

Routing-Aware Head Amplification for Multimodal Long-Chain-of-Thought Reasoning

Chuhao Zhou¹ Zonghan Wu¹ Haolang Lu²
Kun Wang² Weiran Yao³ Linsey Pang⁴
Prakhar Mehrotra⁵ Qingsong Wen^{6,*} Yilei Shao^{1,*}

¹East China Normal University, China ²Nanyang Technological University, Singapore

³actAVA AI ⁴PayPal Inc. ⁵Walmart Labs ⁶Squirrel Ai Learning

*Corresponding authors

Abstract

Multimodal long-chain-of-thought (CoT) generation requires models to dynamically route information between visual evidence and an expanding reasoning history. Existing hallucination mitigation methods often strengthen visual grounding or intervene in decoding, but typically treat grounding as a relatively static property of generation. We propose a training-free dynamic attention intervention framework that exploits the head-level sparsity, layer-level stratification, and trajectory-level dynamics of multimodal long-CoT decoding. Our method selects routing-sensitive heads using perturbation-only calibration, assigns visual-evidence and reasoning-history interventions to different layer regions, and continuously rescales selected head outputs according to their instantaneous token-group attention ratios. Across multiple MLLM architectures and multimodal hallucination benchmarks, our method consistently improves over vanilla decoding and visual contrastive decoding, achieving an average performance gain of 4.6%.

1 Introduction

Multimodal large language models (MLLMs) (Bai et al., 2025; Team et al., 2025) increasingly rely on long chain-of-thought (CoT) generation (Wei et al., 2022; Wang et al., 2025) to solve complex reasoning tasks. Compared with short-answer decoding, long-CoT generation (Zhang et al., 2024b; Chen et al., 2026b) fundamentally changes the structure of the decoding process: each newly generated token is conditioned not only on the input image, but also on an expanding sequence of previously generated reasoning tokens. As decoding proceeds, the relative importance of these information sources

changes continuously. Some reasoning steps require renewed access to visual evidence, while others depend more heavily on the reasoning context. Consequently, multimodal long-CoT generation is not simply a grounding problem, but a dynamic information-routing process in which the model continuously shifts attention between external evidence and internal reasoning history.

Existing studies (Guan et al., 2024; Liu et al., 2024a; Bai et al., 2024; Huang et al., 2025; Samyoun et al., 2026) on multimodal hallucination have mainly focused on insufficient visual grounding, language priors, or decoding-time biases. Existing mitigation methods (Leng et al., 2024; Wang et al., 2024b; Chen et al., 2024b; Yin et al., 2024; Yu et al., 2026) therefore often strengthen the visual modality uniformly, contrast generation against perturbed visual inputs, or intervene in the decoding process at inference time. While effective, these approaches typically treat grounding as a largely static property of generation. In long-CoT reasoning, however, the generated rationale itself becomes an evolving context source that continuously influences subsequent decoding (Wang et al., 2023; Xu et al., 2025). This suggests that hallucination may arise not only from weak visual grounding, but also from ineffective routing between visual evidence and generated reasoning history throughout the decoding trajectory (Zhong et al., 2024).

Head-level routing structure. Recent studies have shown that attention heads often exhibit highly uneven responses to different information sources, suggesting that information routing during decoding is sparse and structured rather than uniformly distributed across all heads (Clark et al., 2019; Kovaleva et al., 2019; Olsson et al., 2022; Todd et al., 2024; Kang et al., 2025a). In multimodal long-CoT

reasoning, some heads consistently allocate substantially larger attention ratios to visual tokens or previous-CoT tokens, while others remain comparatively inactive with respect to these sources (Bi et al., 2025; Kang et al., 2025b). Moreover, the routing behavior of a head is itself dynamic, changing across decoding steps as the model shifts between visual grounding and reasoning-context reuse. This motivates selective intervention on heads that are both strongly and dynamically responsive to specific information sources, rather than uniformly modifying the entire attention module (Zhang et al., 2024a; Sarkar et al., 2025).

Layer-level routing stratification. Beyond head-level sparsity, information routing in multimodal long-CoT reasoning also exhibits clear depth-wise organization (Michel et al., 2019; Rogers et al., 2020; Yin et al., 2026). Prior studies have shown that attention behavior changes systematically across Transformer layers, with different regions of the network displaying different routing tendencies during decoding (Vig and Belinkov, 2019; Tenney et al., 2019). Empirically, attention to visual tokens and previous-CoT tokens often exhibits distinct depth-dependent distributions, suggesting that different routing patterns emerge in different regions of the model. This depth-wise structure further implies that routing-sensitive heads should not be treated uniformly during intervention. Instead, shallower regions can be used to strengthen visual evidence access, while deeper regions can be used to strengthen reuse of previously generated reasoning context.

Trajectory-aware routing dynamics. While head-level sparsity and layer-level stratification describe where routing behavior emerges, long-CoT reasoning further introduces a temporal dimension in which routing patterns evolve throughout decoding (Zou et al., 2026; Jiang et al., 2025a). Unlike short-form generation, long-CoT decoding is highly non-stationary: the model may repeatedly shift attention between visual evidence and previously generated reasoning context as the reasoning trajectory unfolds (Yu and Lee, 2025; Liu et al., 2025). In particular, visual grounding behavior is often highly dynamic rather than monotonic, with the model revisiting visual evidence at different decoding stages when resolving downstream reasoning uncertainty (Man et al., 2025; Li et al., 2025; Jiang et al., 2025b). Such trajectory-level dynamics make static intervention strategies fundamentally mismatched to long-CoT decoding, since a

fixed amplification strength cannot reflect whether a routing-sensitive head is actively participating in information routing at the current step (Chen et al., 2025; Wang et al., 2026). Instead, intervention should adapt to the model’s instantaneous routing state, strengthening selected heads only when they are currently engaged in visual evidence access or reasoning-history reuse (Zhang et al., 2025; Lyu et al., 2026; Yu et al., 2026).

Based on these observations, we propose a dynamic attention intervention framework for multimodal long-CoT hallucination mitigation. Our method continuously regulates selected attention heads throughout the CoT decoding trajectory, strengthening visual evidence access and reasoning-history reuse without introducing additional training or inference overhead. Experiments across multiple MLLMs and multimodal hallucination benchmarks demonstrate consistent improvements, achieving an average performance gain of 4.6%.

2 Related Work

2.1 Multimodal Hallucination in Long-CoT

Multimodal hallucination remains a major reliability challenge for large vision-language models (Fazli et al., 2026). Existing studies attribute hallucinations to insufficient visual grounding (Leng et al., 2024), language priors (Yang et al., 2025; Ouyang et al., 2026), and decoding biases (Xu et al., 2026), while prior mitigation methods improve reliability through contrastive decoding (Chen et al., 2026a), visual-token enhancement (Zhu et al., 2026), attention redistribution (Tang et al., 2026), and post-hoc correction (Liu et al., 2026a). Recent multimodal large language models increasingly rely on long chain-of-thought (CoT) reasoning (Zhang et al., 2024b), where each newly generated token depends not only on the input image, but also on previously generated reasoning tokens (Uppaal et al., 2026). As a result, the generated rationale itself becomes an evolving context source that continuously participates in subsequent decoding, introducing additional challenges for hallucination mitigation during long-CoT generation.

2.2 Structured Attention Routing in LLMs

Prior work has shown that Transformer attention exhibits sparse and structured routing behavior, where different heads and layers respond unevenly to different information sources (Chuang et al., 2024; Kang et al., 2025a,b). In multimodal models, visual

grounding is often concentrated in specific heads or layer regions (Jiang et al., 2025c), motivating inference-time interventions such as image-guided head suppression (SPIN) (Sarkar et al., 2025), shallow visual-head amplification (ClearSight) (Zhang et al., 2024a), and dual-level token/head attention regulation (VisFlow) (Tang et al., 2026). Unlike these primarily visual-attention interventions, our method jointly selects visual-evidence and previous-CoT routing heads, localizes them by layer, and dynamically rescales selected head outputs during long-CoT decoding.

3 Method

We propose a training-free inference-time framework for dynamic attention intervention in multimodal long-CoT generation. Specifically, we use token-group attention ratios to characterize how attention heads route information between visual evidence and previous reasoning context, identify routing-sensitive heads with routing patterns, and dynamically amplify their outputs during decoding. Since the intervention only rescales selected head outputs at inference time, it introduces negligible overhead and requires no additional training.

3.1 Token-Group Routing Signals

To dynamically follow the model’s routing behavior during long-CoT decoding, we first require a lightweight signal that characterizes which information source a head is currently attending to. We use the attention ratio assigned to a token group as such a routing signal.

Let $q = (l, h)$ denote the h -th attention head in transformer layer l . At decoding step t , head q produces an attention distribution $\mathbf{a}_q(t)$ over all available tokens. For a token group G , we define the token-group attention ratio as

$$R_q(G, t) = \sum_{i \in G} \mathbf{a}_q(t)_i. \quad (1)$$

Intuitively, $R_q(G, t)$ measures how much attention mass head q allocates to source group G at decoding step t , and therefore reflects the instantaneous routing intensity toward that source.

We consider two token groups in this work. The first is the visual-token group G_{vis} , containing image tokens from the multimodal input. The second is the previous-CoT group $G_{\text{cot}}(t)$, containing reasoning tokens generated before step t . Accordingly, $R_q(G_{\text{vis}}, t)$ measures the routing intensity toward

visual evidence, while $R_q(G_{\text{cot}}(t), t)$ measures the instantaneous routing intensity toward previously generated reasoning context.

3.2 Routing-Sensitive Head Selection

Since information routing is highly sparse across attention heads (Jin et al., 2025; Voita et al., 2019), uniformly intervening on all heads is both inefficient and unnecessary. Instead, we identify routing-sensitive heads that actively participate in routing to a target information source during decoding.

For each source $\star \in \{\text{vis}, \text{cot}\}$ and each head q , we compute a routing sensitivity score

$$S_\star(q) = \mathbb{E}_{x,t} [R_q(G_\star(t), t)] \cdot \text{Std}_{x,t} [R_q(G_\star(t), t)]. \quad (2)$$

where x indexes calibration samples and t indexes decoding steps.

The expectation term measures how strongly a head participates in routing to a source on average, while the standard-deviation term measures how dynamically this routing behavior changes throughout decoding. Therefore, a high routing sensitivity score indicates that a head is not only strongly engaged with a source, but also dynamically activated during the generation trajectory.

The resulting scores $\{S_{\text{vis}}(q)\}_q$ and $\{S_{\text{cot}}(q)\}_q$ form source-specific routing maps over attention heads, which are later used to localize dynamic interventions.

3.3 Layer-Localized Intervention

Although routing-sensitive heads can be identified from routing maps, information routing in multimodal long-CoT reasoning is not uniformly distributed across network depth. Instead, routing behavior often exhibits a clear layer-wise stratification, with different regions of the network displaying different routing tendencies toward visual evidence and previous-CoT context.

To capture this structure, we estimate a coarse layer boundary from the layer-wise previous-CoT routing profile. For layer l , we compute

$$C_l^{\text{cot}} = \mathbb{E}_{x,t,h} [R_{(l,h)}(G_{\text{cot}}(t), t)], \quad (3)$$

which measures the average amount of previous-CoT routing in layer l over calibration samples and decoding steps.

We then identify the layer boundary as the largest transition point in the routing profile:

$$b = \arg \max_l |C_{l+1}^{\text{cot}} - C_l^{\text{cot}}|. \quad (4)$$

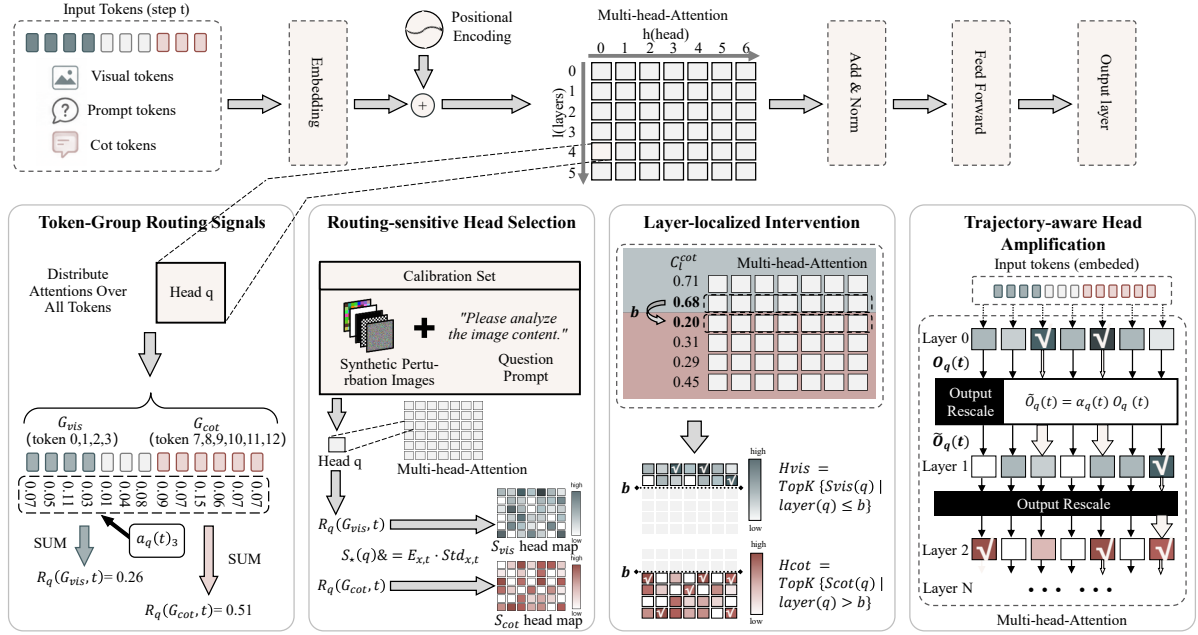


Figure 1: Overview of routing-aware head amplification. The method identifies heads that route to visual evidence or previous CoT tokens, localizes them by layer, and dynamically rescales their outputs during decoding according to their current routing strength.

This boundary provides a coarse shallow/deep stratification of the network and is used only as a structural prior for localizing interventions across depth.

Based on this stratification, visual-sensitive heads are selected from the shallow region,

$$\mathcal{H}_{vis} = \text{TopK} \{S_{vis}(q) \mid \text{layer}(q) \leq b\}, \quad (5)$$

while previous-CoT-sensitive heads are selected from the deep region,

$$\mathcal{H}_{cot} = \text{TopK} \{S_{cot}(q) \mid \text{layer}(q) > b\}. \quad (6)$$

The final intervention set is $\mathcal{H} = \mathcal{H}_{vis} \cup \mathcal{H}_{cot}$.

This layer-localized design preserves the routing structure of the network while avoiding uniform intervention across all layers and heads.

3.4 Trajectory-Aware Head Amplification

Long-CoT decoding is highly non-stationary, and the routing behavior of an attention head can change across decoding steps. A routing-sensitive head may strongly engage with visual evidence or previous-CoT context at certain stages of generation, while remaining inactive at others. Therefore, applying a fixed intervention strength throughout decoding is fundamentally mismatched to the dynamic nature of long-CoT reasoning.

Instead, we use a trajectory-aware conditional intervention strategy that follows the model’s instantaneous routing state. Intuitively, a selected

head should receive stronger amplification only when it is actively routing information to its target source. If the head is not currently engaged with that source, the intervention should remain weak so that the original decoding dynamics are preserved.

Let $\mathbf{o}_q(t)$ denote the output vector of head q at decoding step t before the multi-head output projection. We rescale this vector as

$$\tilde{\mathbf{o}}_q(t) = \alpha_q(t) \mathbf{o}_q(t), \quad (7)$$

where $\alpha_q(t)$ is a routing-aware amplification factor defined as

$$\alpha_q(t) = \begin{cases} 1 + \gamma \lambda R_q(G_{vis}, t), & q \in \mathcal{H}_{vis}, \\ 1 + \lambda R_q(G_{cot}(t), t), & q \in \mathcal{H}_{cot}, \\ 1, & \text{otherwise.} \end{cases} \quad (8)$$

Here, λ controls the overall intervention strength, while γ is an additional scaling factor used to compensate for the intrinsic magnitude gap between visual-routing ratios $R_q(G_{vis}, t)$ and previous-CoT routing ratios $R_q(G_{cot}(t), t)$. In practice, visual-routing ratios are typically much smaller than CoT-routing ratios, and therefore require additional amplification to achieve comparable intervention intensity across the two routing sources.

When a selected head strongly routes attention to its target source at decoding step t , the corresponding routing ratio becomes larger and the head output receives stronger amplification. Conversely,

if the routing ratio is small, the computation remains close to the original model.

Importantly, we do not modify attention logits or attention probabilities. The attention distribution is computed normally, and only the aggregated output of selected heads is rescaled before the output projection. Therefore, the intervention is conditioned on the model’s own routing behavior during decoding, strengthening active visual-evidence and reasoning-history routes while leaving inactive heads and non-selected heads largely unchanged.

4 Experiments

In this section, we design our experiments around five research questions:

- **RQ1:** Can routing-aware attention intervention improve multimodal long-CoT reliability across models and task domains?
- **RQ2:** Can perturbation-only calibration recover routing maps that are aligned with real-task decoding?
- **RQ3:** Does layer-wise routing stratification provide a robust prior for separating visual-evidence and CoT-history interventions?
- **RQ4:** How sensitive is routing-aware intervention to the amplification strength?
- **RQ5:** What do ablations reveal about the performance gains?

4.1 Experimental Setup

Environment. Experiments are conducted on NVIDIA A800-80GB NVLink GPUs with Xeon Platinum 8374C CPUs and 120GB host memory. Details are provided in the supplementary material.

Datasets. To verify the generality of the proposed method beyond hallucination-only settings, we evaluate not only on hallucination benchmarks, but also on mathematical and multimodal reasoning tasks to study whether routing-aware intervention can improve long-CoT reasoning reliability more broadly. Specifically, we consider three categories of benchmarks: **1. Multimodal Integration:** MMStar (Chen et al., 2024a), SEED-Bench (Li et al., 2023a); **2. Mathematical Reasoning:** MathVista (Lu et al., 2024), MathVision (Wang et al., 2024a); **3. Hallucination Evaluation:** HallusionBench (Guan et al., 2024), POPE (Li et al., 2023b), MMHal (Sun et al., 2024).

Baselines. We compare with representative inference-time hallucination mitigation methods, including VCD (Leng et al., 2024), OPERA (Huang et al., 2024), and AGLA (An et al., 2025). We further compare against the vanilla decoding strategy of each base model.

4.2 RQ1: Main Results

We evaluate the proposed routing-aware attention intervention on three representative multimodal long-CoT models, including Kimi-VL A3B-Thinking, Ocean-R1 7B-Instruct, and R1-Onevision 7B. To comprehensively evaluate the generality of the proposed method, we conduct experiments not only on hallucination-oriented benchmarks, but also on mathematical reasoning and general multimodal reasoning datasets. This setting allows us to study whether routing-aware intervention can improve the overall reliability of multimodal long-CoT reasoning beyond hallucination-only scenarios.

As shown in Table 1, our method consistently improves performance across all evaluated models and datasets. Compared with *Vanilla Decoding*, the proposed method achieves an average improvement of approximately **5%** across all benchmarks. Compared with existing inference-time hallucination mitigation methods, including **VCD**, **OPERA**, and **AGLA**, our method achieves the best performance in **85.7%** of all evaluated model–dataset settings. The improvements are particularly clear on **MathVision**, **HallusionBench**, **POPE**, and **MMHal**, suggesting that dynamically strengthening visual-evidence routing and reasoning-history reuse can effectively improve both reasoning accuracy and hallucination robustness during long-CoT generation. Overall, these results demonstrate that routing-aware intervention provides a robust and general mechanism for improving multimodal long-CoT reasoning reliability across different architectures and task domains.

In addition to effectiveness, the proposed method introduces only minimal deployment overhead. The intervention target set $\mathcal{H}$ is computed once per model using 24 non-semantic 640×640 calibration images (white, black, Gaussian-noise, checkerboard, and random-patch images) paired with the same prompt, and is then shared across all downstream datasets. We use one rollout per calibration sample, select the top 30% heads within the corresponding layer region, and fix the visual compensation factor to $\gamma = 10$. During deployment,

Table 1: **Overall results across seven benchmarks covering general multimodal reasoning, mathematical reasoning, and hallucination evaluation.** MHal reports hallucination rate, where lower is better. Bold indicates the best result and underlined numbers indicate the second best. For our method, we use model-specific hyperparameters: $b = 2, \lambda = 0.4$ for Kimi, and $b = 4, \lambda = 0.25$ for Ocean and R1-Onevision.

Method	Model	General Multimodal Tasks		Mathematical Reasoning		Hallucination Evaluation		
		MMStar	SEED-Bench	MathVista	MathVision	HallusionBench	POPE	MMHal (Hal. Rate ↓)
Vanilla	Ocean-R1 7B Instruct	44.21	59.92	54.29	19.74	48.93	64.33	0.31
VCD		45.71	60.19	54.79	20.63	<u>49.92</u>	66.73	0.28
OPERA		47.87	61.90	56.26	22.43	<u>45.07</u>	67.85	0.29
AGLA		<u>48.57</u>	<u>63.36</u>	<u>57.90</u>	24.34	44.95	<u>68.28</u>	<u>0.27</u>
Ours		49.18	65.72	60.72	<u>22.97</u>	55.82	70.67	0.23
Vanilla	R1- Onevision 7B	55.74	68.75	60.66	32.81	57.97	80.33	0.41
VCD		<u>56.43</u>	68.49	60.18	33.11	<u>58.23</u>	81.28	<u>0.40</u>
OPERA		53.14	66.09	59.52	35.97	53.83	81.93	0.41
AGLA		54.17	67.95	<u>60.76</u>	36.84	55.03	<u>82.11</u>	<u>0.40</u>
Ours		59.82	71.20	62.03	39.86	61.75	83.67	0.38
Vanilla	Kimi-VL A3B Thinking	59.43	65.92	62.98	56.17	65.21	84.67	0.42
VCD		59.92	66.34	63.87	55.88	<u>66.13</u>	84.92	0.39
OPERA		63.24	67.65	65.94	57.02	62.83	84.85	0.36
AGLA		<u>64.29</u>	69.31	67.08	<u>58.68</u>	62.74	<u>85.21</u>	<u>0.34</u>
Ours		65.23	<u>69.15</u>	<u>66.41</u>	61.03	69.59	87.00	0.30

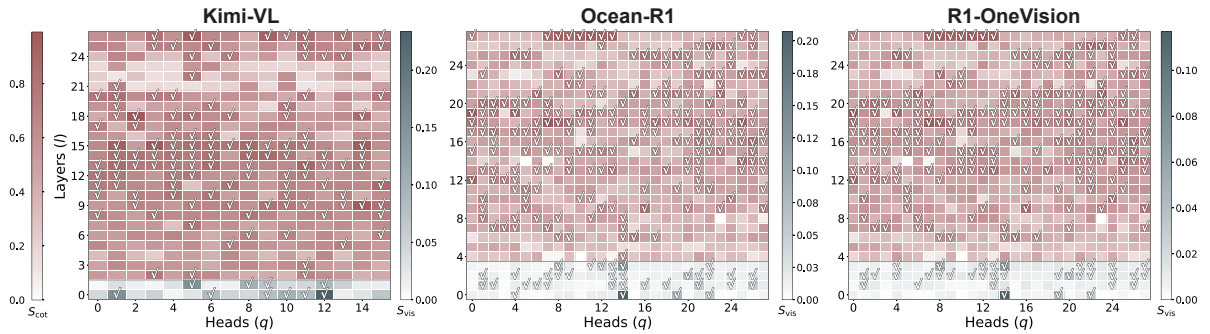


Figure 2: **Visualization of routing-sensitive heads selected from perturbation-based calibration.** For each model, we separately select the top 30% visual-routing heads (Vis) and previous-CoT-routing heads (CoT) according to the routing sensitivity score. The visualization is computed by aggregating routing statistics over all calibration samples, illustrating the global distribution of intervention targets across layers and attention heads.

the method only rescales selected head outputs, increasing end-to-end decoding latency by less than 1% in our implementation with marginal memory and throughput changes.

4.3 RQ2: Validity of Perturbation-Based Routing Maps

Since our method selects intervention heads using only perturbation-based calibration, we further study whether the resulting routing maps remain statistically aligned with routing behavior observed on real-task datasets. Specifically, we compare perturbation-based routing maps with routing maps computed on downstream VQA benchmarks from three complementary perspectives: **Pearson correlation (P)**, which measures the similarity of global routing-score distributions; **Spearman correlation (S)**, which measures whether the relative ranking of

routing-sensitive heads is preserved; and **Top overlap (T)**, which measures whether the final selected intervention heads overlap between perturbation-based and real-task routing maps.

Table 2 reports the PST analysis on Ocean-R1 7B-Instruct, Kimi-VL A3B-Thinking, and R1-Onevision 7B. Overall, perturbation-based routing maps exhibit strong statistical consistency with routing behavior observed on downstream datasets. In particular, the Pearson correlation on **MMStar**, **HallusionBench**, and **MathVista** remains close to **0.9** on average, indicating that the global routing-score distributions are highly aligned between perturbation inputs and real-task decoding. The Spearman correlation further shows that the relative ranking of routing-sensitive heads is largely preserved across datasets. While Ocean-R1 exhibits relatively lower rank-

Table 2: **PST consistency between perturbation-based and real-task routing maps.** The table compares routing maps generated from perturbation-based calibration with routing maps computed on downstream datasets, in order to evaluate whether perturbation-selected routing-sensitive heads remain statistically aligned with real-task routing behavior. **P**, **S**, and **T** denote Pearson correlation, Spearman correlation, and Top-overlap, respectively.

Dataset	Type	Kimi-VL			Ocean-R1			R1-Onevision		
		P	S	T	P	S	T	P	S	T
MMStar	Vis	0.94	0.96	0.90	0.93	0.72	0.65	0.93	0.90	0.74
	CoT	0.95	0.93	0.80	0.89	0.86	0.80	0.88	0.85	0.75
HallusionBench	Vis	0.89	0.94	0.80	0.89	0.55	0.44	0.88	0.80	0.68
	CoT	0.92	0.90	0.77	0.80	0.75	0.69	0.88	0.86	0.75
MathVista	Vis	0.82	0.92	0.90	0.91	0.66	0.56	0.93	0.89	0.71
	CoT	0.88	0.86	0.79	0.72	0.66	0.62	0.90	0.88	0.79

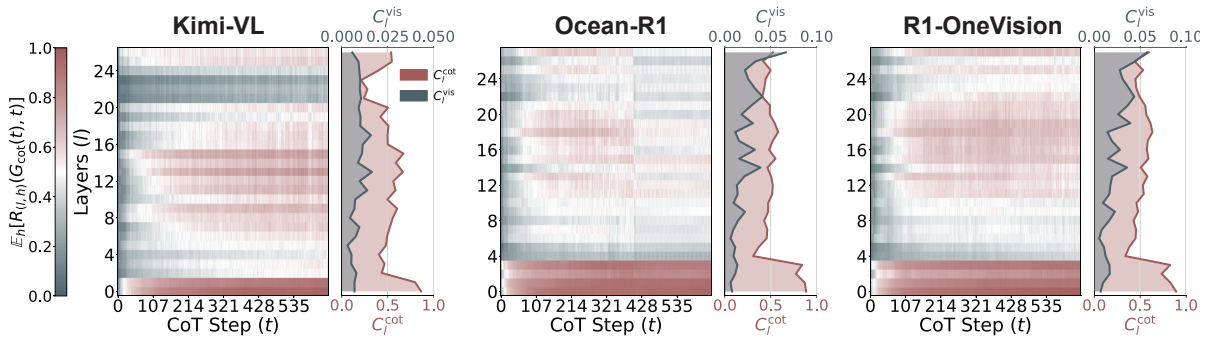


Figure 3: **Layer-wise routing stratification across long-CoT decoding.** The figure visualizes the aggregated routing dynamics on a subset of **HallusionBench** along relative CoT decoding steps. Routing statistics are averaged over all samples within different layer regions, illustrating a clear depth-wise transition from shallow layers to middle layers during long-CoT generation. The observed transition boundary provides empirical evidence for the layer-localized intervention strategy used in our framework.

ing consistency compared with the other models, both Kimi-VL and R1-Onevision maintain consistently high Spearman correlations across benchmarks. Moreover, the Top-overlap metric remains above **50%** across most settings, suggesting that the majority of intervention heads selected from perturbation-based routing maps are also selected under real-task routing distributions.

These results suggest that perturbation inputs can expose stable routing priors of multimodal long-CoT models despite lacking downstream task semantics. Rather than exactly reproducing task-specific routing behavior, perturbation-based calibration captures statistically consistent routing tendencies that remain aligned with real-task decoding patterns, thereby providing an effective lightweight prior for routing-sensitive head selection.

4.4 RQ3: Layer-Wise Routing Stratification

We further analyze whether the proposed layer boundary b reflects a stable structural property of multimodal long-CoT models. Specifically, we aggregate the previous-CoT rout-

ing intensity over all heads within the same layer and compute the layer-wise routing profile C_l^{cot} . Figure 3 visualizes the resulting depth-wise routing distributions for Kimi-VL A3B-Thinking, Ocean-R1 7B-Instruct, and R1-Onevision 7B.

As shown in the figure, all three models exhibit a clear transition pattern between shallow and deep layers. In particular, the transition emerges around layer 2 for Kimi-VL and around layer 4 for both Ocean-R1 and R1-Onevision. Although the exact boundary location differs across architectures, the existence of a clear depth-wise transition remains highly consistent. This suggests that the shallow-to-deep routing boundary is not an accidental phenomenon of a specific model, but rather a recurring structural pattern in multimodal long-CoT reasoning. Such observations provide empirical support for using layer-wise routing stratification as a structural prior for intervention localization.

We further evaluate boundary sensitivity on Ocean-R1 by testing $b - 1 = 3$, the detected boundary $b = 4$, $b + 1 = 5$, and a larger shift

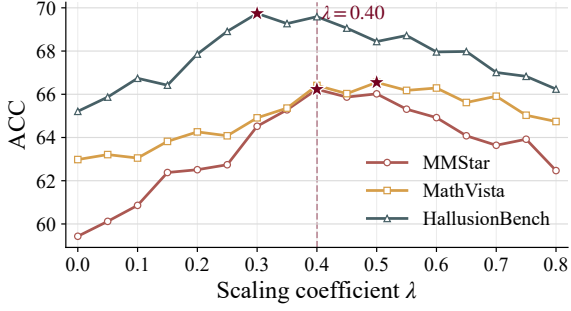


Figure 4: Sensitivity analysis of the scaling coefficient λ on Kimi-VL A3B-Thinking.

$b + 5 = 9$. On MMStar, MathVista, and HallusionBench, these settings obtain (49.91, 58.34, 53.21), (51.43, 60.72, 55.82), (51.92, 61.02, 55.14), and (45.47, 57.96, 52.23), respectively. The neighboring $b + 1$ setting remains close to the detected boundary and slightly improves two datasets, whereas the larger shift degrades more clearly. This suggests that the largest-jump rule identifies a useful boundary region rather than a uniquely optimal layer index.

4.5 RQ4: Sensitivity to Amplification Strength

We further analyze the sensitivity of the amplification factor λ , which controls the strength of routing-aware intervention during decoding. In practice, the optimal λ for each model is selected by evaluating performance on small validation subsets and choosing the configuration that achieves the best overall trade-off across datasets. The final values used in RQ1 are reported in the caption of Table 1. Since the magnitude of $R_q(G_{\text{vis}}, t)$ is relatively small in practice, we additionally apply a $10\times$ scaling factor to the visual-routing branch when computing the final amplification strength.

Figure 4 shows the performance trends of Kimi-VL A3B-Thinking under different λ settings on MMStar, HallusionBench, and MathVista. Although the optimal peak values differ slightly across datasets, the best-performing ranges remain highly consistent within the same model. In particular, the final configuration around $\lambda = 0.4$ achieves stable strong performance across all evaluated benchmarks for Kimi-VL. Moreover, the performance curves exhibit relatively broad plateau regions rather than sharp peaks, indicating that the proposed intervention is not overly sensitive to precise amplification values.

Table 3: Ablation study on Kimi-VL A3B-Thinking. The table reports the performance of different ablation settings.

Variant	MMStar	HallusionBench	MathVista
Vanilla	59.43	65.21	62.98
Random heads	57.24	62.18	61.02
w/o visual heads	63.49	65.91	64.21
w/o CoT heads	64.20	67.83	62.69
Full Ours	66.23	69.59	66.41

4.6 RQ5: Component Ablation

Finally, we conduct component ablations to examine which parts of the proposed intervention are responsible for the final performance gains. We use Kimi-VL A3B-Thinking as the representative model and evaluate on MMStar, HallusionBench, and MathVista. We compare the full method with three ablated variants. First, *Random heads* replaces routing-sensitive heads with randomly selected heads under the same intervention budget and layer regions. Second, *w/o visual heads* removes the visual-evidence branch and only amplifies previous-CoT-sensitive heads. Third, *w/o CoT heads* removes the previous-CoT branch and only amplifies visual-sensitive heads.

As shown in Table 3, the full method achieves the best performance across all three benchmarks. Randomly selected heads perform even worse than vanilla decoding, indicating that the gains do not come from simply rescaling additional attention heads. Instead, amplifying arbitrary heads can disrupt the model’s original routing behavior, which highlights the necessity of selecting heads according to source-specific routing sensitivity.

Removing visual-sensitive heads leads to consistent degradation compared with the full method, especially on MMStar and HallusionBench. This suggests that visual-evidence routing is important for both general multimodal understanding and hallucination mitigation. Meanwhile, removing previous-CoT-sensitive heads causes the largest relative drop on MathVista and even falls slightly below vanilla decoding, indicating that mathematical reasoning relies more heavily on stable reuse of the accumulated reasoning history.

5 Conclusion

We presented a training-free inference-time intervention for multimodal long-CoT reasoning. Rather than enforcing uniform visual attention, our method identifies attention heads that naturally engage with useful information sources and dynami-

cally amplifies their outputs during generation. By strengthening visual-token access in shallow layers and previous-CoT reuse in deeper layers, the method improves multimodal reasoning and hallucination robustness across multiple models and benchmarks. These findings suggest that long-CoT hallucination can be mitigated by stabilizing the model’s own grounding and reasoning mechanisms without additional training or task-specific labels.

Limitation

Although the proposed routing-aware intervention consistently improves multimodal long-CoT reasoning across multiple benchmarks and architectures, several limitations remain. The current framework still relies on manually selected intervention hyperparameters, and the optimal settings may vary across models and tasks. In addition, the method focuses primarily on attention-head-level routing behavior, while other architectural components that may influence hallucination and reasoning stability are not explicitly modeled. Finally, although perturbation-based calibration shows strong alignment with downstream routing behavior, the robustness of the learned routing priors under substantially different domains and interactive reasoning settings remains underexplored.

Ethical Considerations

This work aims to improve the reliability of multimodal long-CoT reasoning by reducing hallucination during generation. The proposed method operates entirely at inference time and does not require additional training data, human annotation, or modification of model parameters. All experiments are conducted on publicly available datasets and open-source models, and no personally identifiable information or sensitive user data is involved. AI assistants were used only for language polishing.

References

- Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. 2025. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29915–29926.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-v1 technical report. *arXiv preprint arXiv:2502.13923*.
- Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.
- Jing Bi, Junjia Guo, Yunlong Tang, Lianggong Bruce Wen, Zhang Liu, Bingjie Wang, and Chenliang Xu. 2025. Unveiling visual perception in language models: An attention head analysis approach. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4135–4144.
- Boqi Chen, Xudong Liu, and Jianing Qiu. 2026a. Mask what matters: Mitigating object hallucinations in multimodal large language models with object-aligned visual contrastive decoding. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 9–16.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and 1 others. 2024a. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087.
- Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. 2026b. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *Science China Information Sciences*, 69(6):161101.
- Xinlong Chen, Yuanxing Zhang, Qiang Liu, Junfei Wu, Fuzheng Zhang, and Tieniu Tan. 2025. Mixture of decoding: An attention-inspired adaptive decoding strategy to mitigate hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8525–8542.
- Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. 2024b. HALC: Object hallucination reduction via adaptive focal-contrast decoding. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*.
- Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations*, volume 2024, pages 54158–54183.
- Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert’s attention. In *Proceedings of the 2019 ACL workshop BlackboxNLP: analyzing and interpreting neural networks for NLP*, pages 276–286.
- Parsa Esmaeilkhani and Longin Jan Latecki. 2026. Direct visual grounding by directing attention of visual tokens. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5787–5797.

- Mehrdad Fazli, Bowen Wei, Ahmet Sari, and Ziwei Zhu. 2026. Caac: Confidence-aware attention calibration to reduce hallucinations in large vision-language models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1242–1251.
- Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14375–14385.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.
- Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427.
- Gangwei Jiang, Yahui Liu, Zhaoyi Li, Wei Bi, Fuzheng Zhang, Linqi Song, Ying Wei, and Defu Lian. 2025a. What makes a good reasoning chain? uncovering structural patterns in long chain-of-thought reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6501–6525.
- Lei Jiang, Chunzhao Xie, Tongxuan Liu, Yuting Zeng, Yunheng Shen, Weizhe Huang, Jing Li, Xiaohua Xu, and 1 others. 2025b. Tarac: Mitigating hallucination in lvlms via temporal attention real-time accumulative connection. *arXiv preprint arXiv:2504.04099*.
- Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. 2025c. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25004–25014.
- Peng Jin, Bo Zhu, Li Yuan, and Shuicheng Yan. 2025. Moh: Multi-head attention as mixture-of-head attention. In *International Conference on Machine Learning*, pages 28233–28255.
- Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. 2025a. See what you are told: Visual attention sink in large multimodal models. In *The Thirteenth International Conference on Learning Representations*.
- Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. 2025b. Your large vision-language model only needs a few attention heads for visual grounding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9339–9350.
- Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of bert. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 4365–4374.
- Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. 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*, pages 13872–13882.
- Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.
- Geng Li, Jinglin Xu, Yunzhen Zhao, and Yuxin Peng. 2025. Dyfo: A training-free dynamic focus visual search for enhancing llms in fine-grained visual understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9098–9108.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 292–305.
- Chengzhi Liu, Yuzhe Yang, Yue Fan, Qingyue Wei, Sheng Liu, and Xin Eric Wang. 2025. Reasoning within the mind: Dynamic multimodal interleaving in latent space. *arXiv preprint arXiv:2512.12623*.
- Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024a. A survey on hallucination in large vision-language models. *CoRR*, abs/2402.00253.
- Jiarui Liu, Weihao Xuan, Zhijing Jin, and Mona Diab. 2026a. Taming object hallucinations with verified atomic confidence estimation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5430–5444.
- Shi Liu, Kecheng Zheng, and Wei Chen. 2024b. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, pages 125–140. Springer.
- Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Jingying Zeng, Zhenwei Dai,

- Zhan Shi, Tianxin Wei, Hanqing Lu, Benoit Dumoulin, and Hanghang Tong. 2026b. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in VLMs. In *The Fourteenth International Conference on Learning Representations*.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, volume 2024, pages 23439–23554.
- Guangtao Lyu, Qi Liu, Chenghao Xu, Jiexi Yan, Muli Yang, Xueting Li, Fen Fang, and Cheng Deng. 2026. Revealing and enhancing core visual regions: Harnessing internal attention dynamics for hallucination mitigation in vlms. *arXiv preprint arXiv:2602.15556*.
- Yunze Man, De-An Huang, Guilin Liu, Shiwei Sheng, Shilong Liu, Liang-Yan Gui, Jan Kautz, Yu-Xiong Wang, and Zhiding Yu. 2025. Argus: Vision-centric reasoning with grounded chain-of-thought. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14268–14280.
- Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32.
- Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, and 1 others. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.
- Shuyi Ouyang, Hongyi Wang, Gongfan Fang, Xinyin Ma, Lanfen Lin, and Xinchao Wang. 2026. Taming the phantom: Token-asymmetric filtering for hallucination mitigation in large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 8206–8214.
- Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Transactions of the association for computational linguistics*, 8:842–866.
- Sirat Samyoun, Yingtai Xiao, and Jian Du. 2026. Attribution-guided multi-object hallucination and bias detection in vision-language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4529–4548.
- Sreetama Sarkar, Yue Che, Alex Gavin, Peter Anthony Beerel, and Souvik Kundu. 2025. Mitigating hallucinations in vision-language models through image-guided head suppression. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12492–12511.
- Arnab Sen Sharma, Giordano Rogers, Natalie Shapira, and David Bau. 2026. LLMs process lists with general filter heads. In *The Fourteenth International Conference on Learning Representations*.
- Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, and 1 others. 2024. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13088–13110.
- Feilong Tang, Chengzhi Liu, Zhongxing Xu, Ming Hu, Zile Huang, Haochen Xue, Ziyang Chen, Zelin Peng, Zhiwei Yang, Sijin Zhou, and 1 others. 2025. Seeing far and clearly: Mitigating hallucinations in mlms with attention causal decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26147–26159.
- Lexiang Tang, Xianwei Zhuang, Bang Yang, Zhiyuan Hu, Hongxiang Li, Lu Ma, Jinghan Ru, and Yuexian Zou. 2026. Not all tokens and heads are equally important: Dual-level attention intervention for hallucination mitigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 9439–9447.
- Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, and 1 others. 2025. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.
- Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4593–4601.
- Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron Wallace, and David Bau. 2024. Function vectors in large language models. In *International conference on learning representations*, volume 2024, pages 17282–17333.
- Rheeya Uppaal, Phu Mon Htut, Min Bai, Nikolaos Pappas, Zheng Qi, and Sandesh Swamy. 2026. Journey before destination: On the importance of visual faithfulness in slow thinking. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4147–4168.
- Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL workshop BlackboxNLP: analyzing and interpreting neural networks for NLP*, pages 63–76.
- Elena Voita, David Talbot, Fedor Moiseev, Rico Senrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 5797–5808.

- Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024a. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169.
- Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024b. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15840–15853.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.
- Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. 2025. Multimodal chain-of-thought reasoning: A comprehensive survey. *CoRR*, abs/2503.12605.
- Yujun Wang, Jinhe Bi, Soren Pirk, Yunpu Ma, and 1 others. 2026. Ascd: Attention-steerable contrastive decoding for reducing hallucination in mllm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 10306–10314.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2025. Llava-cot: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2087–2098.
- Zhe Xu, Zhicai Wang, Junkang Wu, Jinda Lu, and Xiang Wang. 2026. Causal-halbench: Uncovering lvlms object hallucinations through causal intervention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 34169–34177.
- Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. 2025. Understanding and mitigating hallucination in large vision-language models via modular attribution and intervention. In *The Thirteenth International Conference on Learning Representations*.
- Jianghao Yin, Qin Chen, Kedi Chen, Jie Zhou, Xingjiao Wu, and Liang He. 2026. Dynamic multimodal activation steering for hallucination mitigation in large vision-language models. In *The Fourteenth International Conference on Learning Representations*.
- Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2024. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105.
- Liu Yu, Zhonghao Chen, Ping Kuang, Zhikun Feng, Fan Zhou, Lan Wang, and Gillian Dobbie. 2026. Causally-grounded dual-path attention intervention for object hallucination mitigation in lvlms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 36021–36029.
- Zhuoran Yu and Yong Jae Lee. 2025. How multimodal LLMs solve image tasks: A lens on visual grounding, task reasoning, and answer decoding. In *Second Conference on Language Modeling*.
- Xiaofeng Zhang, Yihao Quan, Chaochen Gu, Chen Shen, Xiaosong Yuan, Shaotian Yan, Hao Cheng, Kaijie Wu, and Jieping Ye. 2024a. Seeing clearly by layer two: Enhancing attention heads to alleviate hallucination in lvlms. *CoRR*, abs/2411.09968.
- Xiaofeng Zhang, Yihao Quan, Chen Shen, Chaochen Gu, Xiaosong Yuan, Shaotian Yan, Jiawei Cao, Hao Cheng, Kaijie Wu, and Jieping Ye. 2025. 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*, pages 3512–3534.
- Zhuosheng Zhang, Aston Zhang, Mu Li, hai zhao, George Karypis, and Alex Smola. 2024b. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*.
- Weihong Zhong, Xiaocheng Feng, Liang Zhao, Qiming Li, Lei Huang, Yuxuan Gu, Weitao Ma, Yuan Xu, and Bing Qin. 2024. Investigating and mitigating the multimodal hallucination snowballing in large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11991–12011.
- Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. Analyzing and mitigating object hallucination in large vision-language models. In *International Conference on Learning Representations*, volume 2024, pages 56969–56998.
- Xingyu Zhu, Kesen Zhao, Liang Yi, Shuo Wang, Zhicai Wang, Beier Zhu, Hanwang Zhang, and Xiangnan He. 2026. Look carefully: Adaptive visual reinforcements in multimodal large language models for hallucination mitigation. In *The Fourteenth International Conference on Learning Representations*.
- Jiaru Zou, Ling Yang, Jingwen Gu, Jiahao Qiu, Ke Shen, Jingrui He, and Mengdi Wang. 2026. Reasonflux-prm: Trajectory-aware prms for long chain-of-thought reasoning in llms. *Advances in Neural Information Processing Systems*, 38:26764–26802.