+ScalingVis	59.1 ↑0.3	30.6 ↑1.2
+AdaptVis	**62.7** ↑3.9	**39.3** ↑9.9

Table 2. Results on the VSR dataset (Exact Match and F1) (Metrics in $\times 10^{-2}$). Best-performing method per model and dataset are highlighted in bold.

the image, we sharpen the distribution to concentrate on key objects more effectively. Specifically, we apply the targeted intervention to the attention of the last input token (at the n-th position) directed toward the image tokens as shown in Equation 3a, 3b. Overall, we use a large $\alpha > 1$ when the Confidence $\mathcal{C}$ is large, which sharpens the attention distribution, and the relevant objects are paid more attention to; we use a small $\alpha < 1$ when the confidence $\mathcal{C}$ is small, which mitigates the model's excessive concentration on certain image tokens and makes the overall attention distribution smoother across the image.

Evaluation Settings. We employ the same setting with ScalingVis as shown in 4.2. For hyperparameter choice and robustness, we show in Appendix E.3.

Results Our main results are presented in Table 1 and Table 2. By controlling the distribution of attention weights, we observe a **significant improvement in spatial reasoning ability, with gains of up to 50 absolute points**. In most cases, AdaptVis achieves the best performance, particularly for synthetic datasets like Cont_A and Cont_B, as shown in Table 1, where it significantly outperforms the generalized method ScalingVis. These findings suggest that model performance varies considerably with the label distribution of the dataset, and smoothing the distribution (by applying a coefficient smaller than 1) enhances performance. For real-image datasets like COCO and VG, the adaptive method performs slightly better than the generalized approach, indicating the model's robustness across different label distributions. Example cases are shown in Figure 9 and Figure 10, illustrating the smooth effect and focus effect ($\alpha < 1$ and $\alpha > 1$), respectively. It is important to note that the LLaVA-series models are trained on the COCO dataset, which makes them highly confident and familiar with COCO and VG image types. Hence trusting the model's self-belief and sharpening image attention improves performance. Notably, for datasets containing more unfamiliar images, the adaptive setting proves to be significantly more effective.

6. Related Work

The first line of related work focuses on the attention patterns in LMs. Some studies on attention patterns in LLMs reveal biased attention across context windows, such as ineffective use of the middle context (Liu et al., 2024b) and initial token attention sinks (Xiao et al., 2023). While some approaches use fine-tuning to overcome these biases (An et al., 2024), training-free methods like input-

adaptive calibration (Yu et al., 2024b) and position-specific interventions (Yu et al., 2024a) offer efficient alternatives. PASTA (Zhang et al., 2023), a closely related method, emphasizes attention on selected segments for specific heads; we extend this to VLMs without manual segment specification or multiple validation runs. Our work is also related to failure analysis in VLMs, VLMs have been shown to hallucinate more in multi-object recognition tasks and rely on spurious correlations (Chen et al., 2024c), with systematic visual limitations highlighted from a CLIP perspective (Tong et al., 2024b). Our work also connects to the decoding strategies for reducing hallucinations decoding strategies to mitigate hallucinations include contrastive decoding focusing on image regions (Leng et al., 2024), preference tuning through data augmentation (Wang et al., 2024), and methods leveraging contrastive layers for enhanced knowledge extraction (Chuang et al., 2023), as well as activation-based optimal answer identification (Chen et al., 2024b).

7. Conclusion and Future Work

Our research uncovers the inner working mechanism of VLMs during spatial reasoning, which is a critical limitation in VLMs and constrains their practical utility when requiring geometric understanding of visual scenes. We identify critical insights through an in-depth study of attention behaviors across layers: 1) VLMs allocate surprisingly insufficient attention to image tokens; 2) the location of attention on image tokens is more crucial than quantity; and 3) generation confidence serves as a reliable indicator of its familiarity with the image and the correctness of its attention pattern. Based on these findings, we propose ADAPTVIS, a novel decoding method that dynamically adjusts attention distribution, significantly improving spatial reasoning performance. Future research could focus on further exploring the mechanism of VLMs on complicated geometric structure understanding, such as long-horizon spatial reasoning, and investigate other spatial reasoning bottlenecks.

Impact Statement

Our research into the spatial reasoning capabilities of Large Vision Language Models (VLMs) has significant implications across various domains of artificial intelligence and its real-world applications. First and foremost, our findings highlight a critical limitation in current VLMs: while they excel at object recognition, they struggle with basic spatial relationships. This gap between recognition and spatial understanding has far-reaching consequences for the practical deployment of VLMs in scenarios requiring geometric comprehension of visual scenes. Industries such as robotics, autonomous navigation, and assistive technologies for the visually impaired are particularly affected. For instance, a robot that can identify objects but cannot understand their spatial relationships may struggle with tasks like picking and placing items or navigating complex environments. Similarly, autonomous vehicles might face challenges in interpreting traffic scenarios accurately, potentially compromising safety.

Our development of ADAPTVIS, a novel decoding method that dynamically adjusts attention distribution based on the model's confidence, represents a significant step forward. By enhancing VLMs' performance on spatial reasoning tasks, ADAPTVIS could unlock new possibilities in various fields. In healthcare, improved spatial reasoning could lead to more accurate interpretation of medical imaging, potentially improving diagnostic accuracy. In augmented reality applications, better spatial understanding could enable more immersive and interactive experiences. For assistive technologies, enhanced spatial reasoning could provide more accurate and useful descriptions of environments to visually impaired individuals, significantly improving their independence and quality of life.

Looking ahead, our work opens up new avenues for research in AI and cognitive science. The exploration of mechanism interpretability in VLMs, particularly for complex geometric structures and long-horizon spatial reasoning, could provide insights into how artificial systems process and understand spatial information. This could not only advance AI capabilities but also contribute to our understanding of human spatial cognition. Additionally, investigating the role of training data memorization in spatial reasoning bottlenecks could lead to more efficient and effective training methods for future AI models.

In conclusion, our research not only addresses a fundamental limitation in current VLMs but also paves the way for more versatile and capable AI systems. As we continue to advance VLMs' capabilities in visually-driven tasks requiring nuanced spatial understanding, we have the potential to significantly impact various sectors of society, from healthcare and assistive technologies to urban planning and environmental monitoring. The future research ahead in this field is both exciting and challenging, requiring ongoing collaboration between researchers, ethicists, and policymakers to ensure that these advancements benefit society as a whole.

Acknowledgments

We thank Min Lin, Miao Xiong, Guangtao Zeng, Teng Xiao, Wei Liu, Chi Han, Chang Ma, Zhengxuan Wu, and Junlong Li for their valuable discussions and insights.

References

An, S., Ma, Z., Lin, Z., Zheng, N., and Lou, J.-G. Make your llm fully utilize the context, 2024.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A. C., Korbak, T., and Evans, O. The reversal curse: Llms trained on" a is b" fail to learn" b is a". *arXiv preprint arXiv:2309.12288*, 2023.

Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., and Xia, F. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024a.

Chen, S., Xiong, M., Liu, J., Wu, Z., Xiao, T., Gao, S., and He, J. In-context sharpness as alerts: An inner representation perspective for hallucination mitigation. *arXiv preprint arXiv:2403.01548*, 2024b.

Chen, X., Ma, Z., Zhang, X., Xu, S., Qian, S., Yang, J., Fouhey, D. F., and Chai, J. Multi-object hallucination in vision-language models, 2024c. URL https://arxiv.org/abs/2407.06192.

Chuang, Y.-S., Xie, Y., Luo, H., Kim, Y., Glass, J., and He, P. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*, 2023.

Gao, J., Pi, R., Zhang, J., Ye, J., Zhong, W., Wang, Y., Hong, L., Han, J., Xu, H., Li, Z., et al. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*, 2023.

Hudson, D. A. and Manning, C. D. 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.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Kamath, A., Hessel, J., and Chang, K.-W. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9161–9175, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.568. URL https://aclanthology.org/2023.emnlp-main.568.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., and Bing, L. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024.

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., and 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.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Liu, F., Emerson, G. E. T., and Collier, N. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 2023.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024a.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024b. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9.

OpenAI. .https://cdn.openai.com/papers/GPTV_System_Card.pdf, 2024. [Accessed 06-08-2024].

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.

Redmon, J. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., and 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.

Tong, S., Brown, E., Wu, P., Woo, S., Middepogu, M., Akula, S. C., Yang, J., Yang, S., Iyer, A., Pan, X., et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024a.

Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., and Xie, S. 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, 2024b.

Wang, F., Zhou, W., Huang, J. Y., Xu, N., Zhang, S., Poon, H., and Chen, M. mdpo: Conditional preference optimization for multimodal large language models, 2024. URL https://arxiv.org/abs/2406.11839.

Xiao, G., Tian, Y., Chen, B., Han, S., and Lewis, M. Efficient streaming language models with attention sinks. *arXiv*, 2023.

Xiong, M., Hu, Z., Lu, X., LI, Y., Fu, J., He, J., and Hooi, B. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *The Twelfth International Conference on Learning Representations*, 2024.

Yu, Y., Jiang, H., Luo, X., Wu, Q., Lin, C.-Y., Li, D., Yang, Y., Huang, Y., and Qiu, L. Mitigate position bias in large language models via scaling a single dimension. *arXiv preprint arXiv:2406.02536*, 2024a.

Yu, Z., Wang, Z., Fu, Y., Shi, H., Shaikh, K., and Lin, Y. C. Unveiling and harnessing hidden attention sinks: Enhancing large language models without training through attention calibration. In *International Conference on Machine Learning*, 2024b.

Zhang, Q., Singh, C., Liu, L., Liu, X., Yu, B., Gao, J., and Zhao, T. Tell your model where to attend: Post-hoc attention steering for llms, 2023.

Appendix

A. Limitations

Firstly, our methods, SCALINGVIS and ADAPTVIS, specifically address model-related spatial hallucinations and self-alignment issues but are not designed to handle errors outside the language model's capabilities, such as the CLIP failures discussed by Tong et al. (2024b). Secondly, ADAPTVIS relies on distribution-based confidence to adaptively set the confidence threshold β, we also observe that the optimal α and β is different across different distributions and prompts. This dependence on a validation set for tuning poses a limitation on its applicability.

B. Additional Results on Qwen2-VL

We further expand our intervention experiments and attention analysis on Qwen2-VL, the results on spatial reasoning benchmarks and on general benchmarks are shown in Table 3 and Table 4 respectively. We intervene in the image attention distribution using our temperature-scaling method, showing consistent improvements below, particularly in challenging cases. For example, on VG_two_object, where the baseline performance is the lowest among all benchmarks, our method yields a significant improvement of 10+ absolute points. The gains observed on Qwen2-VL further demonstrate the generalizability of our approach. Experiments on more benchmarks shown in Table 4, including POPE (Li et al., 2023), GQA (Hudson & Manning, 2019), and TextVQA (Singh et al., 2019), which proves that attention intervention can maintain the performance on more general tasks without hurting the performance. Compared with spatial reasoning tasks, general QA tasks achieve relatively smaller improvement. A possible reason is that such tasks are less sensitive to the geometric structured distribution of image attention. For example, given a question like "Is there a dog in this picture?", the model only needs to detect the presence of an object, and is therefore less likely to suffer from misallocated attention across spatial regions.

We also conduct an attention analysis to verify whether our observation of the sparsity of image attention compared to textual tokens still holds. The results are shown in Table 6. Here we calculate the Qwen2-VL's attention scores below (average attention logits for single text/image tokens respectively), which matches our previous claim that image receives much less attention than the text tokens is generally valid for VLMs. And we also conduct the uncertainty experiments to show our uncertainty conclusion is also valid. The results are shown in Table 5. It shows that spatial relationships with lower confidence, such as "Left" and "Under" tend to exhibit lower accuracy when compared to those with higher confidence, like "On" and "Right". After the intervention with AdaptVis, certain spatial relationships like "Left" show improved performance, accompanied by an increase in confidence. This demonstrates a pattern consistent with our observations for LLaVA, as depicted in Figure 8 in Section §4.

Benchmark	Qwen2-VL	+Attention Intervention
VSR	78.96	81.60 ↑2.64
Coco_one_obj	76.64	78.03 ↑1.39
Coco_two_obj	75.28	76.52 ↑1.24
VG_one_obj	74.89	75.11 ↑0.22
VG_two_obj	56.22	66.95 ↑10.73
Controlled_A	98.18	98.18 ↑0.00
Controlled_B	91.73	92.97 ↑1.24

Table 3. Performance of Qwen2-VL before and after the attention intervention on spatial reasoning benchmarks.

Benchmark	Qwen2-VL	+Attention Intervention
POPE-Overall	86.32	87.09 ↑0.77
POPE-P	86.47	87.29 ↑0.82
POPE-A	85.07	85.80 ↑0.73
POPE-R	87.46	88.22 ↑0.76
GQA	62.09	62.17 ↑0.08
TextVQA	79.18	79.26 ↑0.08

Table 4. Performance of Qwen2-VL before and after the attention intervention on additional general benchmarks.

Benchmark	Text attention	Image attention
Controlled_A	1.57e-02	7.59e-05
Controlled_B	1.58e-02	7.69e-05
Coco_one_obj	1.77e-02	4.48e-04
Coco_two_obj	1.65e-02	3.50e-04
VG_one_obj	1.54e-02	4.75e-04
VG_two_obj	1.42e-02	4.19e-04

Table 5. Average text-token and image-token attention scores across benchmarks.

C. Related work

C.1. Attention Patterns in Language Models

Ongoing research has shown how large language models (LLMs) exhibit biased attention across different parts of the context window. Liu et al. (2024b) find that LLMs fail to effectively utilize the information in the middle of a long context window. Meanwhile, Xiao et al. (2023) reveals an

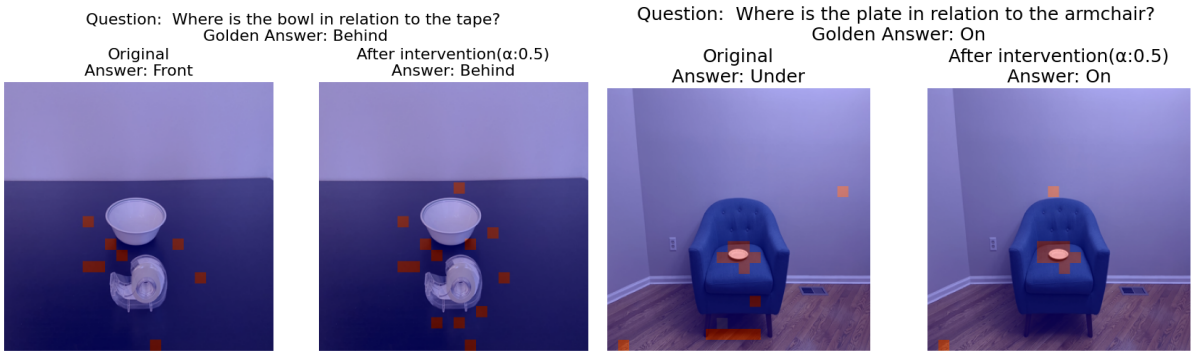

Figure 11. Examples of our fixed case for $\alpha = 0.5$.

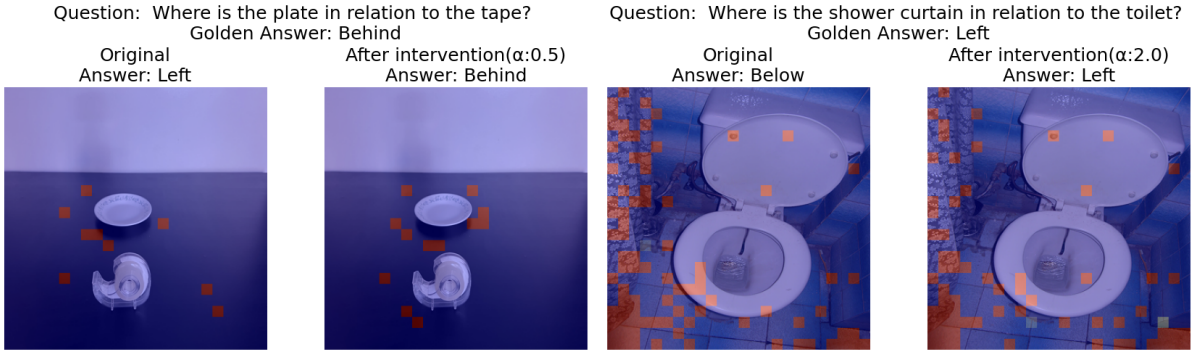

Figure 12. Examples of our fixed case for $\alpha = 2.0$.

attention sink at the initial tokens of the input. Besides fine-tuning methods to overcome such biases (An et al., 2024), some training-free methods have been proposed with the benefit of their efficiency. Yu et al. (2024b) proposes to use input-adaptive calibration to adjust the attention scores, while Yu et al. (2024a) intervenes in position-specific hidden dimensions to alleviate the lost-in-the-middle phenomenon. A closely related work to ours is PASTA (Zhang et al., 2023), which emphasizes the attention scores of specific text segments for selected attention heads. We further develop this motivation on vision language models. Moreover, our method does not require a manual specification of the emphasized segment or multiple validation runs to identify effective attention heads.

C.2. Failure Analysis of Vision-Language Models

Our work relates to research on hallucination detection in VLMs. Chen et al. (2024c) examine multi-object recognition tasks, observing that VLMs exhibit more hallucinations when dealing with multiple objects compared to single-object scenarios. They also note a similar phenomenon to our findings: the distribution of tested object classes impacts hallucination behaviors, suggesting that VLMs may rely on shortcuts and spurious correlations. Additionally, Tong et al. (2024b) analyze VLM failures from a CLIP perspective, highlighting that the visual capabilities of recent VLMs still face systematic shortcomings, partly due to CLIP's limitations in specific cases.

C.3. Decoding Strategies for Reducing Hallucinations

This work is also connected to various decoding and tuning strategies aimed at mitigating hallucinations in VLMs. Leng et al. (2024) introduce a contrastive decoding method that emphasizes certain image regions. Wang et al. (2024) propose a data-augmentation approach to create image-intensive datasets, followed by preference tuning on this enhanced data. Furthermore, knowledge extraction techniques such as the method proposed by Chuang et al. (2023) improve decoding by leveraging contrastive layers for better knowledge extraction. Similarly, Activation Decoding (Chen et al., 2024b) identifies optimal answers as those with the highest activation values within the context.

D. Case Study

We show more case we could fix in Figure 11 and Figure 12. We show more attention examples at Figure 13.

Benchmark	Qwen2-VL Unc.	+ScalingVis Unc.	Qwen2-VL Acc	+ScalingVis Acc
Left	0.4408	0.4474	70.11	74.71
On	0.5758	0.5668	100	100
Right	0.5982	0.5911	100	100
Under	0.5554	0.5418	98.82	98.82

Table 6. Uncertainty (*Unc.*) and accuracy (*Acc*) on spatial-relation categories before and after applying ScalingVis.

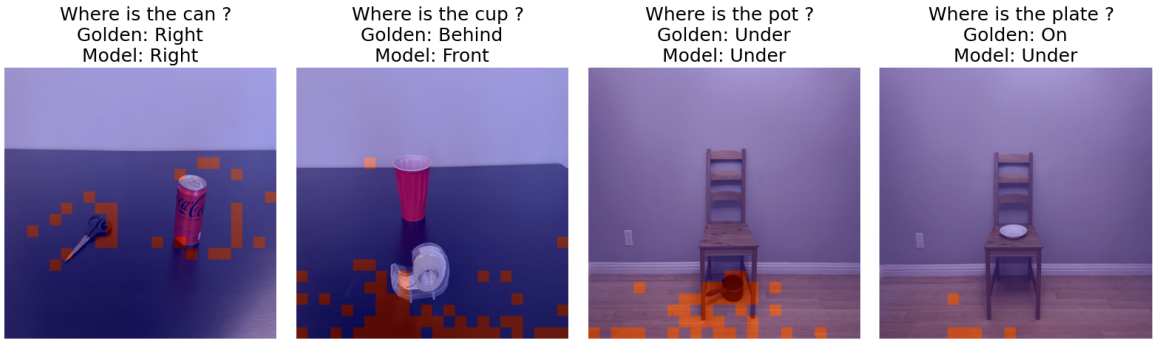

Figure 13. Examples of image attentions. Visualization here is adopted from 14th layers' attention scores.

E. Further Analysis

E.1. Other metrics to distinguish the distributions

We further explored additional metrics to distinguish the distributions and uncover their underlying characteristics during our preliminary experiments. Specifically, we examined entropy and skewness. Entropy was selected based on the hypothesis that parameter differences may stem from the familiarity of attention patterns in real images, which are generally correct, versus synthetic images, where these patterns tend to be incorrect. We posit that the model can express "familiarity" through certain metrics derived from the attention scores. For example, we hypothesize that the entropy of attention will be lower when the model encounters familiar cases.

$$E\left(\mathcal{A}_{n,j}^{(l,h)}\right) = -\sum_{j=1}^{t} \tilde{P}\left(\mathcal{A}_{n,j}^{(l,h)}\right) \log \tilde{P}\left(\mathcal{A}_{n,j}^{(l,h)}\right) \quad (4)$$

In Equation 4,$\mathcal{A}_{n,j}^{(l,h)}$ denotes the attention scores assigned by the h-th head in the l-th layer to the j-th token in sequence n. The summation runs over $j = 1$ to t, where t is the total number of tokens considered for this attention distribution. $\tilde{P}\left(\mathcal{A}_{n,j}^{(l,h)}\right)$ is the normalized probability distribution of these attention scores. This entropy measures the uncertainty or spread of the attention distribution across tokens.

Our experimental results in Figure 14 indicate that the attention distribution is heavily influenced by image features. Notably, the attention distribution is more concentrated in synthetic datasets than in real images. We attribute this to the fact that synthetic images tend to contain fewer objects, resulting in a sharper attention distribution. However, this concentration does not provide a reliable metric for measuring familiarity. Another possible metric is skewness. Another possible metric is skewness, which captures the asymmetry of the attention distribution. A high skewness suggests that the attention is predominantly focused on a few positions, while a low skewness indicates a more balanced spread across multiple regions. By examining skewness, we aim to identify whether the attention is being disproportionately allocated to particular image features, which could provide additional insights into how familiarity is expressed through attention patterns. We could see from Figure 14 that Synthetic datasets show a higher skewness. However, it also related with the object distribution, which is not the real factor behinds the difference.

$$S\left(\mathcal{A}_{n,j}^{(l,h)}\right) = \frac{\sum_{j=1}^{t} \left(j - \mu_{\mathcal{A}}\right)^3 \tilde{P}_j}{\sigma_{\mathcal{A}}^3} \quad (5)$$

The summation runs over $j = 1$ to t, where t is the number of tokens considered in the attention distribution. $\tilde{P}_j$ denotes the probability assigned to the j-th token in the normalized attention distribution. The term $\mu_{\mathcal{A}}$ is the mean of the distribution, and $\sigma_{\mathcal{A}}$ is its standard deviation. The skewness is calculated as the normalized third central moment, which measures the asymmetry of the attention distribution: a positive value indicates a distribution skewed to the right, while a negative value indicates a skew to the left.

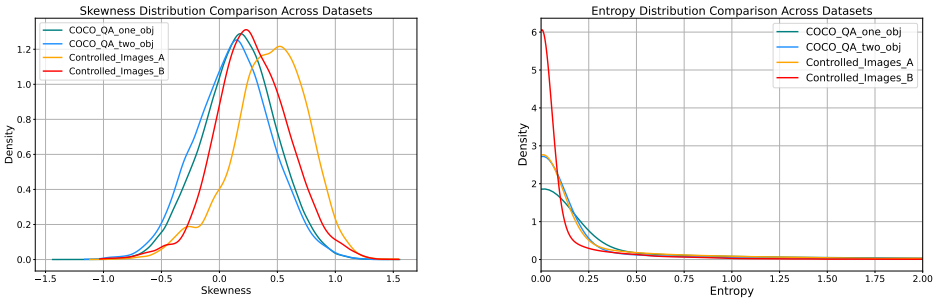

Figure 14. The skewness and entropy distribution comparison between different subsets. Here we use Controlled_Images and COCO datasets due to all of them are in four option label space, which enables us to eliminate the influence of prompts.

Dataset	Relationship Types					
	Right	Left	On	Under	Behind	Front
Controlled_A	92	92	130	92	0	0
Controlled_B	102	102	0	0	102	102
VG_one	376	392	192	198	2	0
VG_two	137	127	3	0	5	19
COCO_one	564	576	363	744	0	0
COCO_two	129	150	86	75	0	0

Table 7. Gold Answer Frequency by Spatial Relation in WhatsUp dataset.

E.2. Label statistics in WhatsUp

We present the golden labels' distribution in Table 7. We can see that the label space in synthetic datasets is balanced while the real image datasets are imbalanced with more "left" and "right" and fewer other relationships.

Label Distribution and the familarity To determine whether to focus more on the image or the text, we begin by analyzing existing datasets to identify potential patterns. Our investigation starts with an examination of label distributions across different subsets of the WhatsUp dataset. As shown in Figure 15, there is a clear label imbalance in the real-image datasets including COCO_two and VG_two, i.e., only a small portion of the samples have the relation of "*behind*" and "*front*". This may be attributed to annotation bias, where left and right are easier to distinguish than other relationships. In contrast, the synthetic datasets, Cont_A and Cont_B, which are carefully curated, display more balanced label distributions. Additionally, we evaluate the model's confidence scores across the ground-truth spatial relations by following the approach of Kadavath et al. (2022), where the probability of the generated output is used to compute confidence. Figure 15 reveals that **the model is unconfident with specific spatial relationships**, such as "*on*" and "*under*" where it shows a colder color at Figure 15, while **demonstrating higher confidence in recognizing more common relationships** where it shows a warmer color. In Figure 8, we observe that the model performs better on confident relationships. For example, in the first figure on the left, when the coefficient is set to 1 (the baseline), the ground-truth samples labeled as *"left"* and *"right"* exhibit significantly better performance compared to those labeled as *"on"* and *"under"*. This observation aligns with our intuition that model tends to be more confident when it performs well and less confident when it struggles. This finding is also consistent with previous work showing that models can convey their uncertainty through confidence scores (Kadavath et al., 2022; Xiong et al., 2024). Motivated by this, we propose using confidence as a metric to gauge the model's familiarity with spatial relationships in images.

Statistics of LLaVA training data To categorize six different spatial relationships, we check whether specific phrases appear in the "llava_v1_5_mix665k.json", which is the training data of LLaVA and increment corresponding counters. For the **left** relationship, we look for phrases such as "left side," "left of," "to the left," and "on the left." Similarly, for the **right** relationship, we detect phrases like "right side," "right of," "to the right," and "on the right." The **on** relationship is identified using phrases like "are on the," "is on the," and "located on," ensuring they do not co-occur with "on the left" or "on the right." To detect the **under** relationship, we check for phrases such as "under the," "beneath the," or "below the." For the **front** relationship, we look for phrases like "are in front of," "is in front of," and "locate in front of." Finally, the **behind** relationship is identified by the phrase "behind the.". The results are shown in Figure 15.

E.3. Hyperparameter choice and robustness

For ADAPTVIS We use the same hyperparameters for SCALINGVIS as in Section 4.2. For ADAPTVIS, we optimize α_1, α_2, and β using the validation set (randomly sampled 20% from each subset) from each distribution. We also show our robustness for hyperparameters in Appendix. Notably, we find that model performance is robust across a range of these hyperparameters, generalizing effectively to

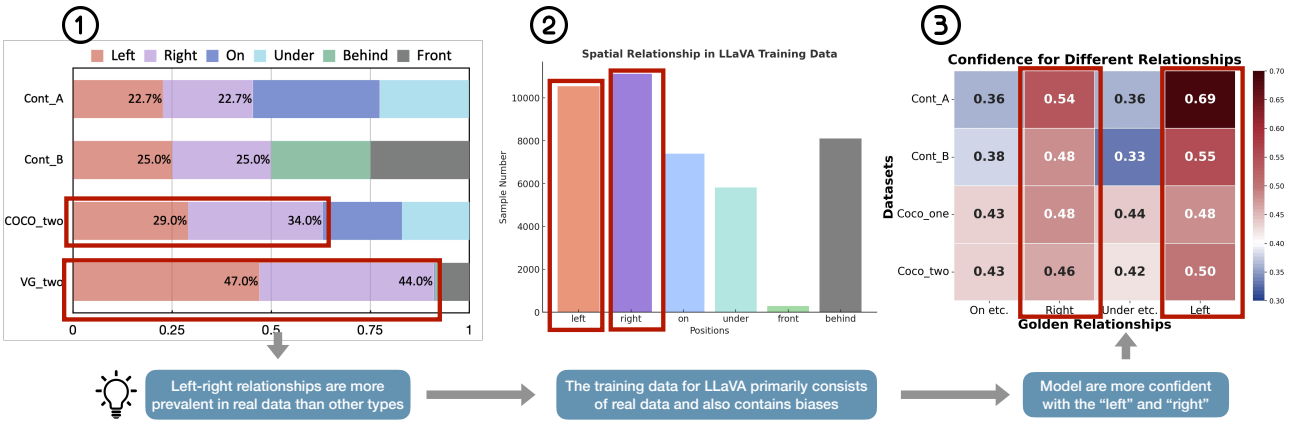

Figure 15. ①: Label distributions across subsets capturing relationships between object pairs in WhatsUp. For **real datasets** like COCO and VG, a severe **label imbalance** is evident, with **"left"** and **"right"** being more common than other relationships.② show the statistics in LLaVA training data. We could see imbalance problem still holds. ③: The model's average confidence in different golden spatial relationships within these four-option subsets in WhatsUp. "On etc" includes "*on*, *above*, *top*", while "Under etc" includes "*under*, *below*, *bottom*". The red heatmap box highlights instances where the model is confident in its generation, while the blue box indicates the opposite. We can observe that **for familiar relationships**, such as "left" and "right" (highlighted with two red boxes), **the model shows higher confidence.**

other subsets within the same distribution (as demonstrated in Table 8). We maintain consistency by using the same range of α values for both methods. For β, we adjust per dataset: for WhatsUP, we select values from $[0.3, 0.65]$ for LLaVA-1.6 and $[0.2, 0.55]$ for LLaVA-1.5 with a grid size of 0.05 (this higher range is due to LLaVA-1.6 generally exhibiting higher confidence than LLaVA-1.5); for VSR, we take the mean value of the average confidence scores corresponding to the two labels.

Out-of-Domain Test of Hyperparameters. We evaluated the generalizability of common hyperparameters across datasets. To this end, we applied the same set of four-option prompts to Controlled_Images and COCO subsets. Results in Table 8 indicate that ADAPTVIS consistently performs well across all subsets, confirming its generalizability.

Model	Cont_A			Cont_B			Coco_one	Coco_two
	Acc	P-Acc	S-Acc	Acc	P-Acc	S-Acc	Acc	Acc
LLaVA	60.3	40.6	0.0	73.1	41.6	3.7	53.0	58.2
+Ours	60.3	41.8 ↑1.2	2.4 ↑2.4	76.5 ↑3.4	48.3 ↑6.7	13.5 ↑9.8	53.6 ↑0.6	59.4 ↑1.2
Best α	$\alpha_1 = 0.5$ $\alpha_2 = 1.2$ $\beta = 0.3$							

Table 8. OOD test results on WhatsUp (Metrics in $\times 10^{-2}$) for LLaVA1.5. Arrows show growth over baseline. P-Acc and S-Acc are Pair and Set Accuracy.

E.4. More Attention Analysis

The sparcity To determine whether our observation about the sparse attention pattern holds across benchmarks, we examine the attention patterns in other subsets of WhatsUp. Figure 16 presents the attention pattern for Cont_B, while

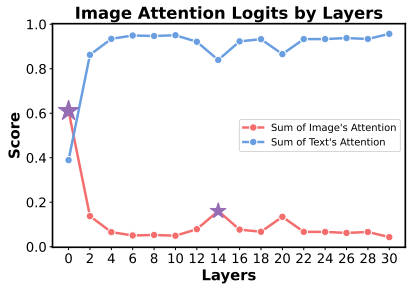

Figure 16. The variance of image token's attention scores through the layers in Cont_B benchmark.

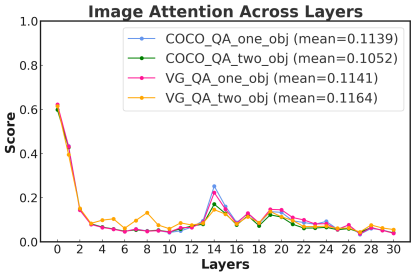

Figure 17. Image attention on COCO and VG dataset.

Figure 17 illustrates the attention pattern for real images. Consistently, we observe that image attention remains sparse across all subsets.

AUROC Analysis of attention score and confidence we conduct a calibration experiment using two statistical ap-

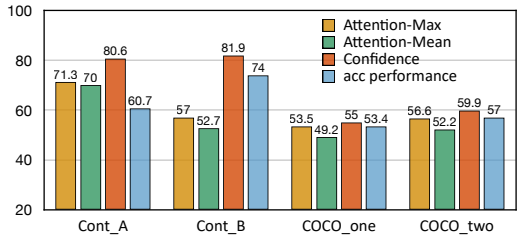

Figure 18. AUROC of attention scores in 17th layer relative to model's confidence.

proaches. The first approach sums the attention scores assigned to the image as a metric, while the second extracts the highest attention score among the image tokens. These metrics are then evaluated to assess their effectiveness in distinguishing between correct and incorrect generations. However, as shown in Figure 18, the AUROC score using attention (the yellow and green bar) is consistently lower than the AUROC score of the model's self-confidence (the yellow bar), which we measure by the probability of the output tokens. Additionally, we observe that the maximum attention score provides better calibration than the average attention score, suggesting that key information aligns more closely with maximal attention values. This suggests that the assumption "*the more attention the model pays to the image, the more accurate the results*" holds only partially true.

E.5. AUROC anlysis of the overlap between YOLO and the attention scores

We present the AUROC of the overlap between YOLO annotations and attention scores on the Controlled_A dataset in Figure 19 and Figure 20. The results demonstrate that the AUROC is remarkably high in the middle-to-high layers, indicating that the attention pattern can serve as a reliable metric for detecting factual errors.

However, in some cases, this relationship is not as apparent, as shown in Figures 21, 22, 23, and 24. In these instances, the AUROC remains relatively modest, which can largely be attributed to errors in YOLO's annotations. We identified four primary error types leading to mismatches between YOLO annotations and actual object instances: Missed Detection, Misclassification, Bounding Box Error, and Ambiguous Refer. These error patterns are illustrated in Figure 25.

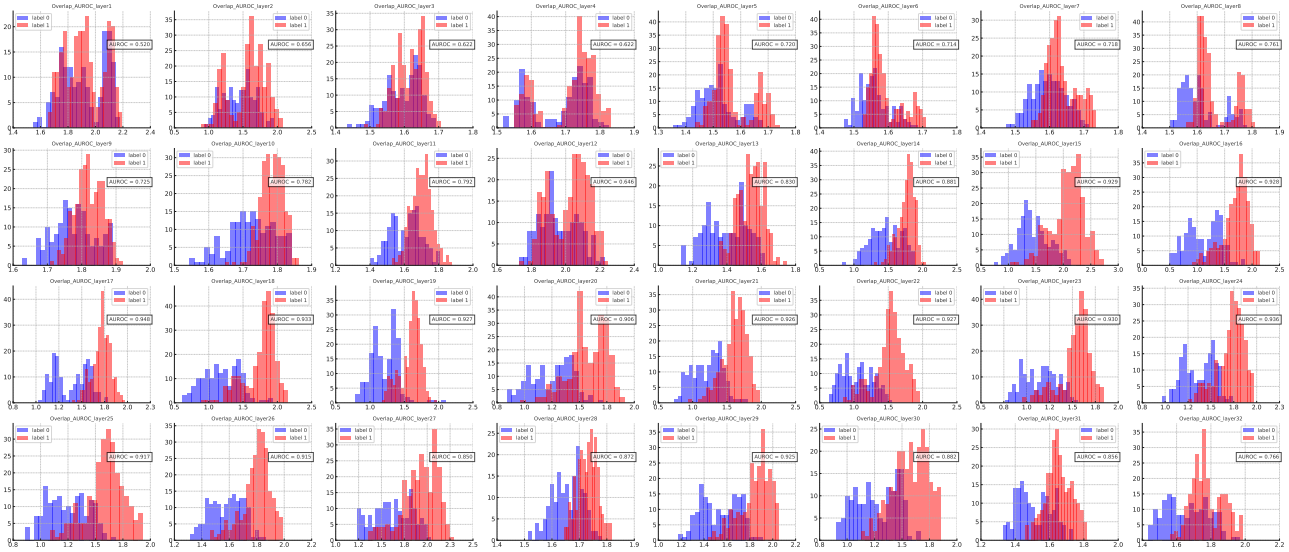

Figure 19. AUROC of the overlap between YOLO annotations and the attention patterns on all layers in Cont_A benchmark.

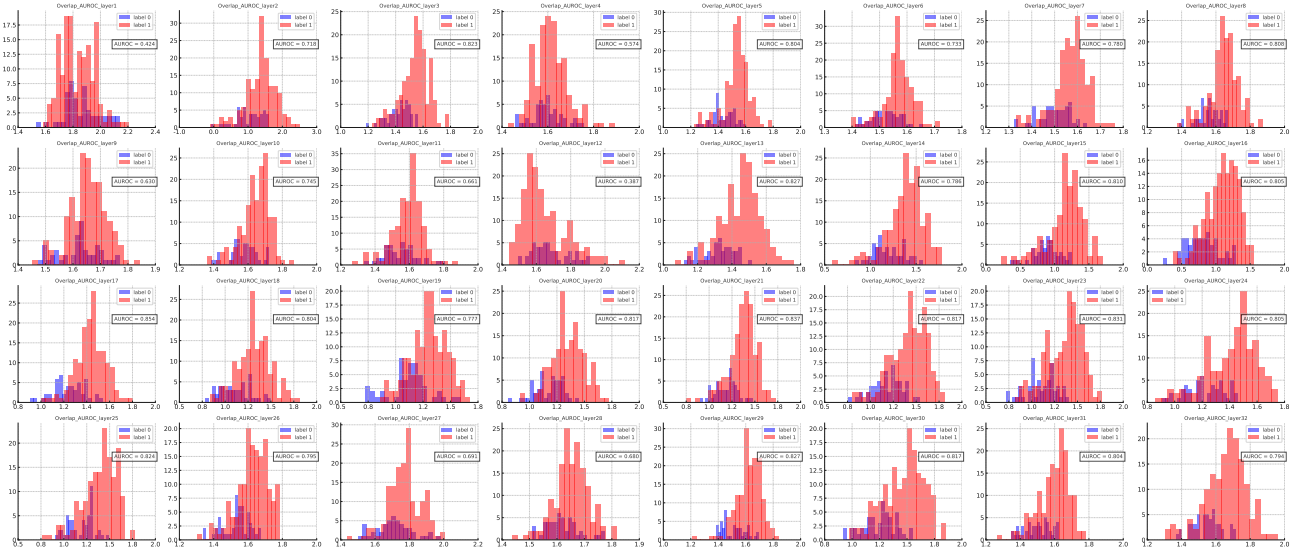

Figure 20. AUROC of the overlap between YOLO annotations and the attention patterns on all layers in Cont_B benchmark.

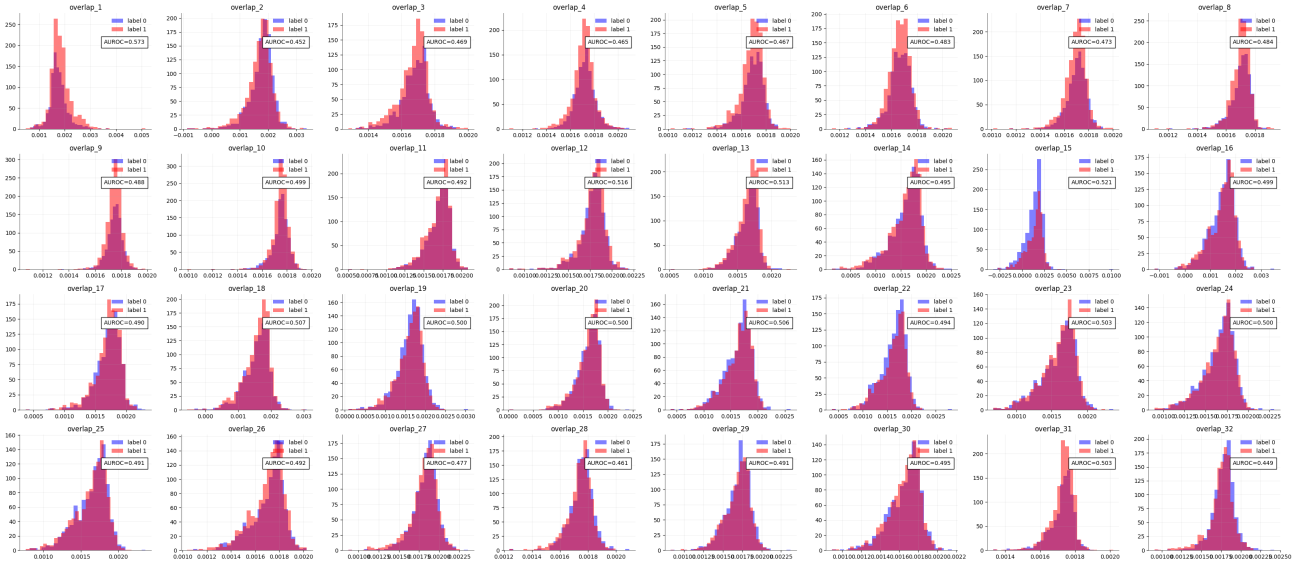

Figure 21. AUROC of the overlap between YOLO annotations and the attention patterns on all layers in COCO_one benchmark.

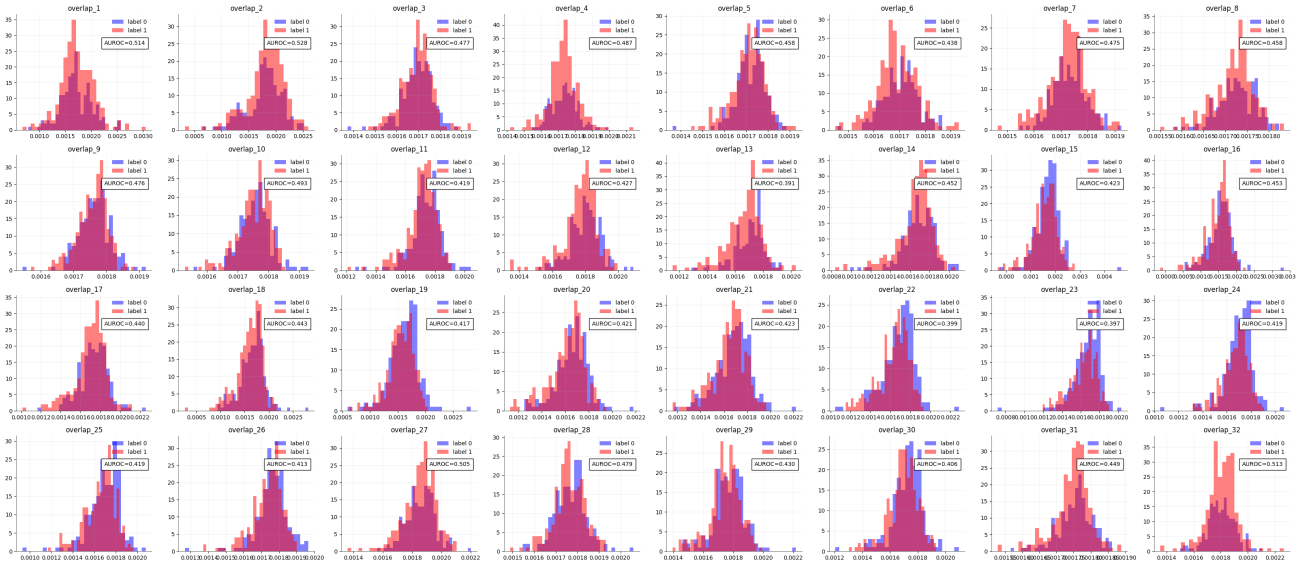

Figure 22. AUROC of the overlap between YOLO annotations and the attention patterns on all layers in COCO_two benchmark.

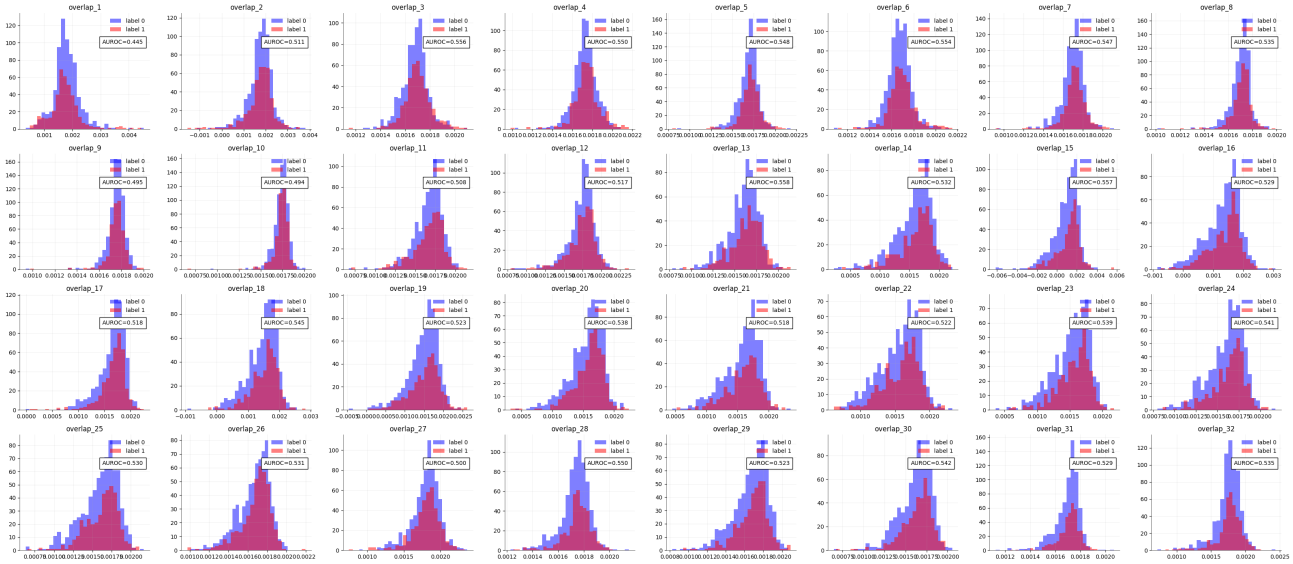

Figure 23. AUROC of the overlap between YOLO annotations and the attention patterns on all layers in VG_one benchmark.

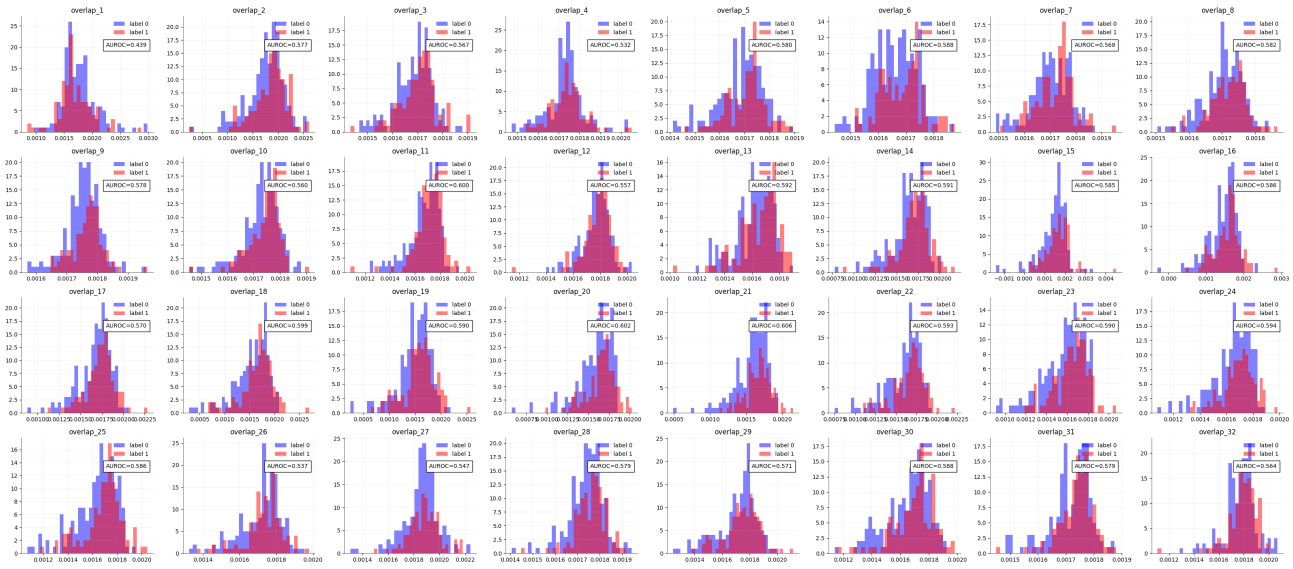

Figure 24. AUROC of the overlap between YOLO annotations and the attention patterns on all layers in VG_two benchmark.

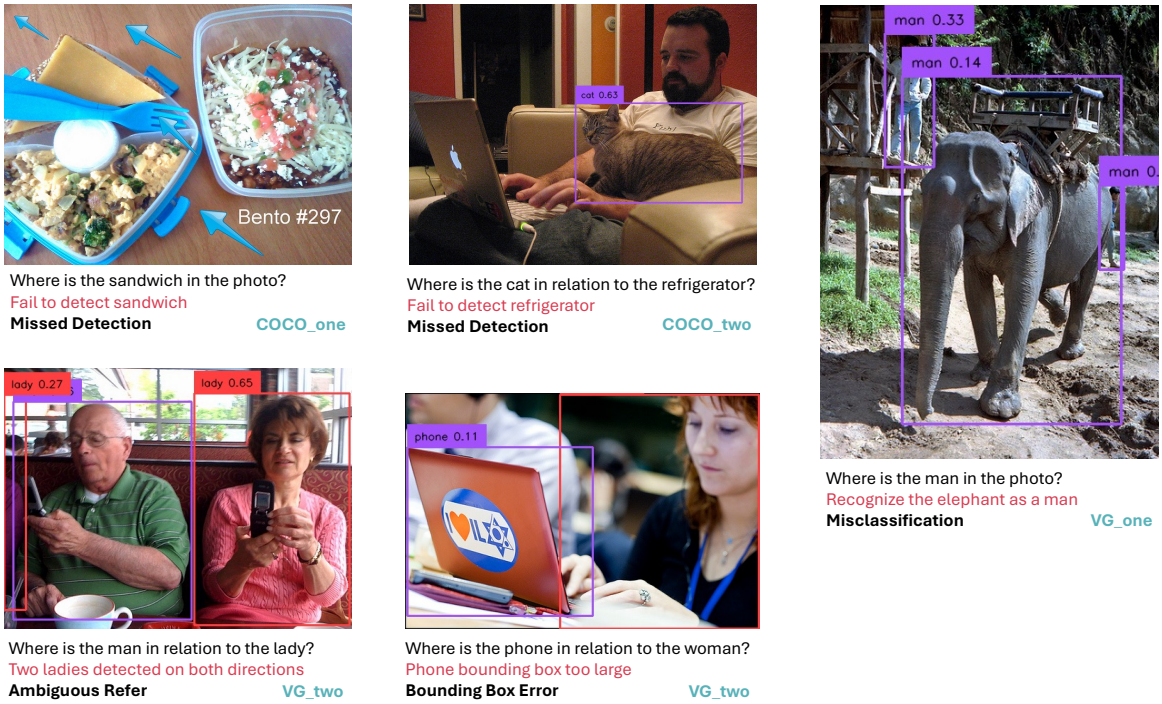

Figure 25. Error cases in YOLO annotation. Under each image we present the related question, the error in YOLO annotation and its type, and the source where the image and question comes from.

E.6. What changes after our intervention?

How the confidence changes by using different αs? To illustrate the impact of coefficients greater or less than 1 on different relationships, Figure 8 shows accuracy and confidence variance for various ground-truth relationships. Results indicate that for familiar relationships like "*left*" (red) and "*right*" (blue), coefficients greater than 1 boost accuracy and confidence. Conversely, for less familiar relationships like "*under*" (green) and "*on*" (yellow), coefficients less than 1 improve accuracy and confidence.

How do the absolute values of attention scores vary before and after intervention? Figure 26 visualizes the attention logits for two-option Cont_A (α=0.5) and six-option VG (α=2). An α of 0.5 decreases the absolute value of logits across layers, while an α of 2 increases them as layers progress. This indicates that an α larger than 1 could strengthen the model's orginal attention pattern.

E.7. Prompt Sensitivity Analysis

To assess the robustness of our method, we varied the number of options in prompts for specific WhatsUp subsets. For Cont_A and Cont_B, we reduce the number of options to two, simplifying the task. Conversely, for COCO subsets, we increase the options to six, making it more challenging. Table 27 shows that our SCALINGVIS method maintains consistent performance across all cases, demonstrating robustness.

E.8. Reverse curse?

Additionally, we observe that the model exhibits a "reverse curse" phenomenon similar to that has been seen in language models (Berglund et al., 2023).

When we reverse the order of the entities in Cont_A (e.g., asking the model, "*Where is the armchair in relation to the beer bottle?*" instead of "*Where is the beer bottle in relation to the armchair?*"), there is a significant drop in performance. As shown in Figure 28, the model's performance declines dramatically from a high score to an exceptionally low one. This reveals that the existing VLMs's attention pattern and generation results could be significantly impacted by the prompt. Our methods, however, consistently improve the model's performance. It suggests that adaptively intervening the attention score is a generalizable method for different prompts.

E.9. Efficiency

We evaluate inference times for different decoding methods in Table 9. ScalingVis introduces only negligible additional computation time compared to the baseline (greedy decoding). However, due to the need for computing the threshold, AdaptVis incurs higher computational overhead.

	Baseline	ScalingVis	AdaptVis
Time (set baseline to 1)	1.00	1.02	1.64

Table 9. Inference-time statistics. It shows that ScalingVis introduces negligible computation compared to the baseline.

Dataset	POPE-A	POPE-P	POPE-R	POPE-AVG	GQA	VQAv2
Baseline	81.0	86.4	88.4	85.3	56.1	74.0
ScalingVis	81.8	86.6	88.6	85.6	56.3	74.2

Table 10. Results on QA benchmarks. POPE-A is POPE-Adversarial. POPE-P is POPE-Popular. POPE-R is POPE-random. POPE-AVG the is average score of the three subsets.

E.10. Results on More Benchmarks

We further evaluate our method on several Question Answering benchmarks, including POPE (Li et al., 2023), GQA (Hudson & Manning, 2019), and VQAv2 (Antol et al., 2015). As shown in Table 10, our method consistently outperforms the baseline across all benchmarks, demonstrating its strong generalization capability in the general Question Answering setting.

F. URLs of Code and Data

We provide the code and data at `https://anonymous.4open.science/r/AdaptVis-B73C`.

The running arguments are configured as follows: `dataset` specifies the evaluation dataset (e.g., `Controlled_Images_A`); `model-name` selects the model (e.g., `llava1.5`); `method` determines the evaluation approach (`scaling_vis` or `adapt_vis`). For `scaling_vis`, `weight1` can be set to values in $[0.5, 0.8, 1.2, 1.5, 2.0]$. For `adapt_vis`, `weight1` ranges in $[0.5, 0.8]$ and `weight2` in $[1.2, 1.5, 2.0]$, with `threshold` controlling the adaptation sensitivity. The `--download` flag enables automatic dataset downloading.

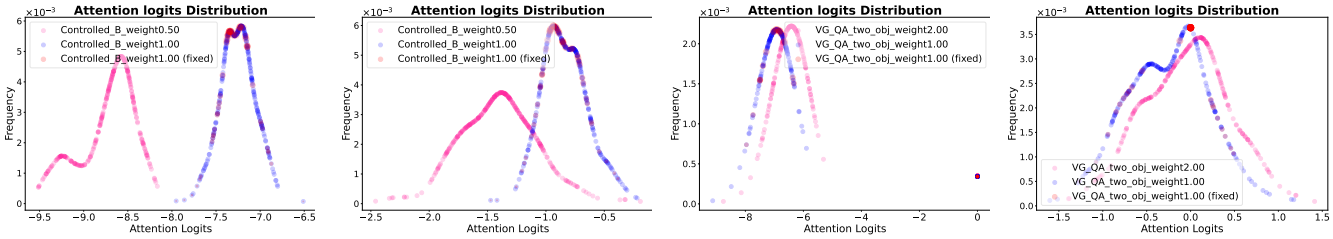

Figure 26. Attention logit distribution before (blue) and after (pink) intervention in the 14th layer (a randomly chosen middle layer). From left to right, the plots represent: mean and max attention values across heads for Cont_A, and mean and max attention for Cont_B. Red dots mark cases corrected by our intervention. We could see that α of 0.5 shifts the line left, while α of 2 shifts it right.

Model	Cont_A			Cont_B			Coco_one	Coco_two
	Acc	P Acc	S Acc	Acc	P Acc	S Acc	Acc	Acc
LLaVA-1.5	76.4	43.0	4.8	74.6	41.0	1.2	30.8	42.6
+Ours	86.4 ↑10.0	61.2 ↑18.2	27.9 ↑23.1	87.8 ↑13.2	59.3 ↑18.3	22.0 ↑20.8	36.0 ↑5.2	48.6 ↑7.3
Best α	0.5			0.5			2	2

Figure 27. Results on WhatsUp (Metrics in $\times 10^{-2}$) when changing prompts for other option numbers. Arrows show improvement over greedy decoding.

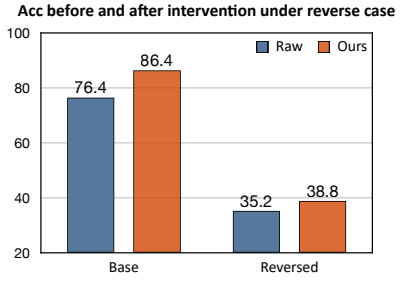

Figure 28. Performance comparison before and after SCALINGVIS intervention ($\alpha = 0.5$).