Intervene-All-Paths: Unified Mitigation of LVLM Hallucinations across Alignment Formats

Jiaye Qian[1,2*] **Ge Zheng**[2*] **Yuchen Zhu**[2] **Sibei Yang**[1†]
[1]School of Computer Science and Engineering, Sun Yat-sen University
[2]ShanghaiTech University
Project Page: `https://github.com/SooLab/AllPath`

Abstract

Despite their impressive performance across a wide range of tasks, Large Vision-Language Models (LVLMs) remain prone to hallucination. In this study, we propose a comprehensive intervention framework aligned with the transformer's causal architecture in LVLMs, integrating the effects of different intervention paths on hallucination. We find that hallucinations in LVLMs do not arise from a single causal path, but rather from the interplay among image-to-input-text, image-to-output-text, and text-to-text pathways. For the first time, we also find that LVLMs rely on different pathways depending on the question–answer alignment format. Building on these insights, we propose simple yet effective methods to identify and intervene on critical hallucination heads within each pathway, tailored to discriminative and generative formats. Experiments across multiple benchmarks demonstrate that our approach consistently reduces hallucinations across diverse alignment types.

1 Introduction

Large Vision-Language Models (LVLMs) [36, 37, 3, 4, 86, 6, 8, 9, 73, 78] have emerged as foundational models [53, 11, 51, 87, 13] for general-purpose visual understanding [67, 66, 68, 69, 70, 84, 71, 32, 52, 21] and reasoning [48, 40, 56, 64, 57, 81], demonstrating strong capabilities across a broad range of open-world tasks [19, 48, 58, 49, 50, 20, 31, 34, 12]. They are able to handle diverse question formats [5], from binary judgments and multiple-choice queries to open-ended captions and dialogues. However, despite these advances, LVLMs often suffer from hallucination, generating textual outputs that are inconsistent with the visual input, such as images [30, 80, 23, 18, 41]. This issue underscores a fundamental limitation of LVLMs and raises concerns about user trust and the potential for misinformation. Therefore, tackling hallucination has become a growing research focus, with increasing efforts to uncover its causes and mitigate its impact to enhance the reliability of LVLMs.

The causes of hallucination in LVLMs are inherently more complex due to their foundation on text-only LLMs. Broadly, hallucinations may be inherited from the underlying LLMs or stem from vision-specific components and training. More specifically, several studies have examined the concrete factors and proposed corresponding mitigation techniques. Contrastive decoding attributes hallucination to over-reliance on statistics and unimodal language priors, mitigating it by calibrating LVLM output distributions via comparisons with logits from distorted visual inputs [28, 27, 61]. Intervention-based techniques diagnose internal attention weights or heads in LVLMs and directly intervene, enabling correction during inference with a single forward pass and no additional computational cost [39, 72, 83]. Among them, weight intervention [39, 83] enhances visual grounding by

[*]Equal contribution. Work done during Jiaye's internship at SYSU.
[†]Corresponding author is Sibei Yang.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

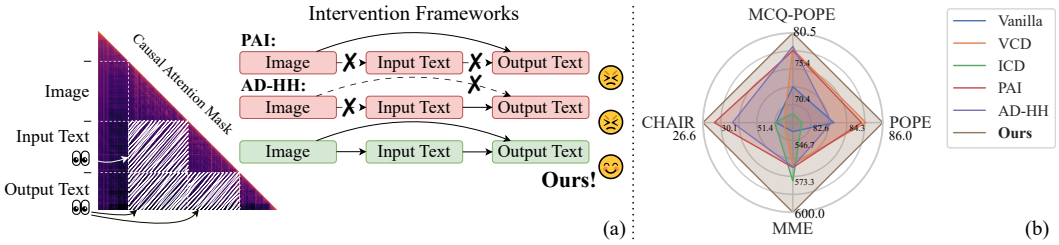


Figure 1: Left: Our AllPath intervention frameowrk comprehensively mitigates hallucinations from image-to-input-text, image-to-output-text, and text-to-text paths. Right: AllPath achieves significant performance improvements over the baselines across all benchmarks.

reinforcing attention to the image, while head intervention [72] mitigates text-dominant behavior by suppressing "lazy" heads that favor text and resemble those in the base LLM after tuning.

In this paper, we explore the causes of hallucination in LVLMs by analyzing their internal components, following the intervention-based line of work. As [72] shows that attention heads have a greater influence on hallucination than MLPs, we focus on heads as the target components. We begin with an empirical observation that motivates our comprehensive intervention framework, which is validated through the identification of critical heads, counterfactual edits, correlation analysis, and empirical evaluation. This process comprises four stages, detailed in below. Notably, our primary analysis focuses on LLaVA-v1.5-7B [36], with additional results for Qwen-VL-Chat [3] and Qwen2.5-VL-7B/72B [4] are provided in the Appendix C.

First, ***we propose a comprehensive intervention framework aligned with the transformer's causal architecture in LVLMs***, based on a thorough evaluation of existing hallucination studies. As shown in Fig. 1(b), a single hallucination mitigation method often struggles to achieve consistently strong performance across different benchmarks [29, 45, 17]. Focusing on intervention methods [27, 61, 39, 72], we observe that they typically intervene along a single causal path—either from input image to output text (PAI in Fig. 1(a)) or from input text to output text (AD-HH in Fig. 1(a))—without considering the system holistically. As a result, certain methods [39, 72] perform better on benchmarks [45] that primarily rely on visual grounding, while others [27] are more effective on benchmarks [29] emphasizing textual coherence. Inspired by this, we propose a comprehensive intervention framework aligned with the input-output sequence and causal transformer architecture of LVLMs, as shown in Fig. 2(Ours). This framework enables joint intervention of the image-to-output-text and image-to-input-text-to-output-text causal pathways. Notably, although the input text is provided rather than generated, its hidden representations are influenced by the input image.

Second, ***we identify the heads in LVLMs that govern distinct intervention pathways within the comprehensive framework by introducing two novel methods***, each targeting the image-to-text and text-to-text causal paths, respectively (detailed in Sec. 2). Although the framework involves multiple causal paths, they fall into two main types: text-to-text (input to output text) and image-to-text (from image to input or output text). For the former type, as textual output logits are primarily influenced by preceding text tokens [38], we identify the path's heads by measuring their differential impact on the logits of hallucinated versus non-hallucinated tokens. For the latter type, the image's influence on text is reflected in the attention contributions from image tokens to text tokens. Path heads are identified by assessing whether they attend to image tokens that are semantically aligned with the corresponding text. Notably, both of our methods are significantly faster than previous approaches for identifying hallucination-related heads, detailed in Sec. 2.4

Third, we validate the impact and roles of the identified heads both within individual causal paths and in their combined pathways, revealing key insights into the comprehensive intervention framework (detailed in Sec. 3). We find that:

- ***Text-to-text intervened heads primarily contribute to LVLM alignment, particularly in adhering to instruction-following formats.*** Heads are highly correlated and often overlap in questions with similar formats, even across different benchmarks, but differ significantly in more distinct formats (e.g., binary vs. open-ended captioning).

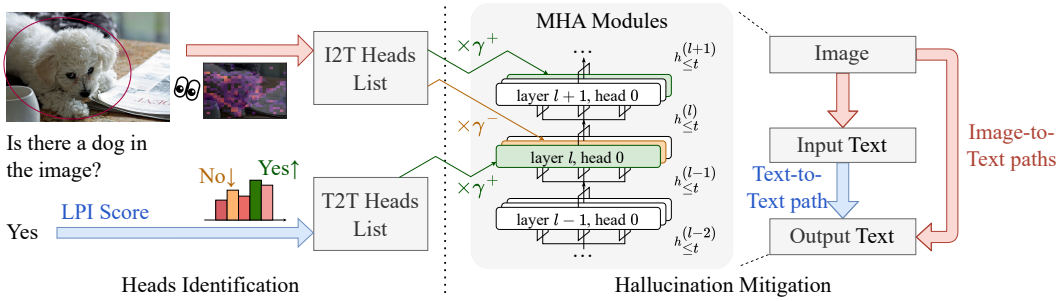


Figure 2: The overview of our proposed AllPath. AllPath first identifies the most critical text-to-text and image-to-text heads contributing to hallucinations using the Log Probability Increase (LPI) score and the ratio of key object attention to total image attention. Then, by applying adaptive heads interventions, AllPath mitigates hallucinations by manipulating the casual pathways in LVLMs.

- ***The intervened heads along the image-to-input-text and image-to-output-text paths differ***, despite both exhibiting stronger attention to image tokens than to text tokens. This suggests that, although both paths "see" the image, the heads governing input and output text function independently.
- ***LVLMs are "smarter" than expected, adaptively selecting the most suitable combined pathways based on the question type***—such as image-to-output-text, image-to-input-text-to-output-text, or both—depending on the relative visual and textual demands of the question.

Finally, ***we propose a simple yet effective intervention method for mitigating hallucination***, grounded in our earlier findings. It adaptively selects the appropriate pathway(s) within the comprehensive intervention framework based on question type and intervenes on the corresponding heads. Experiments demonstrate that our method consistently improves performance across diverse benchmarks while maintaining high efficiency.

In summary, our contributions are multifold: (1) To the best of our knowledge, we are the first to propose a multi-path hallucination intervention framework for LVLMs, going beyond prior single-path approaches. (2) We introduce two novel methods to identify critical heads across different causal pathways. (3) Through counterfactual edits and correlation analysis on attention heads, we uncover several novel findings—reported for the first time—that offer insights into the internal mechanisms of LVLMs. (4) Built upon this framework, we demonstrate that simple interventions on selected heads yield consistent improvements across benchmarks with diverse question and alignment formats, while maintaining high efficiency.

2 Probing image-to-text and text-to-text heads along causal paths

Building on the intervention framework, we propose two lightweight and plug-and-play methods to identify critical attention heads associated with each type of causal pathway—text-to-text and image-to-text. An overview of the our probing method is illustrated in Fig. 2.

2.1 Preliminaries

We consider an LVLM composed of an image encoder, a modality connector and a base LLM with L transformer layers. At each generation time step t, the LVLM first encodes the input image I, text prompt P, and the previously generated response tokens $R_{\leq t}$ into initial embeddings $h^{(0)}_{\leq t}$. This embedding sequence is then progressively updated through the transformer layers, with the hidden states at layer l updated as follows:

$$h^{(l)}_{\leq t} = h^{(l-1)}_{\leq t} + H^{(l)}_{\leq t} + F^{(l)}_{\leq t}$$

where $H^{(l)}_{\leq t}$ and $F^{(l)}_{\leq t}$ are the outputs of the multi-head attention (MHA) module and the feed-forward network (FFN), respectively. The final hidden state at the current time step, $h^{(L)}_t$, is then used to predict the next token r_{t+1} , which is sampled from:

$$\mathbb{P}(r_{t+1} \mid h^{(L)}_t) = \text{softmax}(\text{Unembed}(h^{(L)}_t))$$

Algorithm 1: Get Scores Set for Short Answering Questions

for *each sample, each time step t in the model generation process* **do**
  if *t is the first time step in generation* **then**
    $\mathcal{B}_t^+ \leftarrow$ correct answer token of the question;
    $\mathcal{B}_t^- \leftarrow$ incorrect answer token(s) of the question;
  else
    $\mathcal{B}_t^+, \mathcal{B}_t^- \leftarrow \emptyset$;

Algorithm 2: Get Scores Set for Open-ended Questions

for *each sample, each time step t in the model generation process* **do**
  $\mathcal{B}_t^+, \mathcal{B}_t^- \leftarrow \emptyset$;
  if *r_t is judged as hallucination* **then**
    $\mathcal{B}_t^- \leftarrow \{r_t\}$;
  else if *r_t is judged as non-hallucination* **then**
    $\mathcal{B}_t^+ \leftarrow \{r_t\}$;

In this paper, we focus on the MHA module in transformer layers. Let N denote the number of attention heads employed in each layer. For a specific head indexed by (l, n), where $l \in \{1, 2, \ldots, L\}$ is the layer index and $n \in \{1, 2, \ldots, N\}$ is the head index within that layer, we denote the corresponding query, key, value, and output projection matrices as $W_Q^{(l,n)}$, $W_K^{(l,n)}$, $W_V^{(l,n)}$, and $O^{(l,n)}$, respectively. Given the layer input $h_{\leq t}^{(l-1)}$, the output $H_{\leq t}^{(l,n)}$ is computed as:

$$A_{\leq t}^{(l,n)} = \mathrm{softmax}\left(\frac{W_Q^{(l,n)} h_{\leq t}^{(l-1)} (W_K^{(l,n)} h_{\leq t}^{(l-1)})^\top}{\sqrt{d_k}}\right)$$
$$H_{\leq t}^{(l,n)} = A_{\leq t}^{(l,n)} \cdot W_V^{(l,n)} h_{\leq t}^{(l-1)} \cdot O^{(l,n)}$$

where $\sqrt{d_k}$ is the scaling factor and $A_{\leq t}^{(l,n)}$ is the attention weight for head (l, n). The final output of MHA at layer l is the sum of N head output, and can be expressed as $H_{\leq t}^{(l)} = \sum_{i=1}^{N} H_{\leq t}^{(l,i)}$.

2.2 Discovery of crucial attention heads for the text-to-text path

Given that the textual output relies heavily on preceding text tokens, output probability distribution reflects the model's confidence in hallucinated content, we refine the Log Probability Increase (LPI) score [76, 77] to quantitatively assess the influence of individual model components in promoting or mitigating hallucinations. Specifically, for a generated token r_t at time step t, the LPI score of attention head (l, n) is calculated as:

$$\mathrm{logProb}_\uparrow{}^{(l,n)}(r_t) = \log \mathbb{P}(r_t \mid h_t^{(l-1)} + H_t^{(l,n)}) - \log \mathbb{P}(r_t \mid h_t^{(l-1)})$$

which measures the contribution of head (l, n) to token r_t. To further capture the effect of the head over a broader set of tokens, we extend the definition of LPI score by considering a candidate vocabulary set $\mathcal{B}_t$ at time t, and compute the modified score as:

$$\mathrm{logProb}_\uparrow{}^{(l,n)}(\mathcal{B}_t) = \log \sum_{b \in \mathcal{B}_t} \mathbb{P}(b \mid h_t^{(l-1)} + H_t^{(l,n)}) - \log \sum_{b \in \mathcal{B}_t} \mathbb{P}(b \mid h_t^{(l-1)})$$

At the decoding step t, we designate a hallucination set $\mathcal{B}_t^-$ and a non-hallucination set $\mathcal{B}_t^+$. Specifically, for short answering questions, we prompt the model to produce a direct response and analyze the first decoding step, assigning the correct answer to $\mathcal{B}_t^+$ and all incorrect alternatives to $\mathcal{B}_t^-$. For open-ended questions, we focus on positions that have been identified as hallucinated or non-hallucinated. The corresponding token is assigned exclusively to either $\mathcal{B}_t^-$ or $\mathcal{B}_t^+$, with the other set left empty (see Alg. 1 and 2 for details). Then, for each attention head (l, n), we compute the LPI score on both sets, and then average the scores over all decoding time steps across all samples:

$$S_{\text{T2T}}^{(l,n),-} = \mathbb{E}_{\mathcal{B}_t^- \neq \emptyset}[\mathrm{logProb}_\uparrow{}^{(l,n)}(\mathcal{B}_t^-)], \qquad S_{\text{T2T}}^{(l,n),+} = \mathbb{E}_{\mathcal{B}_t^+ \neq \emptyset}[\mathrm{logProb}_\uparrow{}^{(l,n)}(\mathcal{B}_t^+)]$$

We then define the **Text-to-Text (T2T) Score** for each head as the difference between the two:

$$S_{\text{T2T}}^{(l,n)} = S_{\text{T2T}}^{(l,n),+} - S_{\text{T2T}}^{(l,n),-}$$

where a smaller value indicates a stronger tendency of head (l, n) to promote hallucinations.

2.3 Discovery of crucial attention heads for the image-to-text path

In addition to the textual context, visual information is another crucial factor influencing the model's responses. To identify Image-to-Text (I2T) attention heads, we first select a set of tokens which are most important for grounding the image, denoted as $\mathcal{T}_{\text{I2T}}$. Specifically, since LVLMs are causal models that later tokens integrate information from earlier ones through attention mechanism, we focus only on the *first occurrence* of each object during the question-answering process, either the input text P or the output response $R_{\leq t}$, as these initial mentions are expected to be grounded in the relevant image regions rather than inferred from prior textual context. This makes these selected tokens especially useful for probing I2T heads.

To distinguish whether the corresponding objects of these tokens are present in the image, we then partition $\mathcal{T}_{\text{I2T}}$ into two disjoint subsets: $\mathcal{T}^{+}_{\text{I2T}}$ and $\mathcal{T}^{-}_{\text{I2T}}$, where $\mathcal{T}^{+}_{\text{I2T}}$ contains tokens whose objects are present in the image, and $\mathcal{T}^{-}_{\text{I2T}}$ contains tokens whose objects are absent. When $t \in \mathcal{T}^{+}_{\text{I2T}}$, we expect the attention to be both strong and concentrated within the relevant region. Accordingly, we define the positive score $S^{(l,n),+}_{\text{I2T}}$ as the average summed attention scores within the aligned region M_r across all image tokens. Conversely, for $t \in \mathcal{T}^{-}_{\text{I2T}}$, the attention is expected to be weaker and more diffuse. In this case, we define the negative score $S^{(l,n),-}_{\text{I2T}}$ as the average summed attention scores over the entire image region. Formally, we compute:

$$S^{(l,n),+}_{\text{I2T}} = \mathbb{E}_{r \in \mathcal{T}^{+}_{\text{I2T}}}\Big[\sum_{i \in M_r} A^{(l,n)}_{r,i}\Big], \qquad S^{(l,n),-}_{\text{I2T}} = \mathbb{E}_{r \in \mathcal{T}^{-}_{\text{I2T}}}\Big[\sum_{i \in I} A^{(l,n)}_{r,i}\Big]$$

Finally, similar to Text-to-Text Score, we obtain the **Image-to-Text (I2T) Score** for each head as:

$$S^{(l,n)}_{\text{I2T}} = S^{(l,n),+}_{\text{I2T}} - S^{(l,n),-}_{\text{I2T}}$$

2.4 Efficiency analysis of head probing methods

Previous work identifies hallucination-related attention heads via either a zero-out strategy [72] or training-based methods [15]. However, both approaches incur substantial computational overhead. The zero-out strategy requires ablating each attention head at every transformer layer to estimate its contribution to hallucination, necessitating a full forward pass at each generation time step t for every head. Meanwhile, training-based methods involve training an additional classifier on a large annotated dataset.

In contrast, our method computes Text-to-Text (T2T) Score and Image-to-Text (I2T) Score for each attention head based solely on local behavior—the log-probability increase at that layer or attention shifts within the layer—without requiring full inference. This allows us to obtain all heads' T2T and I2T scores with only a single forward pass, significantly enhancing both efficiency and practicality.

3 Analysis and discussion on identified heads in different paths

In this section, we determine the specific roles of the heads we identified by analyzing their statistical information and behaviors. First, in Sec. 3.1, we investigate the behaviors of the heads identified by our two methods and examine how they contribute to the text-to-text and image-to-text causal pathways, respectively. Subsequently, in Sec. 3.2, we show that LVLMs adaptively select different pathways across benchmarks, and that pathway choice significantly affects hallucination occurrence and mitigation across question types.

3.1 Functional interpretation of the identified heads

Text-to-text causal path is related to the alignment. To investigate the relationships between the heads identified in different datasets, we independently identify heads within each dataset and analyze the correlations between them. Specifically, we focus on comparing the heads identified in the POPE [29] COCO [33] adversarial dataset with those found in other datasets. For each dataset, we rank the Importance Scores $S^{(l,n)}_{\text{T2T}}$ and $S^{(l,n)}_{\text{I2T}}$ for each head. We then pairwisely plot the ranking positions of each head on a two-dimensional graph and calculate the linear regression correlation

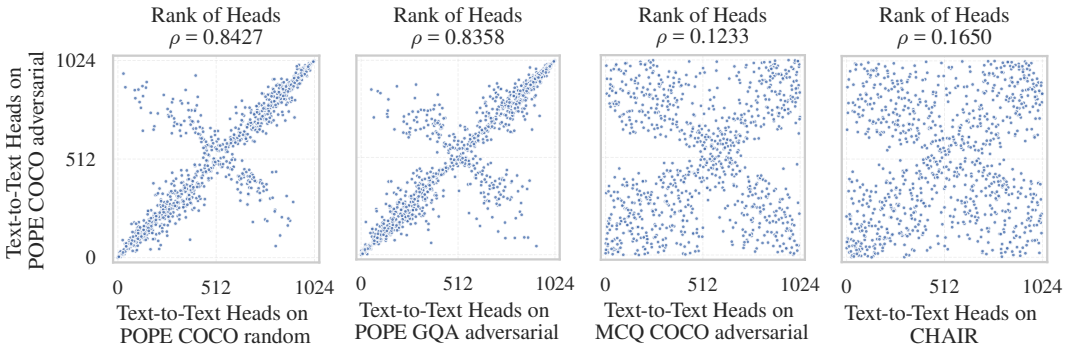


Figure 3: Visualization of the rank distributions of text-to-text heads extracted from different datasets. ρ denotes the correlation coefficient; a higher value (i.e., points concentrated along the diagonal) indicates greater similarity between the two sets of heads. Heads identified from datasets with similar question–answer alignment formats exhibit strong correlation and substantial overlap.

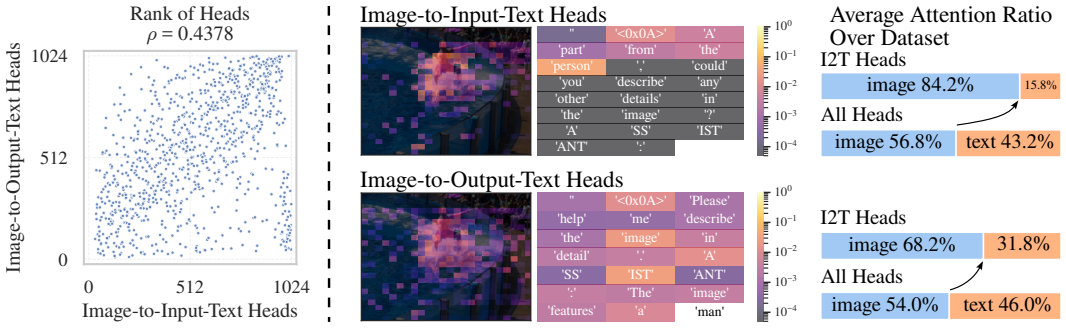


Figure 4: Left: Visualization of the rank distributions of image-to-input-text heads and image-to-output-text heads, showing that the two are largely uncorrelated. Right: Compared to the average of all heads, all image-to-text heads we identified, exhibit a stronger focus on visual content.

coefficient ρ between them. A higher ρ value indicates a greater similarity in the ranking of heads across the datasets being compared.

As shown in Fig. 3, we observe that ***the correlation between text-to-text heads under the same alignment format is very high, but the correlation is significantly lower under different formats.*** Specifically, for Yes/No questions, the correlation coefficient between the heads identified by POPE COCO adversarial and POPE COCO random reaches 0.8247, and the correlation coefficient between the heads identified by POPE COCO adversarial and POPE GQA [22] adversarial reaches 0.8358. However, even with the same images and similar questions, when switching to MCQ format, the correlation coefficient between heads drops to only 0.1233. This demonstrates that our T2T heads exhibit a strong correlation with the alignment format.

Image-to-input-text causal path is different from image-to-output-text causal path. For the image-to-text path, we further investigate the behavioral differences between the heads in the image-to-input-text and image-to-output-text pathways. Specifically, we conduct experiments using CHAIR. Unlike the standard CHAIR [45] evaluation, we design a modified prompting strategy: we make certain objects to appear in the input while prompting the LVLM to generate captions for other objects. We then extract the heads separately from the input and output paths and calculated the correlation coefficient ρ between these two sets of heads. Additionally, we calculate the average attention weights of the top-10 identified heads and compared them with the average attention weights of all heads.

As illustrated in Fig. 4, we observe that both image-to-input-text heads and image-to-output-text heads attend to the image more significantly than other heads. ***However, even within the image-to-text causal path, the heads identified in the image-to-input-text and image-to-output-text pathways differ considerably.*** This suggests that focusing solely on the output text when mitigating hallucinations may be insufficient, even though hallucinations only occur in the output.

Method	POPE (COCO)						CHAIR	
	Random		*Popular*		*Advers.*			
	Acc.	F1	Acc.	F1	Acc.	F1	$C_S\downarrow$	$C_I\downarrow$
w/o mask	85.1	83.7	83.7	82.4	80.9	80.0	52.2	14.6
w/ mask	83.2	81.2	82.0	80.2	79.7	78.1	62.4	32.4
$\Delta\%$	$\downarrow$1.9	$\downarrow$2.5	$\downarrow$1.7	$\downarrow$2.2	$\downarrow$1.2	$\downarrow$1.9	$\uparrow$10.2	$\uparrow$17.8

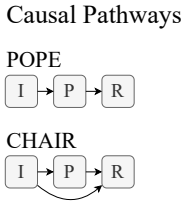


Figure 5: Left: When we entirely knocked out the attention weights of output tokens attending to image tokens, POPE's performance worsens very little, whereas CHAIR's performance gets worse significantly. Right: This indicates that LVLMs utilize different pathways for different question formats.

3.2 LVLMs adaptively select different causal pathways for different question formats

Motivation & experiment setting. Given that both the image-to-input-text-to-output-text and image-to-output-text pathways may contribute to the prediction of final output tokens (*i.e.*, the response), we further investigate whether LVLMs exhibit different pathway utilization patterns across different types of hallucination benchmarks. To demonstrate the significance of both causal pathways in LVLM hallucinations, following [26], we attempt to remove the image-to-text causal path from the image input to the output while preserving the image-to-text causal path from image input to text input and the text-to-text causal path from text input to output. Specifically, we utilize LLaVA [36] to mask the attention from output tokens to image tokens on both POPE and CHAIR, as shown in Fig. 5.

Finding & discussion. We observe that on POPE, even when the image input to the output attention is entirely knocked out, performance only decreases by approximately 2%, underscoring the importance of the text-to-text causal path. However, on CHAIR, consistent with the findings of [26], the performance worsens by about 10% after knocking out, indicating the substantial contribution of the image-to-text causal path. This experiment also suggests that ***the causal paths taken by LVLM in answering questions differ across different question formats.*** Based on these findings, we contend that focusing solely on strengthening one causal path is insufficient for mitigating LVLM hallucinations comprehensively.

4 Mitigating hallucinations through multi-path head intervention

In this section, building on the heads identified in our intervention framework (Sec. 2) and their roles in different causal pathways (Sec. 3), we propose a simple yet effective intervention method to reduce hallucinations.

Important claim: To demonstrate that our study is general and not specific on questions, **when identifying heads, we use images and questions that are entirely different from those used during testing.** Specifically, for each tested dataset, we generate a new dataset and calculate importance scores on it, ensuring that this newly generated dataset shares only the same format with the original dataset without sharing any images.

To identify attention heads with the most pronounced impact on hallucination, we use Importance Scores $S_{\text{T2T}}^{(l,n)}$ and $S_{\text{I2T}}^{(l,n)}$ as indicators for head selection. Specifically, we select Z_{T2T}^{+} and Z_{T2T}^{-}, the top and bottom ξ attention heads ranked by alignment scores, which are associated with hallucination suppression and promotion, respectively. Similarly, we select Z_{I2T}^{+}, the top ζ attention heads ranked by image scores $S_{\text{I2T}}^{(l,n)}$, which exhibit the desirable attention patterns over visual inputs.

Finally, we derive the final sets of attention heads by combining the respective conditions:

$$Z^{-} = Z_{\text{T2T}}^{-}, \qquad Z^{+} = Z_{\text{I2T}}^{+} \cup Z_{\text{T2T}}^{+}$$

To effectively mitigating hallucinations, we apply different scaling factors $\lambda^{(l,n)}$ to different heads according to Z^{-} and Z^{+}. Specifically, we apply this modification to MHA, the updated output of the

MHA module can be expressed as follows:

$$\tilde{H}^{(l)}_{\leq t} = \sum_{n=1}^{N} \lambda^{(l,n)} H^{(l,n)}_{\leq t}, \qquad \lambda^{(l,n)} = \begin{cases} \gamma^+, & \text{if } (l,n) \in Z^+ \\ \gamma^-, & \text{if } (l,n) \in Z^- \\ 1, & \text{otherwise} \end{cases}$$

where γ^+ and γ^- are hyperparameters.

5 Experiments

5.1 Experimental settings

5.1.1 Datasets and metrics

POPE [29], the Polling-based Object Probing Evaluation (POPE) benchmark is a widely utilized benchmark for evaluating object hallucination in LVLMs. Built upon three established datasets, COCO [33], A-OKVQA [46], and GQA [22], POPE samples 500 images from each dataset and selects three ground-truth objects per image as positive instances. Correspondingly, three absent objects are sampled per image using one of three strategies: random, popular, and adversarial sampling. Each object is then queried in the binary format: "*Is there a(n) {object} in the image?*" This yields 3,000 questions per dataset, with a balanced distribution of positive and negative instances.

MCQ-POPE. To extend the format of questions, we adapted POPE into multiple-choice questions, creating a new benchmark named MCQ-POPE. Specifically, for each image in each dataset each split, we extract all the three objects labeling Yes and three labeling No. For each object, we pair it with three other objects whose labels differ from the target object, shuffle the four objects, and present them in a multiple-choice question. The resulting MCQ-POPE benchmark also maintains a total of 3,000 questions for each dataset each split, ensuring consistency with the original POPE benchmark.

CHAIR [45], the Caption Hallucination Assessment with Image Relevance is another widely used metric for evaluating hallucinations in LVLMs, focusing on open-ended image captioning tasks. Objects mentioned in the generated captions but absent from the ground truth are regarded as hallucinations. CHAIR evaluates such errors along two dimensions: sentence-level and instance-level, with their respective scores computed as follows:

$$\mathrm{C_S} = \frac{|\{\text{captions with hallucinated objects}\}|}{|\{\text{all captions}\}|} \quad \text{and} \quad \mathrm{C_I} = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all mentioned objects}\}|}$$

MME [17], the MLLM Evaluation Benchmark is a benchmark designed to comprehensively evaluate the performance of LVLMs. Similar to POPE, MME also requires the model to answer yes or no questions. Following [74, 27], and given our focus on hallucination, we report results on the hallucination subsets, specifically including existence, count, position, and color subsets. Consistent with their official implementation, we use the sum of accuracy and accuracy+ as the evaluation metric.

Baselines. We conduct experiments on the LLaVA-1.5-7B [36]. In addition to the vanilla baseline, we also compared AllPath with some other training-free methods, including contrastive-decoding-based VCD [27] and ICD [61], intervention-based PAI [39] and AD-HH [72].

5.1.2 Implementation details

Unless otherwise specified, we set the scaling factors to $\gamma^+ = 2.0$ and $\gamma^- = 0.0$. For short-answering tasks, we set $\xi = 20$ and $\zeta = 10$, while for open-ended tasks, we set $\xi = 40$ and $\zeta = 50$. For MME, we employ the same attention heads identified from POPE. See Appendix A for more details.

5.2 Comparison with state-of-the-art methods

As shown in Table 1, our AllPath exhibits the highest performance across all benchmarks with different formats. It can be observed that on the POPE dataset, our method achieves at least a 2.1% increase in F1, significantly higher than the best baseline improvement of 0.9%. On MCQ-POPE, our method also achieves at least a 5.2% increase. On the CHAIR dataset, we also achieve the highest F1 increase of 2.6%, exceeding the best baseline improvement of 1.9%. Additionally, we find that while

Method	POPE (COCO)						MCQ-POPE (COCO)						CHAIR	
	Random		*Popular*		*Advers.*		*Random*		*Popular*		*Advers.*			
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	$C_S\downarrow$	$C_I\downarrow$
Vanilla	85.1	83.7	83.7	82.4	80.9	80.0	72.8	72.9	68.0	68.0	64.4	64.5	52.2	14.6
VCD [27]	86.3	85.2	84.5	83.5	81.4	80.9	78.2	78.2	72.2	72.2	67.9	67.9	58.2	16.1
ICD [61]	84.4	82.2	83.8	81.6	81.5	79.5	69.1	69.1	67.2	67.2	64.7	64.6	51.4	14.7
PAI [39]	86.4	85.0	84.9	83.6	82.5	81.4	78.0	78.0	71.2	71.2	68.0	67.9	28.8	7.9
AD-HH [72]	85.0	83.6	83.7	82.4	80.9	79.9	78.5	78.6	71.1	71.1	68.1	68.1	33.2	7.5
AllPath	**87.2**	**86.0**	**86.0**	**84.9**	**82.8**	**82.1**	**80.5**	**80.5**	**74.0**	**74.0**	**69.7**	**69.7**	**26.6**	**7.2**

Table 1: Results on POPE, MCQ-POPE, and CHAIR benchmark. The F1 score for MCQ-POPE represents Macro F1.

Method	Object-Level		Attribute-Level		Total↑
	Existence↑	*Count*↑	*Position*↑	*Color*↑	
Vanilla	180.0 (±8.66)	113.9 (±10.05)	116.7 (±15.90)	129.4 (±18.28)	540.0 (±39.69)
VCD [27]	177.8 (±2.55)	122.8 (±20.84)	122.2 (±1.92)	141.7 (±2.89)	564.4 (±20.84)
ICD [61]	185.0 (±5.00)	121.7 (±2.89)	118.3 (±6.01)	151.7 (±7.64)	576.7 (±11.67)
PAI [39]	185.0 (±5.00)	122.8 (±14.94)	114.4 (±4.19)	144.4 (±5.09)	566.7 (±15.90)
AD-HH [72]	179.4 (±5.09)	117.8 (±12.51)	116.1 (±13.37)	152.8 (±10.72)	566.1 (±22.19)
AllPath	**188.3** (±2.89)	**126.1** (±0.96)	**132.2** (±8.39)	**153.3** (±7.64)	**600.0** (±8.66)

Table 2: Performance on MME Benchmark [17]. Results are reported as the average of 3 test runs, with ± indicating the sample standard deviation.

VCD shows noticeable improvements on POPE and MCQ-POPE, its performance on CHAIR suffers a significant decline of 1.4%. AD-HH, on the other hand, exhibits substantial performance gains on CHAIR and MCQ-POPE, but its performance on POPE is similar to that of the vanilla method. Our method achieves substantial performance gains across three benchmarks of completely different formats.

Our results on MME is shown in Table 2. It can be observed that our method still achieves stable performance improvement on the MME hallucination subset. Considering that we still use the POPE heads for MME without extracting independent heads specifically for MME, this demonstrating the generalizability of our method under a unified task format.

5.3 Ablation study

Ablation study on pathways. To evaluate the two key components of our method, the T2T heads and I2T heads, we conduct an ablation study on POPE. The results are shown in the upper part of Table 3. We observe that employing either T2T heads or I2T heads individually leads to a certain degree of performance improvement. However, the most substantial performance gain is achieved only when all heads are utilized, significantly outperforming the previous two configurations. This result substantiates our claim that both proposed pathways are present and exert a considerable influence on LVLM hallucinations. Therefore, considering only one pathway is insufficient for effectively mitigating LVLM hallucinations.

Ablation study on hyperparameters. To validate that our AllPath is insensitive to hyperparameter selection, we perform an ablation study on POPE COCO. As results report in the lower part of Table 3, our method is largely insensitive to the choice of hyperparameters. Substantial changes to the hyperparameters lead to only minor variations in performance. For instance, the accuracy on the random subset fluctuates only within the range of 85.6–88.3, consistently outperforming the baseline of 85.1. Notably, some hyperparameter settings in this table result in better performance on POPE. This is because rather than tuning hyperparameters for a specific dataset, we select a single set of hyperparameters that works best across various discriminative tasks. This shows that our method is effective and stable without extensive tuning.

Hyperparameters				*Random*		*Popular*		*Adversarial*	
ξ	ζ	γ^+	γ^-	Acc.	F1	Acc.	F1	Acc.	F1
20	10	2.0	0.0	87.2	86.0	86.0	84.9	82.8	82.1
0	10	2.0	0.0	86.2	85.2	84.5	83.6	81.6	81.1
20	**0**	2.0	0.0	86.4	84.9	85.2	83.7	82.5	81.3
0	**0**	2.0	0.0	85.1	83.7	83.7	82.4	80.9	80.0
10	10	2.0	0.0	87.3	86.2	86.0	85.0	82.6	82.0
30	10	2.0	0.0	88.3	87.7	86.7	86.3	82.9	83.0
40	10	2.0	0.0	88.1	87.9	85.9	86.0	81.2	82.2
20	**5**	2.0	0.0	86.2	84.8	85.8	84.6	82.7	81.9
20	**15**	2.0	0.0	88.2	87.4	86.6	85.9	83.0	82.8
20	**20**	2.0	0.0	88.1	87.3	86.7	86.0	83.0	82.8
20	10	**1.2**	0.0	85.6	85.3	85.3	84.1	82.3	81.5
20	10	**1.5**	0.0	86.9	85.7	85.6	84.5	82.5	81.7
20	10	**2.5**	0.0	86.8	85.6	86.3	85.4	83.1	82.5
20	10	2.0	**0.2**	87.1	85.9	85.8	84.7	82.5	81.8
20	10	2.0	**0.5**	87.0	85.8	85.6	84.5	82.2	81.5
20	10	2.0	**0.8**	86.8	85.5	85.4	84.2	82.0	81.3

Table 3: Ablation studies. Hyperparameters different from the default setting are highlighted.

6 Related work

Large Vision-Language Models. With the success of LLMs [1, 59, 60, 10, 44, 2], an increasing number of studies focus on LVLMs [36, 3, 86, 8, 9, 73]. A typical LVLM architecture comprises a vision encoder [43] with an LLM, connected via a modality alignment module such as a linear projection [36] or a Q-former [86, 14]. To support effective cross-modal reasoning, these models follow a multi-stage training pipeline. The pretraining stage focuses on aligning visual and language modalities using large-scale image-text pairs. Building on this foundation, supervised fine-tuning (SFT) with instruction-following data enhances the model's ability to generate coherent responses conditioned on multimodal prompts. Beyond standard training recipe, several works have further explored advanced adaptation strategies, including multi-phase pre-training [4], knowledge learning stage [37], and Reinforcement Learning from Human Feedback [55].

Mitigating hallucinations in LVLMs. Extensive researches have been conducted to mitigate LVLM hallucinations from various perspectives, including curating training datasets of models [35, 75, 79], modifying training objectives [24], adjusting model architectures [65], and applying post-processing methods to model outputs [85, 63]. Nevertheless, many studies also focus on mitigating hallucinations during inference. Some work [16, 27, 62, 61, 82] primarily focused on Contrastive Decoding [28] methods, while more recent approaches have focused on attention mechanisms, such as applying straightforward scaling techniques [39] or by dynamically adjusting scaling weights [25] on attention scores for a range of layers and tokens. Meanwhile, some approaches focus on modifying attention heads. [72] identifies hallucinatory heads by zero-ablation, and [7] identifies hallucinatory heads by modifying input image tokens and training a binary classifier.

Understanding LVLMs. Currently, methods for understanding LVLMs can mainly be categorized into three types. The first type is based on averaging transformer weights. However, [54, 72] demonstrates that they do not effectively capture the specific regions that the model actually attends to. Additionally, Grad-CAM [47] visualize the model's focus by calculating the gradient of the output with respect to the input. Another approach is the early exit method [42, 72], which obtains hidden state representations by extracting them before the unembedding process.

7 Conclusion

In this paper, we propose a unified framework for mitigating hallucinations in LVLMs. By distinguishing between Text-to-Text and Image-to-Text pathways and analyzing their corresponding attention heads, we develop AllPath, a simple yet powerful method. Experimental results demonstrate that AllPath consistently achieves superior performance across all benchmarks.

Acknowledgment: This work is supported by the National Natural Science Foundation of China under Grant No.62206174 and No.62576365.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[5] Chaoqi Chen, Yushuang Wu, Qiyuan Dai, Hong-Yu Zhou, Mutian Xu, Sibei Yang, Xiaoguang Han, and Yizhou Yu. A survey on graph neural networks and graph transformers in computer vision: A task-oriented perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10297–10318, 2024. doi: 10.1109/TPAMI.2024.3445463.

[6] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechu Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023.

[7] Junzhe Chen, Tianshu Zhang, Shiyu Huang, Yuwei Niu, Linfeng Zhang, Lijie Wen, and Xuming Hu. Ict: Image-object cross-level trusted intervention for mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2411.15268*, 2024.

[8] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

[9] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024.

[10] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

[11] Qiyuan Dai and Sibei Yang. Curriculum point prompting for weakly-supervised referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13711–13722, June 2024.

[12] Qiyuan Dai and Sibei Yang. Free on the fly: Enhancing flexibility in test-time adaptation with online em, 2025. URL `https://arxiv.org/abs/2507.06973`.

[13] Qiyuan Dai, Hanzhuo Huang, Yu Wu, and Sibei Yang. Adaptive part learning for fine-grained generalized category discovery: A plug-and-play enhancement, 2025. URL `https://arxiv.org/abs/2507.06928`.

[14] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. URL `https://arxiv.org/abs/2305.06500`.

[15] Jinhao Duan, Fei Kong, Hao Cheng, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. Truthprint: Mitigating lvlm object hallucination via latent truthful-guided pre-intervention, 2025. URL `https://arxiv.org/abs/2503.10602`.

[16] Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding, 2024. URL `https://arxiv.org/abs/2403.14003`.

[17] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

[18] Anish Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *AAAI Conference on Artificial Intelligence*, 2023. URL https://api.semanticscholar.org/CorpusID:260887222.

[19] Xiang He, Sibei Yang, Guanbin Li, Haofeng Li, Huiyou Chang, and Yizhou Yu. Non-local context encoder: Robust biomedical image segmentation against adversarial attacks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):8417–8424, Jul. 2019. doi: 10.1609/aaai.v33i01.33018417. URL https://ojs.aaai.org/index.php/AAAI/article/view/4857.

[20] Hanzhuo Huang, Yufan Feng, Cheng Shi, Lan Xu, Jingyi Yu, and Sibei Yang. Free-bloom: Zero-shot text-to-video generator with llm director and ldm animator. *arXiv preprint arXiv:2309.14494*, 2023.

[21] Hanzhuo Huang, Yuan Liu, Ge Zheng, Jiepeng Wang, Zhiyang Dou, and Sibei Yang. MVTokenflow: High-quality 4d content generation using multiview token flow. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=zu7cBTPsDb.

[22] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[23] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

[24] Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model, 2023.

[25] Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. *arXiv preprint arXiv:2411.16724*, 2024.

[26] Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the image? a deep-dive into the vision of vision language models, 2024. URL https://arxiv.org/abs/2411.17491.

[27] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *arXiv preprint arXiv:2311.16922*, 2023. URL https://arxiv.org/abs/2311.16922.

[28] Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization, 2023. URL https://arxiv.org/abs/2210.15097.

[29] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=xozJw0kZXF.

[30] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

[31] Han Liang, Jiacheng Bao, Ruichi Zhang, Sihan Ren, Yuecheng Xu, Sibei Yang, Xin Chen, Jingyi Yu, and Lan Xu. Omg: Towards open-vocabulary motion generation via mixture of controllers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 482–493, June 2024.

[32] Liang Lin, Pengxiang Yan, Xiaoqian Xu, Sibei Yang, Kun Zeng, and Guanbin Li. Structured attention network for referring image segmentation. *IEEE Transactions on Multimedia*, 24:1922–1932, 2022. doi: 10.1109/TMM.2021.3074008.

[33] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll'a r, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. URL http://arxiv.org/abs/1405.0312.

[34] Zhenxiang Lin, Xidong Peng, Peishan Cong, Ge Zheng, Yujing Sun, Yuenan Hou, Xinge Zhu, Sibei Yang, and Yuexin Ma. Wildrefer: 3d object localization in large-scale dynamic scenes with multi-modal visual data and natural language. In *ECCV (46)*, pages 456–473, 2024. URL https://doi.org/10.1007/978-3-031-72952-2_26.

[35] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023.

[36] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

[37] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

[38] Jiazhen Liu, Yuhan Fu, Ruobing Xie, Runquan Xie, Xingwu Sun, Fengzong Lian, Zhanhui Kang, and Xirong Li. Phd: A chatgpt-prompted visual hallucination evaluation dataset. In *CVPR*, 2025.

[39] Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. *arXiv preprint arXiv:2407.21771*, 2024.

[40] Xuyang Liu, Bingbing Wen, and Sibei Yang. Ccq: Cross-class query network for partially labeled organ segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):1755–1763, Jun. 2023. doi: 10.1609/aaai.v37i2.25264. URL https://ojs.aaai.org/index.php/AAAI/article/view/25264.

[41] Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. Negative object presence evaluation (nope) to measure object hallucination in vision-language models. *ArXiv*, abs/2310.05338, 2023. URL https://api.semanticscholar.org/CorpusID:263829510.

[42] nostalgebraist. Interpreting gpt: The logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens, 2020.

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

[44] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[45] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018.

[46] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge, 2022. URL https://arxiv.org/abs/2206.01718.

[47] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. doi: 10.1109/ICCV.2017.74.

[48] Cheng Shi and Sibei Yang. *Spatial and Visual Perspective-Taking via View Rotation and Relation Reasoning for Embodied Reference Understanding*, page 201–218. Springer-Verlag, Berlin, Heidelberg, 2022. ISBN 978-3-031-20058-8. doi: 10.1007/978-3-031-20059-5_12. URL https://doi.org/10.1007/978-3-031-20059-5_12.

[49] Cheng Shi and Sibei Yang. Edadet: Open-vocabulary object detection using early dense alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023.

[50] Cheng Shi and Sibei Yang. Logoprompt:synthetic text images can be good visual prompts for vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023.

[51] Cheng Shi and Sibei Yang. The devil is in the object boundary: Towards annotation-free instance segmentation using foundation models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=4JbrdrHxYy.

[52] Cheng Shi, Yulin Zhang, Bin Yang, Jiajin Tang, Yuexin Ma, and Sibei Yang. Part2object: Hierarchical unsupervised 3d instance segmentation. *arXiv preprint arXiv:2407.10084*, 2024.

[53] Cheng Shi, Yuchen Zhu, and Sibei Yang. Plain-det: A plain multi-dataset object detector. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 210–226, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72652-1.

[54] Gabriela Ben Melech Stan, Raanan Yehezkel Rohekar, Yaniv Gurwicz, Matthew Lyle Olson, Anahita Bhiwandiwalla, Estelle Aflalo, Chenfei Wu, Nan Duan, Shao-Yen Tseng, and Vasudev Lal. Lvlm-intrepret: An interpretability tool for large vision-language models, 2024.

[55] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023.

[56] Jiajin Tang, Ge Zheng, Cheng Shi, and Sibei Yang. Contrastive grouping with transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23570–23580, June 2023.

[57] Jiajin Tang, Ge Zheng, and Sibei Yang. Temporal collection and distribution for referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15466–15476, 2023.

[58] Jiajin Tang, Ge Zheng, Jingyi Yu, and Sibei Yang. Cotdet: Affordance knowledge prompting for task driven object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3068–3078, 2023.

[59] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[60] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[61] Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15840–15853, 2024. URL `https://aclanthology.org/2024.findings-acl.937`.

[62] Sangmin Woo, Donguk Kim, Jaehyuk Jang, Yubin Choi, and Changick Kim. Don't miss the forest for the trees: Attentional vision calibration for large vision language models. *arXiv preprint arXiv:2405.17820*, 2024.

[63] Junfei Wu, Qiang Liu, Ding Wang, Jinghao Zhang, Shu Wu, Liang Wang, and Tieniu Tan. Logical closed loop: Uncovering object hallucinations in large vision-language models. *arXiv preprint arXiv:2402.11622*, 2024.

[64] Yu Wu, Yana Wei, Haozhe Wang, Yongfei Liu, Sibei Yang, and Xuming He. Grounded image text matching with mismatched relation reasoning, 2023.

[65] Yun Xing, Yiheng Li, Ivan Laptev, and Shijian Lu. Mitigating object hallucination via concentric causal attention. *arXiv preprint arXiv:2410.15926*, 2024.

[66] Sibei Yang, Guanbin Li, and Yizhou Yu. Dynamic graph attention for referring expression comprehension. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4643–4652, 2019. doi: 10.1109/ICCV.2019.00474.

[67] Sibei Yang, Guanbin Li, and Yizhou Yu. Cross-modal relationship inference for grounding referring expressions. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4140–4149, 2019. doi: 10.1109/CVPR.2019.00427.

[68] Sibei Yang, Guanbin Li, and Yizhou Yu. Propagating over phrase relations for one-stage visual grounding. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 589–605, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58529-7.

[69] Sibei Yang, Guanbin Li, and Yizhou Yu. Graph-structured referring expression reasoning in the wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9949–9958, 2020. doi: 10.1109/CVPR42600.2020.00997.

[70] Sibei Yang, Guanbin Li, and Yizhou Yu. Relationship-embedded representation learning for grounding referring expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8):2765–2779, 2021. doi: 10.1109/TPAMI.2020.2973983.

[71] Sibei Yang, Meng Xia, Guanbin Li, Hong-Yu Zhou, and Yizhou Yu. Bottom-up shift and reasoning for referring image segmentation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11261–11270, 2021. doi: 10.1109/CVPR46437.2021.01111.

[72] Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. Understanding and mitigating hallucination in large vision-language models via modular attribution and intervention. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=Bjq4W7P2Us`.

[73] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chaoya Jiang, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-owl: Modularization empowers large language models with multimodality, 2023.

[74] Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*, 2023.

[75] Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data, 2023.

[76] Zeping Yu and Sophia Ananiadou. Neuron-level knowledge attribution in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3267–3280, 2024.

[77] Zeping Yu and Sophia Ananiadou. Understanding multimodal llms: the mechanistic interpretability of llava in visual question answering. *arXiv preprint arXiv:2411.10950*, 2024.

[78] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. URL `https://arxiv.org/abs/2306.02858`.

[79] Jinrui Zhang, Teng Wang, Haigang Zhang, Ping Lu, and Feng Zheng. Reflective instruction tuning: Mitigating hallucinations in large vision-language models. *arXiv preprint arXiv:2407.11422*, 2024.

[80] Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*, 2023.

[81] Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *arXiv preprint arXiv:2310.16436*, 2023.

[82] Ge Zheng, Jiaye Qian, Jiajin Tang, and Sibei Yang. Why lvlms are more prone to hallucinations in longer responses: The role of context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4101–4113, October 2025.

[83] Guanyu Zhou, Yibo Yan, Xin Zou, Kun Wang, Aiwei Liu, and Xuming Hu. Mitigating modality prior-induced hallucinations in multimodal large language models via deciphering attention causality. In *The Thirteenth International Conference on Learning Representations*, 2025.

[84] Hong-Yu Zhou, Chixiang Lu, Sibei Yang, Xiaoguang Han, and Yizhou Yu. Preservational learning improves self-supervised medical image models by reconstructing diverse contexts. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3479–3489, 2021. doi: 10.1109/ICCV48922.2021.00348.

[85] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2023.

[86] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[87] Yuchen Zhu, Cheng Shi, Dingyou Wang, Jiajin Tang, Zhengxuan Wei, Yu Wu, Guanbin Li, and Sibei Yang. Rethinking query-based transformer for continual image segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4595–4606, 2025.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the main contributions and align well with the theoretical and experimental results presented in the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Appendix E for our limitation.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not contain any theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides detailed descriptions of the experimental setup, data preprocessing, model architecture, and training procedures, ensuring that the main results can be reliably reproduced.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Code will be released after acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides detailed experimental settings.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The performance improvements are significant, and the paper provides detailed analyses to support our conclusions.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The experimental results are not sensitive to compute resources.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite all the original paper that produced the code package or dataset.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Appendix

The appendix includes supplementary details, analyses, and experiments in support of the main text, structured as follows:

A Additional experimental and implementation details

A.1 Additional implementation details

For POPE [29] and MCQ-POPE, attention heads are identified using samples generated via an adversarial strategy, and subsequently applied across all three test splits: random, popular, and adversarial. Following the official LLaVA evaluation setup, prompts are appended with "Please answer this question with one word." All evaluations on POPE, MCQ-POPE, and MME [17] utilized nucleus sampling with a temperature of 1, and does not use beam search.

For CHAIR [45], we randomly sample 500 images from the COCO [33] dataset validation subset to identify attention heads. For both attention head identification and final evaluation, we pair each image with the prompt "*Please help me describe the image in detail.*", and responses are generated using greedy decoding with a maximum length of 512 tokens.

A.2 Prompt for MCQ-POPE

Which of the following {appears | does not appear} in the image?
A. {object1}
B. {object2}
C. {object3}
D. {object4}

B Additional analysis

In this section, we first supplement the main text with additional analyses based on LLaVA-v1.5-7B [36] in Section B.1 to further support the conclusion that LVLMs adaptively select different causal pathways for different questions. We then replicate the same series of analyses using Qwen-VL-Chat [3] in Section B.2 and Section B.3. Implementation details for Qwen-VL-Chat are provided in Section C.1.

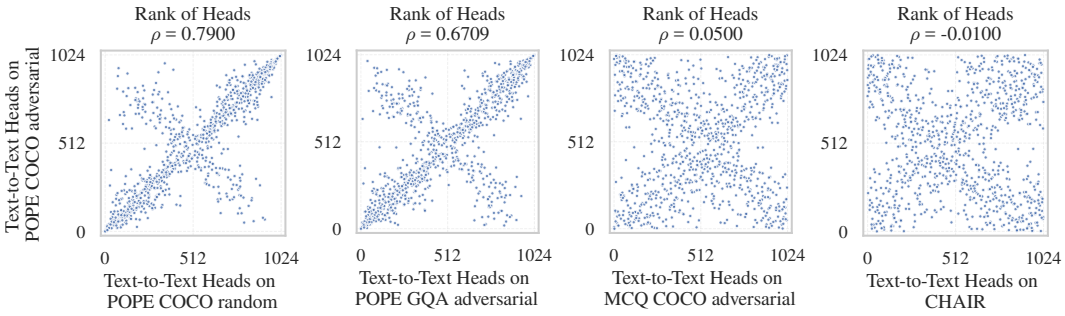


Figure 6: Visualization of the rank distributions of text-to-text heads extracted from different datasets on Qwen-VL-Chat [3]. ρ denotes the correlation coefficient; a higher value (*i.e.*, points concentrated along the diagonal) indicates greater similarity between the two sets of heads. Heads identified from datasets with similar question–answer alignment formats exhibit strong correlation and substantial overlap.

Method	**LLaVA on POPE (COCO)**						**Qwen on POPE (COCO)**					
	Random		*Popular*		*Advers.*		*Random*		*Popular*		*Advers.*	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
w/o mask	85.1	83.7	83.7	82.4	80.9	80.0	84.5	82.3	83.4	81.4	80.5	78.9
w/ mask	52.9	59.2	47.8	56.7	48.9	56.6	48.3	31.0	51.2	31.3	50.6	33.5
Δ%	↓32.2	↓24.5	↓35.9	↓25.7	↓32.0	↓23.4	↓35.0	↓51.3	↓32.2	↓50.1	↓29.9	↓45.4

Table 4: When we knock out the attention weights of ***input text*** tokens attending to image tokens on LLaVA-v1.5-7b [36] and Qwen-VL-Chat [3], POPE's performance worsens significantly.

B.1 Adaptive selection of causal path in LLaVA-v1.5-7B

As a supplement to the main experiments, we further knock out the image-to-text causal path from the image input to the text input while preserving the image-to-text path from image input to text output, as well as the text-to-text path from text input to output. Specifically, our experiments are conducted on LLaVA, where we mask the attention from text input tokens to image tokens.

As shown in Table 4 (left), this modification results in a substantial performance drop on POPE, with accuracy falling to **around 50%, comparable to random guessing**. This suggests that the model's prediction on POPE heavily depends on the causal path from the image input to the text input.

B.2 Functional interpretation of the identified heads in Qwen-VL-Chat

Text-to-text causal path is related to the alignment. We present the correlation maps of heads identified in Qwen-VL-Chat in Fig. 6. Consistent with the findings from LLaVA, we can observe that ***the correlation between text-to-text heads under the same alignment format is very high, but the correlation is significantly lower under different formats*** in Qwen-VL-Chat. Specifically, for Yes/No questions, the heads identified from POPE COCO adversarial show strong correlations with those from POPE COCO random and POPE GQA [22] adversarial, with correlation coefficients of 0.79 and 0.67, respectively. However, the correlation drops markedly when comparing POPE COCO adversarial with heads identified from the MCQ-POPE COCO (0.05), and also declines when compared to those from generation tasks (-0.01). These observations indicate that the T2T heads in Qwen-VL-Chat also exhibit a strong correlation with the alignment format.

Image-to-input-text causal path is different from image-to-input-text causal path. Consistent with the LLaVA-based analysis in the main text, Fig. 7 presents the correlation between image-to-input-text and image-to-output-text heads in Qwen-VL-Chat under the modified prompting strategy on CHAIR, along with the visualization of their focus on visual content. The results reveal that Qwen-VL-Chat exhibits a similar pattern: although both the image-to-input-text and image-to-output-text heads attend more strongly to the image than other heads, their attention patterns differ markedly. This suggests that the two pathways modulate the model's behavior in distinct.

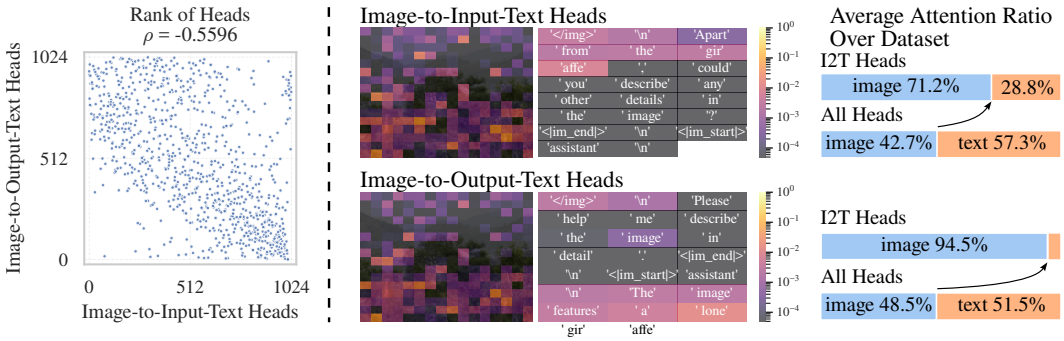


Figure 7: Left: Visualization of the rank distributions of image-to-input-text heads and image-to-output-text heads on Qwen-VL-Chat [3], showing that the two are largely uncorrelated. Right: Compared to the average of all heads, all image-to-text heads we identified, exhibit a stronger focus on visual content on Qwen-VL-Chat.

Method	**POPE (COCO)** *Random* Acc.	*Random* F1	*Popular* Acc.	*Popular* F1	*Advers.* Acc.	*Advers.* F1	**CHAIR** $C_S\downarrow$	**CHAIR** $C_I\downarrow$
w/o mask	84.5	82.3	83.4	81.4	80.5	78.9	43.4	13.5
w/ mask	83.4	81.1	83.1	80.9	80.9	78.9	78.4	48.0
Δ%	$\downarrow$1.1	$\downarrow$1.2	$\downarrow$0.3	$\downarrow$0.5	$\uparrow$0.4	$\downarrow$0.0	$\uparrow$35.0	$\uparrow$34.5

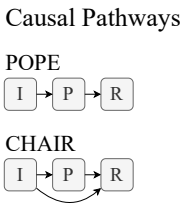


Figure 8: Left: When we mask the attention weights of ***output*** tokens attending to image tokens on Qwen-VL-Chat [3], POPE's performance worsens very little, whereas CHAIR's performance gets worse significantly. Right: This indicates that LVLMs utilize different pathways for different question formats.

B.3 Adaptive selection of causal path in Qwen-VL-Chat

We perform two counterfactual interventions based on Qwen-VL-Chat: (1) removing the image-to-output-text path by masking attention from text output tokens to image tokens, while preserving the image-to-input-text and text-to-text paths; and (2) removing the image-to-input-text path by masking attention from text input tokens to image tokens, while keeping the image-to-output-text and text-to-text paths. The effects of both interventions are illustrated in Fig. 8 and Table 4 (right).

(1) As shown in the left panel of Fig. 8, the first counterfactual intervention results in a negligible impact on POPE, with an average performance drop of less than 0.5 points, whereas it causes a substantial degradation on CHAIR. This suggests that the image-to-output path is not critical for POPE, but plays a crucial role in CHAIR. (2) Table 4 provides further evidence of the critical role played by the image-to-input-text causal path in POPE: masking this pathway reduces the model's accuracy to chance level, underscoring its essential contribution to successful prediction.

C Additional experimental results

C.1 Experimental setup for Qwen-VL-Chat

Following a similar methodology to LLaVA-v1.5-7B [36], we identify attention heads in Qwen-VL-Chat [3] using a generated dataset that is entirely disjoint from the test set. This dataset is employed to compute both the T2T and I2T Scores for head probing. However, due to substantial differences between the base models of Qwen-VL-Chat and LLaVA, we introduce two key adaptations in our implementation: (1) We modify the I2T Score to $S_{\text{I2T}}^{(l,n)} = S_{\text{I2T}}^{(l,n),+}$. This change is necessitated by the observation that attention heads in Qwen-VL-Chat generally assign non-negligible attention to most image patches. As a result, penalizing background attention would otherwise favor heads with uniformly low attention to visual inputs, which is not desirable. (2) We adopt $\xi = 20$ and $\zeta = 10$ for both discriminative and generative tasks.

Method	POPE (COCO)						MCQ-POPE (COCO)						CHAIR			
	Random		*Popular*		*Advers.*		*Random*		*Popular*		*Advers.*					
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	$C_S\downarrow$	$C_I\downarrow$	F1	Len
Vanilla	84.5	82.3	83.4	81.4	80.5	78.9	61.8	67.3	55.9	58.9	52.5	56.1	43.4	13.5	76.4	98.1
VCD [27]	85.5	83.4	84.3	82.3	82.0	80.2	66.6	71.9	60.9	63.6	55.3	58.3	46.0	12.9	76.9	100.8
ICD [61]	85.4	83.6	84.1	82.4	81.7	80.3	74.9	**77.7**	67.3	**68.8**	62.0	63.3	50.4	14.4	73.9	100.5
PAI [39]	85.8	84.1	84.9	83.3	**82.4**	80.9	57.9	66.1	52.0	56.6	48.6	54.6	17.2	7.9	70.1	36.8
AllPath	**86.6**	**85.3**	**85.6**	**84.3**	**82.4**	**81.5**	**75.4**	**77.7**	**68.8**	**68.8**	**62.6**	**63.6**	**34.2**	**9.0**	**77.2**	91.0

Table 5: Results of Qwen-VL-Chat [3] model on POPE [29], MCQ-POPE, and CHAIR [45] benchmark. The F1 score for MCQ-POPE represents Macro F1. Notably, PAI compromises generative capability, resulting in a substantial reduction in generation length. In contrast, our method significantly improves performance across all benchmarks while largely maintaining generative capacity.

Method	POPE (A-OKVQA)						POPE (GQA)					
	Random		*Popular*		*Advers.*		*Random*		*Popular*		*Advers.*	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Vanilla	86.3	85.8	82.7	82.7	75.9	77.5	85.1	84.5	79.6	80.0	76.6	77.6
VCD [27]	86.2	86.2	82.3	83.0	76.2	78.5	86.4	86.4	78.3	79.8	75.6	77.7
ICD [61]	86.0	85.1	83.2	82.6	77.3	77.8	84.5	83.3	80.6	80.0	77.5	77.6
PAI [39]	87.8	87.3	84.1	84.0	77.2	78.5	86.2	85.6	80.5	80.8	77.5	78.4
AD-HH [72]	86.2	85.7	82.7	82.7	76.0	77.6	85.0	84.5	79.6	80.0	76.5	77.5
AllPath	**88.1**	**87.9**	**84.2**	**84.5**	**77.6**	**79.4**	**87.5**	**87.2**	**81.9**	**82.5**	**78.2**	**79.5**

Table 6: Results of LLaVA-v1.5-7b [36] model on POPE [29] A-OKVQA [46] and GQA [22] split. Our method achieves consistent improvements over previous approaches across all metrics.

C.2 Results of Qwen-VL-Chat on POPE, MCQ-POPE and CHAIR

The results of Qwen-VL-Chat on POPE [29], MCQ-POPE, CHAIR [45] are shown in Table 5. Notably, our method outperforms each method on at least one benchmark by a significant margin. Specifically, compared to VCD [27], our method achieves an average improvement of **8.0%** and **5.4%** in terms of accuracy and F1 score on MCQ-POPE, respectively, and yields a gain of **11.8** and **3.9** on CHAIR_S and CHAIR_I. Relative to ICD [61], our approach demonstrates a significant margin of **16.2** on CHAIR_S and **5.4** on CHAIR_I. As for the PAI method [39], it is noteworthy that it degrades the generative capability of Qwen-VL-Chat, leading to a substantial reduction (from 98.1 to 36.8) in generation length on CHAIR. In contrast, our method not only achieves a significant average improvement of **16.1%** on accuracy and **10.9%** on F1 score on MCQ-POPE over PAI, but also substantially reduces the metrics on CHAIR_S and CHAIR_I while largely preserving the original generation length of Qwen-VL-Chat.

C.3 Results of Qwen2.5-VL on POPE

To verify that our AllPath remains effective on newer and larger models, we evaluate its performance on POPE on Qwen2.5-VL-7B/72B [4], as shown in Table 7. It is worth noting that since 72B version of Qwen2.5-VL has 5 times as many heads as LLaVA, we accordingly scale the number of heads used by a factor of five, *i.e.*, $\xi = 100$ and $\zeta = 50$. As can be seen from the table, our method also achieves consistent improvements on Qwen2.5.

D Qualitative results

D.1 Qualitative Results on POPE

Fig. 9 and 10 each illustrate two qualitative results from POPE [29], where the ground-truth answers to the given questions are Yes and No, respectively. They also show attention distributions of top- and bottom-ranked heads by T2T Score during answer generation, and I2T Scores at the object token, with average attention for comparison. From these visualizations, we observe that heads with the highest T2T Scores consistently attend to object tokens, differing from the average pattern. Moreover,

Method	Qwen2.5-VL-7B / POPE (COCO)						Qwen2.5-VL-72B / POPE (COCO)					
	Random		*Popular*		*Advers.*		*Random*		*Popular*		*Advers.*	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Vanilla	87.6	86.0	86.6	85.1	85.5	84.0	87.7	86.2	86.6	85.1	85.6	84.1
AllPath	**90.6**	**90.0**	**88.6**	**88.1**	**86.6**	**86.3**	**90.7**	**90.2**	**88.5**	**88.1**	**85.7**	**85.7**

Table 7: Results of Qwen2.5-VL [4] model on POPE [29] COCO [33] split. Our method achieves consistent improvements across all metrics.

when the ground-truth answer is Yes, *i.e.*, the queried object indeed exists in the image, the heads with the highest I2T Scores focus on the corresponding visual regions, again showing a distinct difference from the mean attention. These qualitative findings further demonstrate that our two core metrics, the I2T Score and T2T Score can effectively identify the critical attention heads along the two pathways. This provides strong evidence that our method, AllPath, mitigates hallucinations by enhancing both pathways.

D.2 Qualitative Results on CHAIR

To further illustrate the behavior of our intervention, we present qualitative examples highlighting both successful (Fig. 11) and failure (Fig. 12) cases of our AllPath on CHAIR [45]. We emphasize that our method is particularly effective when the LVLM demonstrates correct reasoning even along just one of the two pathways (e.g., image-to-text or text-to-text). In such scenarios, our intervention adaptively enhances the pathway to make it more reliable, thereby reducing hallucination. However, our method is less effective when both pathways in the LVLM exhibit substantial errors, making it difficult for the intervention to fully correct the output.

- **Success Case.** The baseline model generates a hallucination "potted plants", whereas our AllPath correctly describes the scene without hallucination. Visualization shows that when baseline is generating "potted plants", the heads with top I2T Scores correctly attend to the flower in the image, while the heads with bottom T2T scores distribute attention across general objects such as "dog", "bench", and "fence". Our method adaptively amplified both the image-to-text and text-to-text pathways, reinforcing useful information and effectively suppressing hallucination.
- **Failure Case.** The baseline response hallucinates two "bowls", and our AllPath still produces a hallucinated "cup". Visualization shows that when baseline is generating "bowls", the heads with top I2T Score fail to focus on the relevant area in the image, while the heads with bottom T2T Score remain strongly influenced by textual priors, particularly the token "blender". Since neither pathway is reliable, our intervention could not mitigate this hallucination.

E Limitations

This work focuses on understanding how hallucinations in LVLMs relate to their internal causal pathways. However, we do not investigate what causal pathways are most effective for different types of questions, nor do we explore training strategies that could systematically improve the use of such pathways. Additionally, our analysis is constrained by the scope of existing hallucination benchmarks, which may not capture a broader range of unrecognized or subtle hallucinations. These remain important open questions for future research.

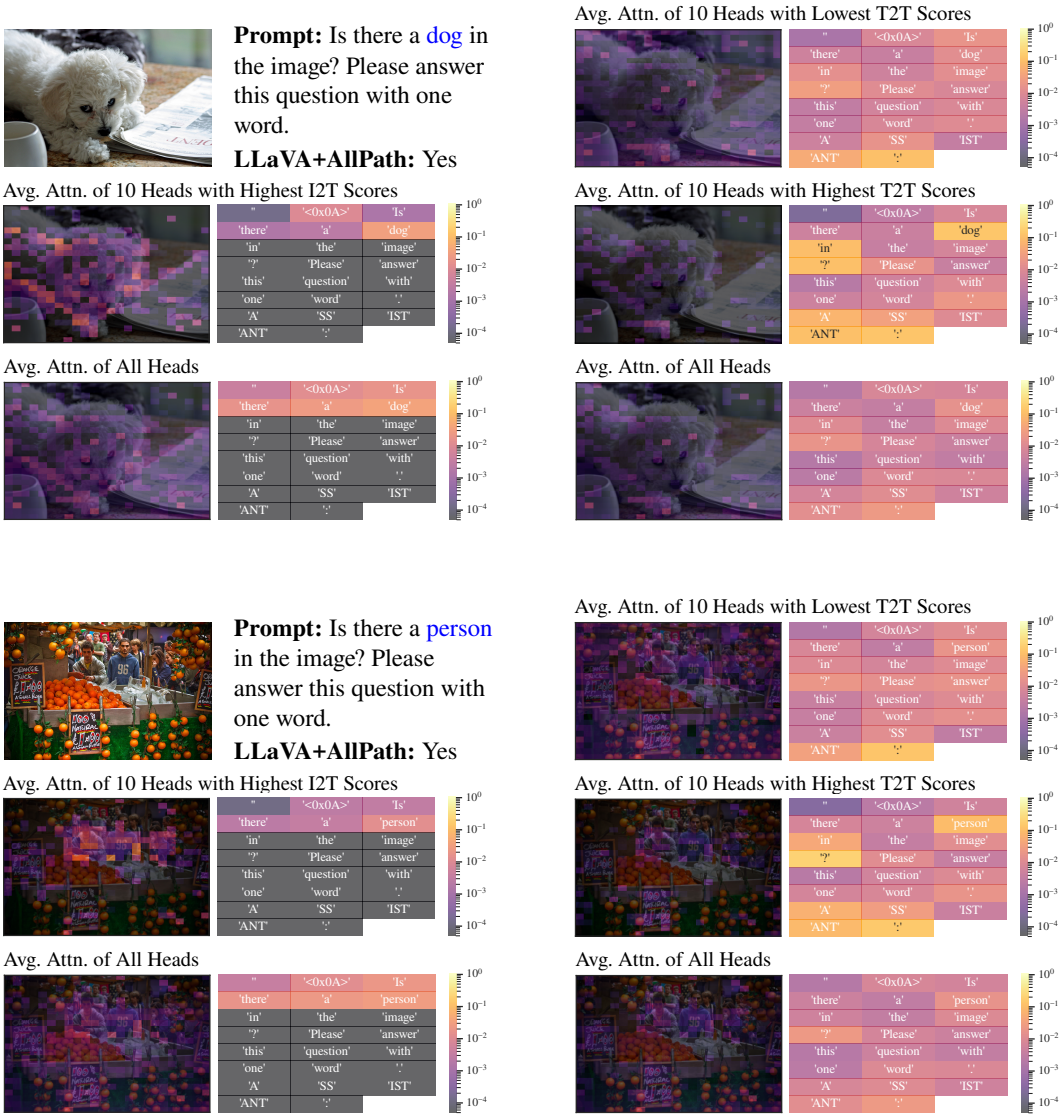


Figure 9: Qualitative visualization of attention weights of I2T and T2T heads on the POPE dataset [29]. Each case contains three rows and two columns. Left: The top row shows the input image, the corresponding question, and the model's predicted answer, where the object token in the question is highlighted in blue. The second and third rows visualize the attention from the object token to previous tokens: the second row shows the average attention of the 10 heads with the highest I2T scores, and the third row shows the average attention of all heads. Right: The attention weights when generating the final answer token. From top to bottom: average attention of the 10 heads with the lowest T2T scores, average attention of the 10 heads with the highest T2T scores, and average attention of all heads.

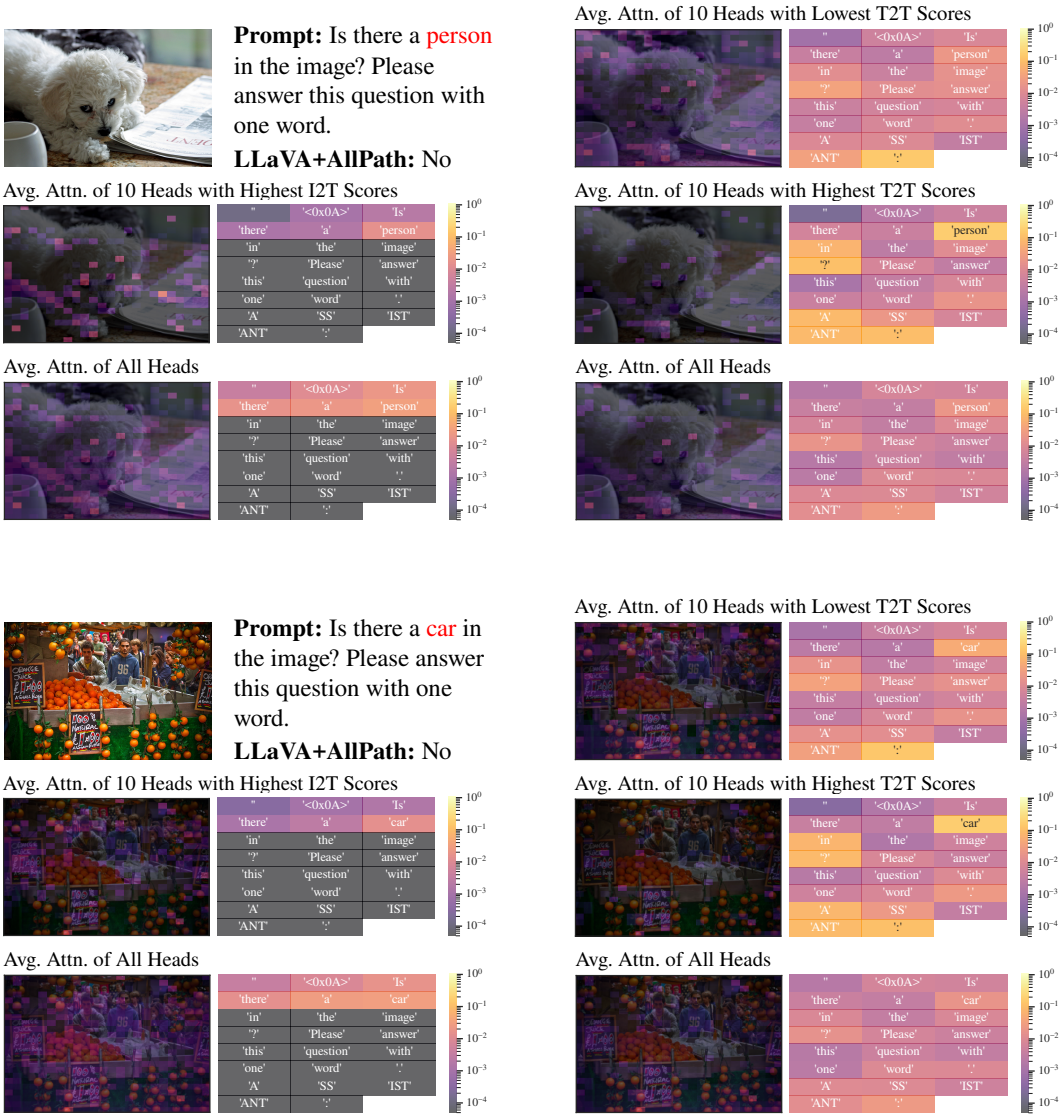
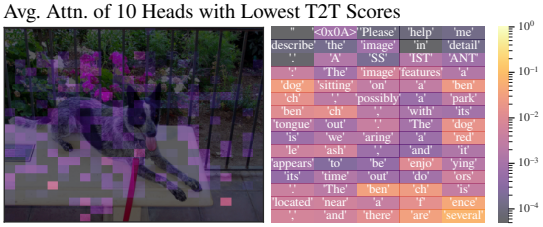
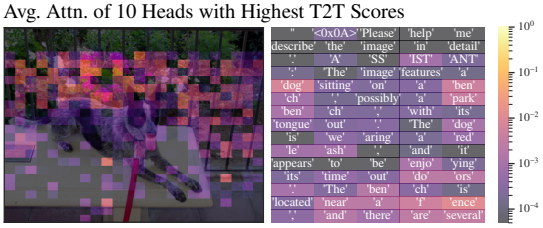


Figure 10: Qualitative visualization of attention weights of I2T and T2T heads on the POPE dataset [29]. Each case contains three rows and two columns. Left: The top row shows the input image, the corresponding question, and the model's predicted answer, where the object token in the question is highlighted in red. The second and third rows visualize the attention from the object token to previous tokens: the second row shows the average attention of the 10 heads with the highest I2T scores, and the third row shows the average attention of all heads. Right: The attention weights when generating the final answer token. From top to bottom: average attention of the 10 heads with the lowest T2T scores, average attention of the 10 heads with the highest T2T scores, and average attention of all heads.

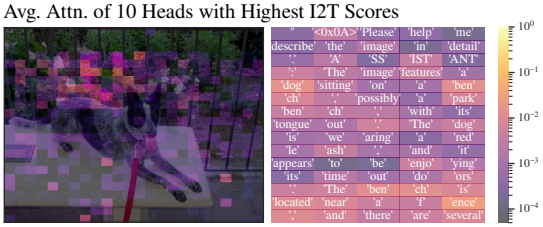
Avg. Attn. of 10 Heads with Lowest T2T Scores



Avg. Attn. of 10 Heads with Highest T2T Scores



Avg. Attn. of 10 Heads with Highest I2T Scores



Prompt:

Please help me describe the image in detail.

LLaVA (Baseline):

The image features a dog sitting on a bench, possibly a park bench, with its tongue out. The dog is wearing a red leash, and it appears to be enjoying its time outdoors. The bench is located near a fence, and there are several potted plants surrounding the area.

In addition to the main dog, there are two other dogs in the scene, one located near the left edge of the image and the other towards the right side. The presence of multiple dogs suggests that this might be a popular spot for pet owners to bring their pets for a walk or a rest.

LLaVA+AllPath:

The image features a dog sitting on a bench, possibly a park bench, with its tongue out. The dog is wearing a red leash, and it appears to be enjoying its time outdoors. The bench is located near a fence, and there are several potted plants surrounding the area.

In addition to the main dog, there are two other dogs in the scene, one located near the left edge of the image and the other towards the right side. The presence of multiple dogs suggests that this might be a popular spot for pet owners to bring their pets for a walk or a rest.

Figure 11: Qualitative success case of our intervention on CHAIR [45]. Non-hallucinated objects in model's response are highlighted in blue, while hallucinated objects are highlighted in red. The upper part of this figure presents the baseline output, with right column visualizes the model's average attention of different set of heads when generating the hallucinated object "potted plants". From top to bottom: average attention of the 10 heads with the lowest T2T scores, average attention of the 10 heads with the highest T2T scores, and average attention of the 10 heads with the highest I2T scores.

Prompt:

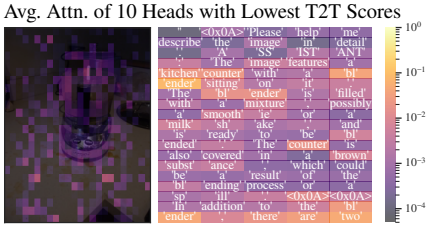
Please help me describe the image in detail.

LLaVA (Baseline):

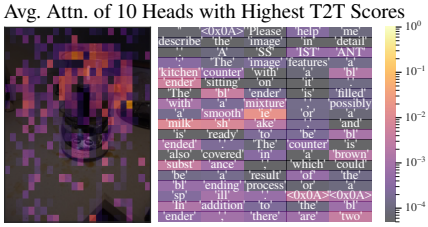
The image features a kitchen counter with a blender sitting on it. The blender is filled with a mixture, possibly a smoothie or a milkshake, and is ready to be blended. The counter is also covered in a brown substance, which could be a result of the blending process or a spill.

In addition to the blender, there are two bowls on the counter. One bowl is located on the left side of the blender, while the other is positioned on the right side. The kitchen counter appears to be a bit messy, with the blender and the brown substance taking up a significant portion of the space.

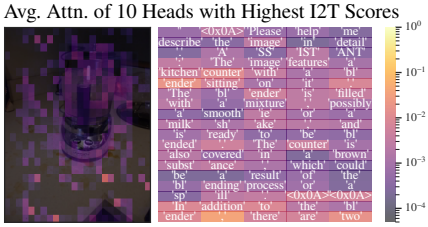
Avg. Attn. of 10 Heads with Lowest T2T Scores



Avg. Attn. of 10 Heads with Highest T2T Scores



Avg. Attn. of 10 Heads with Highest I2T Scores



LLaVA+AllPath:

The image features a kitchen counter with a blender sitting on it. The blender is filled with a mixture, possibly a smoothie or a smoothie-like substance. The blender is turned on, and the blender cup is filled with the mixture. The counter is also covered in a bit of mess, with a few drops of liquid spread across the counter.

Figure 12: Qualitative failure case of our intervention on CHAIR [45]. Non-hallucinated objects in model's response are highlighted in blue, while hallucinated objects are highlighted in red. The upper part of this figure presents the baseline output, with right column visualizes the model's average attention of different set of heads when generating the hallucinated object "bowls". From top to bottom: average attention of the 10 heads with the lowest T2T scores, average attention of the 10 heads with the highest T2T scores, and average attention of the 10 heads with the highest I2T scores.