

Mechanisms of Prompt-Induced Hallucination in Vision–Language Models

William Rudman^{*1}, Michal Golovanevsky^{*2}, Dana Arad³,
Yonatan Belinkov^{3,5}, Ritambhara Singh², Carsten Eickhoff⁴, Kyle Mahowald¹

¹The University of Texas at Austin, ²Brown University, ³Technion,

⁴University of Tübingen, ⁵Harvard University

michal_golovanevsky@brown.edu & william.rudman@utexas.edu

Abstract

Large vision–language models (VLMs) are highly capable, yet often hallucinate by favoring textual prompts over visual evidence. We study this failure mode in a controlled object-counting setting, where the prompt overstates the number of objects in the image (e.g., asking a model to describe *four* waterlilies when only *three* are present). At low object counts, models often correct the overestimation, but as the number of objects increases, they increasingly conform to the prompt regardless of the discrepancy. Through mechanistic analysis of three VLMs, we identify a small set of attention heads whose ablation substantially reduces prompt-induced hallucinations (PIH) by at least 40% without additional training. Across models, PIH-heads mediate prompt copying in model-specific ways. We characterize these differences and show that PIH ablation increases correction toward visual evidence. Our findings offer insights into the internal mechanisms driving prompt-induced hallucinations, revealing model-specific differences in how these behaviors are implemented. *

1 Introduction

Vision–language models (VLMs) often follow textual prompts even when they contradict visual evidence (Deng et al., 2025; Zhang et al., 2025; Hua et al., 2025). These hallucinations are a growing concern as VLMs are deployed in real-world settings with noisy or inaccurate user inputs. To better understand this behavior, we study *prompt-induced hallucinations* (PIH) in a controlled object-counting setting, where prompts systematically overestimate the number of objects in an image.

Figure 1 illustrates the phenomenon where, although the model correctly counts three waterlilies when asked directly, it hallucinates additional,

*Equal contribution. Order determined by coin flip.

*Code available at: https://github.com/michalg04/prompt-induced_hallucinations.git

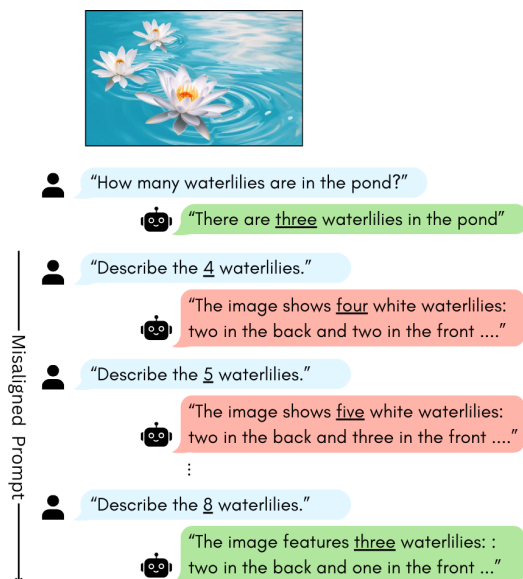


Figure 1: **Example of prompt-induced hallucination (PIH) in numerical reasoning.** The model correctly answers the baseline counting question. However, it conforms to the prompt and hallucinates additional waterlilies when asked to describe a mismatched number of objects. As the mismatch between the prompt and visual evidence grows, the model increasingly favors the image and recovers the correct count.

nonexistent flowers when asked to describe a mismatched count (e.g., “Describe the four waterlilies”). This undesirable behavior is unlikely to be driven by limitations in numerical reasoning (Zhang et al., 2023a; Rahmanzadehgervi et al., 2024; Rudman et al., 2025), since the model correctly solves the underlying counting task. Instead, this is consistent with a known tendency of VLMs to rely on textual prompts over conflicting visual evidence (Deng et al., 2025).

To characterize PIH, we study **when** such errors occur and **how** they can be reduced via targeted analysis of attention-heads. Across models, PIH outputs follow a consistent and structured pattern. When the ground-truth number of objects

exceeds four, VLMs are more likely to conform to the prompt and hallucinate missing objects. In contrast, at lower object counts, models are more likely to copy the prompt under small discrepancies but transition to visually grounded responses as the mismatch increases (Figure 1).

Through mechanistic analysis, we identify attention heads associated with PIH. Ablating these heads reduces hallucinations and restores visually grounded counts by up to 54%. Importantly, this intervention does not require additional data or training and alters only the object count, leaving the described objects unchanged. To test generalization beyond counting, we apply the same PIH head ablations to a color identification task (Golovanevsky et al., 2025a) in which the prompt specifies an incorrect color. We find that these heads similarly mediate PIH in the color domain, mitigating prompt-color copying by up to 94.25%.

We further investigate the functional role of PIH attention heads and find that their ablation consistently reduces prompt-based copying, increases correction toward ground-truth responses, and shifts attention toward the image. While these effects are shared across all tested models, the mechanisms by which copying is suppressed differ. We make this distinction by examining how models format their responses, how attention shifts between text and image inputs, and how output probabilities change after ablation. Across models, these analyses show that PIH ablation consistently reduces direct copying from the prompt. While the way this reduction appears varies by model, it is always accompanied by increased reliance on visual information and improved correctness. LLaVA-OneVision shows the strongest reduction in prompt-following after PIH-heads ablation, with clear suppression of prompt-consistent formatting, together with the largest increase in image attention and improved general counting performance.

Our contributions are threefold:

1. We introduce and analyze PIH, a failure mode in which VLMs prioritize prompt information over conflicting visual evidence. (Section 3).
2. We identify a small subset of attention heads whose ablation substantially reduces PIH across models and tasks, without additional training or performance degradation on aligned prompts (Section 4 & Section 5).
3. We characterize the function of PIH-heads and show their ablation suppresses prompt

copying while increasing correction toward visual evidence. (Section 6).

2 Related Work

2.1 Prompt-Induced Hallucinations in LLMs

A substantial body of work has investigated hallucinations in large language models (LLMs) (Ji et al., 2023; Tonmoy et al., 2024; Huang et al., 2025). Prior work has shown that hallucinations can be reliably induced through out-of-distribution prompting (Yao et al., 2023), long context windows (Li et al., 2024b; Xu et al., 2024a), or the use of explicit personas (Joshi et al., 2024; Simhi et al., 2025). However, these settings often reflect adversarial or non-realistic usage patterns, rather than the models’ behavior under typical user interactions. Previous studies have demonstrated that LLMs may produce incorrect outputs when prompts contradict internal knowledge (Xie et al., 2023; Xu et al., 2024b) or contain errors subsequently repeated or amplified by the model (Zhang et al., 2023b; Simhi et al., 2025). This behavior has been linked to cognitive-like biases, including anchoring bias, where models rely disproportionately on information introduced in the prompt (Jones and Steinhardt, 2022; Echterhoff et al., 2024; Malberg et al., 2025), and sycophancy, where models conform to user assumptions despite conflicting evidence (Sharma et al., 2024).

2.2 Modality Conflicts in VLMs

Vision–language models (VLMs) integrate textual and visual information, introducing additional complexity. Prior work shows that VLMs achieve higher accuracy on textual tasks compared to visual equivalents (Frank et al., 2021; Fu et al., 2024; van Sprang et al., 2025). This gap is attributed to distinct internal mechanisms for visual and textual processing (Nikankin et al., 2025; Golovanevsky et al., 2025b) and to modality-specific parametric knowledge (Zhu et al., 2024). VLMs also struggle to disentangle different signals, both across their input modalities (Hua et al., 2025) and between their input and parametric knowledge (Tang et al., 2023; Golovanevsky et al., 2025a; Ortu et al., 2025).

Recent work investigated how VLMs perform under inconsistent or conflicting textual and visual inputs (Yan et al., 2025; Zhao et al., 2025). Given conflicting information, current VLMs tend to disproportionately prefer textual information (Salin et al., 2022; Deng et al., 2025), which can manifest

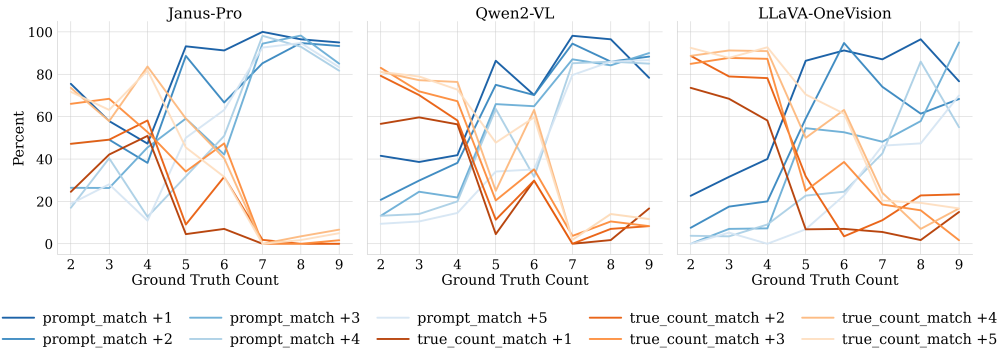


Figure 2: **PIH rates across different numbers of ground-truth objects.** Blue lines show the percentage of outputs matching the prompted number for different offsets, k . Orange lines show matches to the true count. As the object count increases, models stop correcting the prompt.

in sycophantic behavior (Pi et al., 2025). Further exploring modality conflicts, Zhang et al. investigated modality preference through the lens of unimodal uncertainty. In contrast, our work studies modality conflicts stemming from minor, realistic inconsistencies and examines a more complex task that more closely reflects real-world usage.

3 Prompt-Induced Hallucinations in Object Counting

In this section, we present Prompt-Induced Hallucinations (PIH) and study when they manifest in the context of object counting. Object counting provides a clean setting for isolating prompt-image conflicts: ground-truth answers are verifiable, numerical deviations can be precisely controlled, and model errors can be explicitly measured. While a language model could legitimately answer the question "Describe 4 waterlilies" as part of an instruction-following task without hallucinating, for VLMs these responses constitute hallucinations: (1) we explicitly ground prompts with the phrase "in the image," and (2) the models describe additional objects not present in the image (see examples in Table 11).

3.1 Experimental Setup

We investigate three VLMs covering a range of state-of-the-art model families: LLaVA-OneVision-7B (Li et al., 2024a), Qwen2-VL-7B (Wang et al., 2024), and Janus-Pro-7B (Chen et al., 2025). We study counting using *CountBench* (Paiss et al., 2023), a benchmark designed to measure object-count understanding in VLMs. For each image with ground-truth object count N , we first present a *baseline prompt*, $P_B =$ "How many [objects]

are there in the image?", retaining only samples where the model correctly predicts N . We then use *misaligned prompts* asking the model to "Describe the $N + k$ [objects] in the image," where $k \in \{1, 2, \dots, 5\}$ denotes the *discrepancy distance* between the prompt and the visual input. PIH occur when the model describes $N + k$ objects without correcting the count, despite contradictory visual evidence. See Appendix A for additional details.

3.2 PIH at Different Object Counts

Figure 2 shows the proportion of outputs that either match the prompted count ($N + k$, blue curves) or the true count (N , orange curves) across a range of object quantities and prompt offsets (k). The x -axis denotes the true number of objects (N), and the y -axis shows the fraction of responses aligned with either the prompt or ground truth. We observe a consistent pattern across all three models. For images containing small numbers of objects (2-4 objects), models typically resist misleading prompts and report the true count, demonstrated by the orange curves being mostly higher than the blue curves. However, models still produce PIH in approximately 20-40% of cases, which primarily occur when the discrepancy is small ($N + 1$).

Beyond four objects, this behavior changes. Models increasingly conform to the prompt regardless of the discrepancy size, with prompt-matching responses approaching 80-90% even when the prompt substantially overestimates the true count. Accordingly, ground truth responses drop to near zero. In these cases, the discrepancy distance ceases to play a critical role, as shown by the convergence of the blue prompt-aligned lines. This effect persists for extreme offsets (with $k \in \{10, 20, 50\}$), with models readily describing,

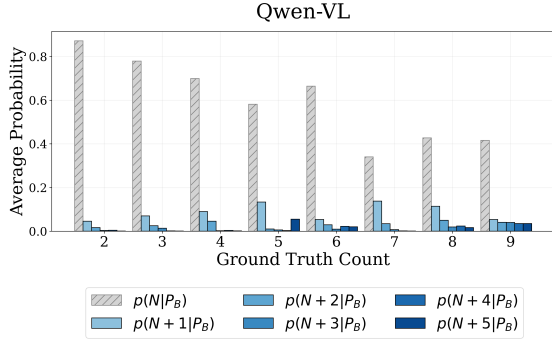


Figure 3: **Base probabilities on the object counting task.** We plot the probability assigned to the true count N and neighboring counts $N+k$ ($k \in \{1, \dots, 5\}$) given the prompt “How many [objects] are in the image?”. As object counts increase, confidence in N decreases and probability mass spreads to neighboring counts.

for example, fifty-nine cats in an image containing only nine (more details in Appendix B).

3.3 PIH Increase as Confidence Decreases

We hypothesize that models are more confident when counting fewer objects and therefore more resistant to misaligned prompts, with failures concentrated at minimal discrepancies ($k = 1$). We test this by collecting the probabilities for the true count N and neighboring counts $N+k$ under the baseline prompt (P_B). Figure 3 shows results for Qwen-VL, with similar trends for Janus-Pro and LLaVA-OneVision shown in Appendix C. We filter out all samples for which P_B is answered incorrectly, ensuring that the reported values reflect base confidence rather than raw accuracy. Indeed, for $N \leq 4$, the probability $p(N | P_B)$ is high while $p(N+k | P_B)$ is low, indicating strong confidence in the predicted count. At $N \geq 5$, base confidence drops and $p(N+k | P_B)$ increases, matching our PIH findings. Computing the Pearson correlation, we observe a moderately high ρ between the base probability and confirmation to the prompt (computed per-sample), with $\rho = 0.37, 0.46$ for Qwen-VL, and Janus-Pro, respectively, as well as a small positive correlation of 0.1 for LLaVA-OneVision. Overall, these results demonstrate that base probabilities are correlated with model behavior under prompt-image mismatches, suggesting that PIH is more prevalent when visual confidence is low.

4 Locating PIH Mechanisms

To better understand and reduce PIH, we aim to locate model components responsible for PIH. Prior

work has shown that attention heads often serve specific functional roles, such as information copying (Olsson et al., 2022; Feucht et al., 2025) and context tracking (Prakash et al., 2025), whereas MLP layers are more closely associated with storing parametric knowledge (Meng et al., 2023; Arad et al., 2024). Therefore, we focus our analysis on identifying attention heads associated with PIH.

4.1 Method: Attention Head Knockouts

To analyze the causal role of individual attention heads in PIH, we perform attention head knockouts using mean ablation (Wang et al. (2022); Mueller et al. (2025), see Appendix D for technical details).

Intuitively, this intervention removes token-specific information carried by the head while preserving its overall activation magnitude. We apply this procedure head-by-head and quantify knockout success as the fraction of samples that switch from the prompted count to the true count under the misaligned prompt. We rank heads by success and select the top- m heads per model. We then perform a second-stage grouped knockout in which all top- m heads are ablated simultaneously for $m \in \{1, 3, 5, 10\}$, selecting $m = 3$ for Qwen-VL and $m = 10$ for LLaVA-OneVision and Janus-Pro as a per-model hyperparameter. We refer to the final set of per-model attention heads as **PIH attention heads**. See Appendix E for a full list of top-10 heads per model and their success scores.

4.2 Attention Knockouts for PIH Mitigation

Table 1 shows models’ outputs given the baseline and misaligned prompts, before and after knockouts on the model-specific sets of PIH-heads. To verify that our results are specific to PIH heads rather than a general effect of ablation, we compare PIH heads to an equal number of randomly selected heads from the same layers in which the PIH heads reside. First, our results show that ablating PIH-heads does not disrupt standard counting behavior, as demonstrated by the exact-match accuracies on the baseline prompt. Notably, LLaVA-OneVision exhibits an improvement after ablation (+4.35 points), whereas Qwen-VL and Janus-Pro show minor changes of approximately 1 point. Additionally, we verify that the object mentioned in the response remains unchanged and correct. This confirms that ablating PIH-heads does not impact general counting capabilities.

Under the misaligned prompts, ablating PIH-heads substantially reduces prompt-following be-

	LLaVA-OV	Qwen-VL	Janus-Pro
Baseline Prompt: Exact Match (\uparrow)			
Before	76.89	78.49	80.32
After (Random)	77.30	78.70	80.00
After (PIH)	81.24	79.29	79.41
Misaligned Prompt: Prompt Match (\downarrow)			
Before	42.58	56.51	64.10
After (Random)	37.80	54.60	58.30
After (PIH)	1.42	3.22	10.19
Misaligned Prompt: True Count Match (\uparrow)			
Before	45.68	37.70	30.54
After (Random)	48.50	39.00	34.00
After (PIH)	77.80	70.66	70.90

Table 1: **Response accuracies on baseline and misaligned prompts before and after ablations.** The baseline prompt is “How many [objects] are there in the image?”. Misaligned prompts take the form “Describe the $N + k$ [objects] in the image” ($k \in \{1, 2, \dots, 5\}$) with results averaged across all k . As a random baseline, we sample an equal number of heads from the same layers as the PIH heads. Lower prompt match indicates reduced PIH, while a higher true-count match reflects stronger correction toward visual evidence. Prompt and true-count matches do not sum to 100%, as some responses contain no numerical value.

behavior, while random ablations produce only marginal reductions. Prompt match rates (previously corresponding to the blue curves in Figure 2) drop from 42-64% before intervention to below 11% for all models after ablation. Models also show a complementary increase in true count match rates, reaching 70-78% after intervention. This shift indicates that the ablated heads are responsible for propagating the incorrect numerical information specified by the prompt. Qualitative analysis of model generations before and after PIH-head ablation shows not only a reduction in incorrect numerical repetition but also the disappearance of spurious objects that were previously hallucinated and described in detail (examples are shown in Appendix F).

Our findings show that PIH is causally mediated by a small set of attention heads that propagate tokens from the prompt. Removing these heads shifts behavior from prompt reliance toward image reliance without harming general counting abilities, indicating that observed hallucinations are prompt-induced rather than due to counting failures.

4.3 Do PIH-Head Ablations Affect General Capabilities?

While Table 1 provides initial evidence that PIH-head ablation does not degrade performance on

Model	CalTech101		MM-Vet		POPE	
	Before	After	Before	After	Before	After
LLaVA-OV	100.0	100.0	50.5	50.7	89.3	89.3
Qwen-VL	100.0	96.0	43.9	42.8	85.0	86.0
Janus-Pro	99.0	99.0	46.5	44.8	86.3	87.3

Table 2: Accuracy on CalTech101, MM-Vet, and POPE before and after PIH-head ablation.

CountBench, we further evaluate whether our intervention impacts broader capabilities.

First, we test whether ablating PIH heads interferes with instruction following or general copying behavior. We construct a controlled task that does not introduce mismatches between the image and the prompt, in which copying is desired. Specifically, we adapt CalTech101 (Li et al., 2022) and use the prompt: “Repeat the following sentence if the image contains $a(n)$ [OBJECT]. There is $a(n)$ [OBJECT] in the image.”. We measure whether the model correctly copies the second sentence conditioned on the image label.

We further evaluate models on MM-Vet (Yu et al., 2023) and POPE (Li et al., 2023) to assess broader multimodal capabilities. MM-Vet covers diverse skills such as recognition, OCR, spatial reasoning, and knowledge-based reasoning, while POPE primarily consists of existence-based questions (e.g., “Is there a bottle in the image?”). Since these prompts are image-grounded and do not introduce explicit conflicts between the prompt and visual evidence, we do not expect PIH-head ablation to significantly affect performance.

Performance remains broadly stable across all three benchmarks, with only minor fluctuations. This indicates that PIH-head ablation does not introduce negative side effects on general multimodal reasoning or standard image-grounded question answering.

4.4 Early-Layer LM Heads Mediate PIH

We observe that PIH-heads exhibit a consistent pattern in their layer-wise localization. Across all three models, top-10 PIH-heads (detailed in Table 9) emerge in early and middle layers, suggesting that prompt-induced numerical information is introduced early and subsequently propagated forward. In both Qwen-VL and LLaVA-OneVision, the top-ranked PIH-heads are L0H3 and L0H6, while in Janus-Pro the most influential head is likewise located in the first layer (L0H20). This pattern extends beyond the top ranks: among

Model	PIH Heads
LLaVA-OV	<u>L0H3</u> , <u>L0H6</u> , L0H26, <u>L11H15</u> , L0H23, <u>L0H24</u> , L0H15, L14H9, <u>L0H11</u> , L17H22
Qwen-VL	<u>L0H3</u> , <u>L0H6</u> , <u>L11H15</u> , <u>L0H11</u> , L15H8, L14H13, <u>L0H10</u> , L7H3, <u>L0H24</u> , L8H13
Janus-Pro	<u>L0H20</u> , <u>L1H7</u> , L14H27, L12H9, L11H18, L13H2, <u>L0H24</u> , L14H28, L8H11, L8H3

Table 3: Top-10 attention heads per model. Heads shared by Qwen-VL and LLaVA-OneVision are in underline. Early-layer heads (L0-1) are **highlighted**.

the top-10 PIH-heads, 5/10 occur in the 0th layer for Qwen-VL, 7/10 for LLaVA-OneVision, and 3/10 fall within the first two layers (layers 0–1) for Janus-Pro. The early-layer concentration of PIH-heads, highlighted in Table 9, suggests that numerical information from the prompt is injected before extensive multimodal integration or higher-level reasoning takes place, enabling it to shape downstream computation even when it conflicts with visual evidence.

This early-layer localization raises a related question about the source of these mechanisms: namely, whether PIH arises from vision–language interaction or from the language model itself. One way to interpret this question is through the overlap of PIH-heads between Qwen-VL and LLaVA-OneVision, which share the same underlying Qwen2 language model but use different vision backbones and image–language fusion architectures.

The top-ranked PIH-heads exhibit substantial overlap across the two models, underlined in Table 9: the top-1 and top-2 heads overlap perfectly, and at $m = 10$, half of the identified heads are shared. This level of overlap strongly suggests that PIH is primarily mediated by language-model attention heads, rather than arising from vision-specific components or cross-modal alignment layers. This is consistent with prior work showing that fine-tuning tends to amplify or repurpose existing mechanisms rather than introduce entirely new mechanisms (e.g., Prakash et al., 2024). Further, this supports the view that PIH reflects language-internal information routing that persists across multimodal models.

5 Generalization Beyond Counting

Given that PIH appears to be mediated by shared language-model attention heads that persist across multimodal architectures, we next ask whether these same heads generalize beyond counting to mitigate prompt-induced hallucinations in other tasks. Specifically, we use a color prediction task based on the Visual CounterFact dataset (Golovanevsky et al., 2025a), as color naturally allows gradual discrepancies. We use “What color is the [object] in the image?” as the baseline prompt, and “Describe the $C + k$ [object]” as the misaligned prompt, where C denotes the ground-truth color of the object and $k \in \{1, 2, 3\}$ denotes the perceptual “difference” from the C on the color wheel. Additional details on the color task are available in Appendix G. To test generalization, we use the same set of PIH-heads per model (Table 1) on the color task. Table 4 show that ablating the same PIH heads yields large improvements under misaligned color prompts across all models, reducing prompt-induced hallucinations by between 40% and 95%. This demonstrates that PIH heads identified through counting generalize beyond numerical reasoning and are not task-specific.

6 Analyzing the Function of PIH-heads

In this section, we analyze PIH-heads to better understand their functionality. We hypothesize that ablating PIH-heads induces one of two mechanisms: PIH-heads could reduce hallucinations by inhibiting copying behavior or by increasing attention flow to the image. Namely, PIH-heads could enable the model to propagate incorrect information from the prompt without referencing the image. Alternatively, PIH-heads may influence hallucinations by modulating the model’s reliance on the image, so that removing these heads increases reliance on visual evidence.

To determine the functionality of PIH heads, we first investigate their attention patterns over the input prompts and find that they do not reveal a consistent pattern connecting the generated answer to either copying the prompt or relying on the image (Examples shown in Appendix E.1). This is in line with several studies showing that influential attention heads may not exhibit human-interpretable patterns (Jain and Wallace, 2019; Serrano and Smith, 2019; Brunner et al., 2019; Grimsley et al., 2020). Therefore, we study the function of PIH-heads by examining the generated texts before and af-

Response	LLaVA-OV		Janus-Pro		Qwen-VL	
	Before	After	Before	After	Before	After
NC (correct)	0.96	46.01	5.17	20.94	8.93	30.24
FC (correct)	0.00	33.23	9.61	24.14	11.34	39.86
NC (no color)	0.00	15.97	0.00	10.34	0.00	9.62
No PIH	0.96	95.21	14.78	55.42	20.27	79.72
EC (incorrect)	93.61	3.51	82.76	41.13	28.87	14.09
SC (incorrect)	5.43	1.28	2.46	3.45	50.86	6.19
PIH	99.04	4.79	85.22	44.58	79.73	20.28

Table 4: **Distribution of PIH and no PIH cases before and after PIH attention head ablation for the color task and semantic response types (%)**. NC, FC, SC, and EC denote no-, format-, soft-, and exact-copying, respectively. Correct denotes the ground-truth image color, while incorrect denotes the color in the misaligned prompt. Summed rows highlight the shift from PIH to image-grounded behavior after ablation.

Prompt: “Describe the 3 cats.” (N=2)	
Response	Copying Form
“There are 3 cats.”	Exact copying
“There are three cats.”	Soft copying
“There are 2 cats.”	Format copying
“There are two cats.”	No copying
Prompt: “Describe the purple banana.” (N=yellow)	
Response	Copying Form
“The purple banana . . .”	Exact copying
“The banana is purple.”	Soft copying
“The yellow banana . . .”	Format copying
“The banana is yellow.”	No copying

Table 5: **Definition of copying forms under misaligned prompts for count and color tasks.**

ter knockouts, shifts in attention mass after PIH-ablation, and the impact of PIH-ablation on model confidence. Through these analyses, we show that PIH ablation reduces hallucinations by both **inhibiting copying** of information from the prompt and **increasing attention** to the image.

6.1 Ablation Impacts Copying Form

To characterize how models incorporate misaligned prompt information, we categorize several forms of copying behavior, summarized in Table 5. *Soft-copying* occurs when the response contains the misaligned prompt information but deviates in structure from the prompt’s structure. *Format copying* occurs when the response contains the correct (visually grounded) information, while maintaining the same prompt format. *Exact copying* occurs when both the prompt information and the format are present. Using these definitions, we analyze

Model	Accuracy	Digit (%)		Word (%)	
		Before	After	Before	After
LLaVA-OV	Incorrect	7.57	0.25	45.62	15.29
	Correct	0.30	0.25	46.51	84.21
	Total	7.87	0.50	92.13	99.50
Qwen-VL	Incorrect	28.44	4.53	33.11	16.94
	Correct	0.05	0.00	38.40	78.53
	Total	28.49	4.53	71.51	95.47
Janus-Pro	Incorrect	43.45	2.67	25.40	23.30
	Correct	0.50	0.14	30.65	73.89
	Total	43.95	2.81	56.05	97.19

Table 6: **Response format and correctness before and after PIH attention head ablation.** Digit responses indicate copying format used in the prompt, while word responses correspond to spelled-out number expressions.

copying behavior first in the counting task and then in the color recognition task.

Table 6 shows copying patterns in the counting task, where the input prompts always include the number as a digit (i.e., “Describe the N_{digit} [objects]”). Prior to ablation, responses that include the number as a digit are more likely to be incorrect. For example, Janus-Pro outputs a digit in 43.95% of the cases, consisting of 43.45% of the incorrect cases, and 0.05% correct cases. This indicates that incorrect responses stem from exact copying of misaligned prompt information. After ablating PIH-heads, digit responses are almost entirely eliminated across all models. Notably, the reduction in digit responses occurs for both incorrect and correct outputs, showing that the suppression of format copying is not merely a byproduct of improved numerical accuracy.

The different copying forms in Table 4 show that

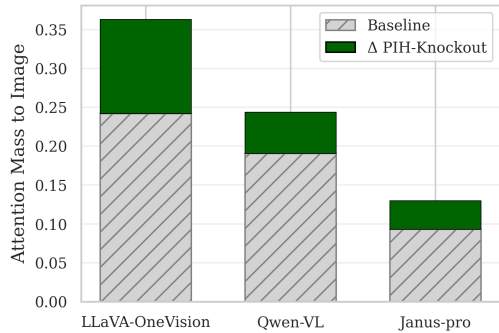


Figure 4: Image attention mass at the layer with the largest intervention-induced change for each model. Namely, LLaVA-OneVision (layer 2, $\Delta=0.121$), Qwen-VL (layer 27, $\Delta=0.053$), and Janus-pro (layer 22, $\Delta=0.037$). Gray bars indicate the baseline attention mass for image tokens, while green bars show an increase after PIH-knockout.

PIH in the color task is likewise associated with copying the format of the prompt. Before ablation, models frequently reproduce the prompt’s structure through format copying responses, accounting for over 90% of outputs for both LLaVA-OneVision and Janus-Pro. After PIH head ablation, this behavior is substantially reduced and accompanied by a corresponding increase in free-form descriptions.

These results suggest that when PIH-heads are active, models tend to reproduce both the content and the structure of misaligned prompts; when they are ablated, responses shift toward freer, less prompt-constrained descriptions that better reflect visual evidence. This parallel between counting and color tasks indicates that PIH is not limited to numerical reasoning, but reflects a broader tendency to copy information from the prompt, even when it is inconsistent with the image.

6.2 PIH-Ablation Shifts Attention Mass to Image Tokens

One notable outcome of PIH head ablation is observed in LLaVA-OneVision, where performance improves even for non-conflicting prompt-image cases. As shown in Table 1, baseline counting accuracy increases by 4.35%, despite the absence of misleading prompt information. Similarly, in the color task, PIH ablation yields a strong reduction in prompt-copying, with PIH reduced by 94%. These gains suggest that PIH ablation affects more than prompt-following, potentially altering how models balance textual and visual inputs.

To investigate this effect, we analyze changes in *attention mass* from assistant tokens to image

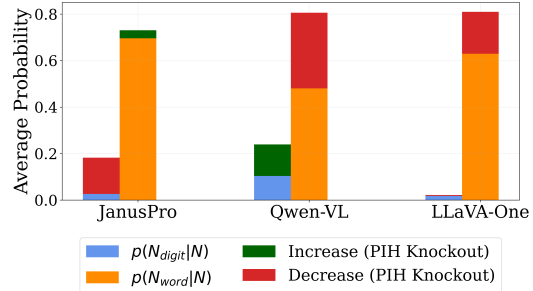


Figure 5: **Impact of knockout on the probability of the correct count in digit (blue) and word form (orange).** We measure the probabilities given the prompt “Describe the N_{digit} [objects]”. The knockout effects are shown in green (increase) and red (decrease).

tokens. For each layer, we compute the fraction of attention allocated to image tokens, with the remainder allocated to text tokens (summing to 1). Figure 4 shows the layer with the largest shift in attention from text to image, after PIH ablation: layer 2 in LLaVA-OneVision, 22 in Qwen-VL, and 27 in Janus-Pro; see Appendix H for all layers. PIH head ablation increases image attention across all models, most strongly in LLaVA-OneVision ($\Delta = 0.121$) which shows an increase in an early model layer. This is significant, as previous work has shown the importance of visual information processing in early layer for multimodal integration (Nikankin et al., 2025). These indicate that, particularly in LLaVA-OneVision, PIH head ablation induces a reallocation of attention toward visual inputs, providing a plausible explanation for the observed improvements in both baseline counting accuracy and robustness to PIH.

6.3 PIH Heads Mediate Model-Specific Copying Mechanisms

Beyond analyzing changes in model outputs, we seek to understand how PIH head ablation affects ground-truth probabilities when presented with prompts that do not conflict with the image. We examine the conditional probability of the ground-truth count N when expressed as either a digit (N_{digit}) or a word (N_{word}) under prompts of the form “Describe the N_{digit} [objects]”, shown in Figure 5. We use a digit-word analysis because it offers a controlled way to distinguish format-copying from soft-copying by analyzing a single token.

Janus-Pro inhibits format copying when correcting misaligned prompts. Figure 5 demonstrates that $P(N_{\text{digit}} | N_{\text{digit}})$ significantly decreases when

ablating PIH-Heads, which indicates a reduction in format-copying in the counting task. This is supported by the finding that PIH ablation in Janus-Pro reduces format copying by nearly 30%, a shift that is consistently accompanied by an increase in correct, free-form responses. While $P(N_{\text{digit}} | N_{\text{digit}})$ decreases, PIH ablation increases $P(N_{\text{word}} | N_{\text{digit}})$, indicating that when correcting misaligned prompts, Janus-Pro more frequently produces the ground-truth response in a word-based, free-form format rather than copying the digit form used in the prompt. Additionally, our findings show that PIH ablation produces only a modest shift in attention mass toward the image, consistent with almost no change in baseline counting performance. Together, these results suggest that PIH ablation in Janus-Pro does not increase visual reliance but primarily inhibits prompt-based format copying when correcting misaligned prompts.

Qwen-VL encourages format copying when correcting misaligned prompts. In contrast to Janus-Pro, we find that PIH-ablation causes Qwen-VL to more heavily rely on the prompt *format*. Figure 5 shows a 32.6% decrease in probability of $P(N_{\text{word}} | N_{\text{digit}})$ with a corresponding *increase* in corrected responses that follow the digit format of the prompt. Namely, $P(N_{\text{digit}} | N_{\text{digit}})$ increases by 13.6%. This is consistent with Table 4, which shows that Qwen-VL is the only model for which PIH head ablation increases format copying, with copying rates rising from 40.21% to 53.95%. While PIH ablation increases image attention and improves correction performance in Qwen-VL, Figure 5 reveals that these gains arise from a shift away from soft copying toward increased format copying, rather than a strict reduction in copying.

LLaVA-OneVision inhibits all forms of copying and shows a greater reliance on the image when correcting misaligned prompts. Figure 5 and Table 6 show that LLaVA-OneVision rarely produces digit-form responses even prior to PIH ablation, indicating format copying suppression. Next, Figure 5 demonstrates that $P(N_{\text{word}} | N_{\text{digit}})$ heavily decreases (18%), indicating LLaVA-OneVision also suppresses soft copying. In terms of image reliance, LLaVA-OneVision is the only model that, after PIH ablations, showed a substantial increase in performance on the baseline counting prompt, indicating an improvement in general counting abilities. In parallel, we find that PIH-ablation leads to the largest shift in attention mass toward the image

(+12%), indicating a greater reliance on the image.

Although implemented differently across models, PIH head ablation consistently reduces prompt copying and increases visual reliance, highlighting the promise of targeted interventions.

7 Conclusion

In this work, we study prompt-induced hallucinations in VLMs through a controlled object-counting task, isolating text-image conflicts. Through mechanistic analysis of three models, we show that such hallucinations stem from systematic prompt-based copying rather than failures of visual perception, and can be substantially reduced by ablating a small set of attention heads. We find that PIH is localized primarily to early layers of the language model, where prompt-based numerical information is likely processed and propagated before multimodal integration. Additionally, we find that Qwen-VL and LLaVA-OneVision share many of these influential heads, indicating they largely originate from the shared language-model backbone rather than vision-specific components.

Ablating PIH heads consistently shifts behavior toward visually grounded corrections, generalizes beyond counting to color prediction, and does not degrade general performance. Finally, we find that PIH heads mediate different forms of prompt copying with model-specific implementations. LLaVA-OneVision exhibits the strongest effects, including broad suppression of all forms of prompt-copying and substantially increased reliance on visual information after PIH head ablations. Our results highlight PIH as a mechanistically localized and model-dependent failure mode in VLMs, suggesting targeted interventions can improve visual grounding under prompt-image mismatches.

Limitations

Our study focuses on medium-scale VLMs ($\sim 7\text{B}$ parameters). While these models are representative of widely used multimodal architectures, the specific PIH heads identified here and the effects of their ablation may not generalize directly to substantially larger models.

Furthermore, our analysis relies on attention-based methods, which offer only a partial view of internal computation. As noted in prior work, attention patterns are often not directly interpretable (Jain and Wallace, 2019; Serrano and Smith, 2019; Brunner et al., 2019; Grimsley et al., 2020), and

heads with strong behavioral influence may not exhibit clear or human-aligned attention behaviors. Consistent with this, we do not observe simple or consistent attention patterns within PIH heads themselves.

Further, ablating PIH heads may induce secondary changes in the behavior of other attention heads or layers. These secondary effects on other attention heads may be more human-interpretable than the original PIH heads. We do not explicitly trace how these secondary changes emerge in this work, and leave their detailed analysis to future studies.

While we demonstrate all models follow the same general trend of a reduction in exact copying and an increase in attention mass to the image, we find that the manner in which copying is reduced is distinct for all three models. Determining the exact mechanism as to why PIH ablations causes a reduction in soft copying, while encouraging format copying in Qwen-VL and results in a strict decrease in format copying with an increase of soft copying Janus-Pro, is left for future works. Developing a better understanding the exact reason, whether its due to differences in architecture, patterns in selected heads or training data would help to give a more complete picture of prompt-induced hallucinations and the model components that mediate them.

Acknowledgments

This research was partly funded by the European Union (ERC, Control-LM, 101165402). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. YB and DA were partially supported by the Israel Science Foundation (grant no. 2942/25) and by Coefficient Giving. DA is supported by the Ariane de Rothschild Women Doctoral Program. K.M. and W.R. were supported by NSF grant 2313027. This work was partly supported by the National Science Foundation under Cooperative Agreement 2421782 and the Simons Foundation award MPS-AI-00010515 (NSF-Simons AI Institute for Cosmic Origins - CosmicAI, <https://www.cosmicai.org/>)

References

- Dana Arad, Hadas Orgad, and Yonatan Belinkov. 2024. Refact: Updating text-to-image models by editing the text encoder. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2537–2558.
- Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. 2019. On identifiability in transformers. *arXiv preprint arXiv:1908.04211*.
- Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025. [Janus-pro: Unified multimodal understanding and generation with data and model scaling](#). *Preprint*, arXiv:2501.17811.
- Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. 2025. Words or vision: Do vision-language models have blind faith in text? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3867–3876.
- Jessica Maria Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. 2024. Cognitive bias in decision-making with llms. In *Findings of the association for computational linguistics: EMNLP 2024*, pages 12640–12653.
- Sheridan Feucht, Eric Todd, Byron Wallace, and David Bau. 2025. The dual-route model of induction. *arXiv preprint arXiv:2504.03022*.
- Stella Frank, Emanuele Bugliarello, and Desmond Elliott. 2021. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. *arXiv preprint arXiv:2109.04448*.
- Deqing Fu, Ruohao Guo, Ghazal Khalighinejad, Ollie Liu, Bhuwan Dhingra, Dani Yogatama, Robin Jia, and Willie Neiswanger. 2024. Isobench: Benchmarking multimodal foundation models on isomorphic representations. *arXiv preprint arXiv:2404.01266*.
- Michal Golovanevsky, William Rudman, Michael Lepori, Amir Bar, Ritambhara Singh, and Carsten Eickhoff. 2025a. Pixels versus priors: Controlling knowledge priors in vision-language models through visual counterfactuals. *arXiv preprint arXiv:2505.17127*.
- Michal Golovanevsky, William Rudman, Vedant Palit, Ritambhara Singh, and Carsten Eickhoff. 2025b. [What do vlms notice? a mechanistic interpretability pipeline for gaussian-noise-free text-image corruption and evaluation](#). *Preprint*, arXiv:2406.16320.
- Christopher Grimsley, Elijah Mayfield, and Julia RS Bursten. 2020. Why attention is not explanation: Surgical intervention and causal reasoning about neural models. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1780–1790.

- Tianze Hua, Tian Yun, and Ellie Pavlick. 2025. How do vision-language models process conflicting information across modalities? *arXiv preprint arXiv:2507.01790*.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.
- Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.
- Erik Jones and Jacob Steinhardt. 2022. Capturing failures of large language models via human cognitive biases. *Advances in Neural Information Processing Systems*, 35:11785–11799.
- Nitish Joshi, Javier Rando, Abulhair Saparov, Najoung Kim, and He He. 2024. Personas as a way to model truthfulness in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6346–6359.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. [Llava-onevision: Easy visual task transfer](#). *Preprint*, arXiv:2408.03326.
- Fei-Fei Li, Marco Andreeto, Marc’Aurelio Ranzato, and Pietro Perona. 2022. [Caltech 101](#).
- Kenneth Li, Tianle Liu, Naomi Bashkinsky, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024b. Measuring and controlling instruction (in) stability in language model dialogs. *The Conference on Language Modeling COLM 2024*.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 292–305.
- Simon Malberg, Roman Poletukhin, Carolin Schuster, and Georg Groh Groh. 2025. A comprehensive evaluation of cognitive biases in llms. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 578–613.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2023. [Locating and editing factual associations in gpt](#). *Preprint*, arXiv:2202.05262.
- Aaron Mueller, Atticus Geiger, Sarah Wiegrefe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fiotto-Kaufman, Tal Haklay, Michael Hanna, et al. 2025. [Mib: A mechanistic interpretability benchmark](#). *arXiv preprint arXiv:2504.13151*.
- Yaniv Nikankin, Dana Arad, Yossi Gandelsman, and Yonatan Belinkov. 2025. Same task, different circuits: Disentangling modality-specific mechanisms in vlms. *arXiv preprint arXiv:2506.09047*.
- Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.
- Francesco Ortu, Zhijing Jin, Diego Doimo, and Alberto Cazzaniga. 2025. When seeing overrides knowing: Disentangling knowledge conflicts in vision-language models. *arXiv preprint arXiv:2507.13868*.
- Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. 2023. Teaching clip to count to ten. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3170–3180.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.
- Renjie Pi, Kehao Miao, Li Peihang, Runtao Liu, Jiahui Gao, Jipeng Zhang, and Xiaofang Zhou. 2025. Pointing to a llama and call it a camel: On the sycophancy of multimodal large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20177–20191.
- Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. 2024. Fine-tuning enhances existing mechanisms: A case study on entity tracking. *arXiv preprint arXiv:2402.14811*.
- Nikhil Prakash, Natalie Shapira, Arnab Sen Sharma, Christoph Riedl, Yonatan Belinkov, Tamar Rott Shaham, David Bau, and Atticus Geiger. 2025. Language models use lookbacks to track beliefs. *arXiv preprint arXiv:2505.14685*.
- Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. 2024. Vision language models are blind. In *Proceedings of the Asian Conference on Computer Vision*, pages 18–34.
- William Rudman, Michal Golovanevsky, Amir Bar, Vedant Palit, Yann LeCun, Carsten Eickhoff, and Ritambhara Singh. 2025. Forgotten polygons: Multimodal large language models are shape-blind. *arXiv preprint arXiv:2502.15969*.

- Emmanuelle Salin, Badreddine Farah, Stéphane Ayache, and Benoit Favre. 2022. Are vision-language transformers learning multimodal representations? a probing perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11248–11257.
- Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? *arXiv preprint arXiv:1906.03731*.
- Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.
- Adi Simhi, Jonathan Herzig, Itay Itzhak, Dana Arad, Zorik Gekhman, Roi Reichart, Fazl Barez, Gabriel Stanovsky, Idan Szpektor, and Yonatan Belinkov. 2025. Hack: Hallucinations along certainty and knowledge axes. *arXiv preprint arXiv:2510.24222*.
- Yingtian Tang, Yutaro Yamada, Yoyo Zhang, and Ilker Yildirim. 2023. [When are lemons purple? the concept association bias of vision-language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 14333–14348. Association for Computational Linguistics.
- SMTI Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 6.
- Angela van Sprang, Laurens Samson, Ana Lucic, Erman Acar, Sennay Ghebreab, and Yuki M Asano. 2025. Same content, different answers: Cross-modal inconsistency in mllms. *arXiv preprint arXiv:2512.08923*.
- An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. 2025. Vision language models are biased. *arXiv preprint arXiv:2505.23941*.
- Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. [Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution](#). *Preprint, arXiv:2409.12191*.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.
- Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.
- Rongwu Xu, Brian Lin, Shujian Yang, Tianqi Zhang, Weiyang Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2024a. The earth is flat because...: Investigating llms’ belief towards misinformation via persuasive conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16259–16303.
- Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024b. Knowledge conflicts for llms: A survey. *arXiv preprint arXiv:2403.08319*.
- Qianqi Yan, Yue Fan, Hongquan Li, Shan Jiang, Yang Zhao, Xinze Guan, Ching-Chen Kuo, and Xin Eric Wang. 2025. Multimodal inconsistency reasoning (mmir): A new benchmark for multimodal reasoning models. *arXiv preprint arXiv:2502.16033*.
- Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, Yu-Yang Liu, and Li Yuan. 2023. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469*.
- Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.
- Huixuan Zhang, Junzhe Zhang, and Xiaojun Wan. 2023a. Quantity matters: Towards assessing and mitigating number hallucination in large vision-language models. *arXiv preprint arXiv:2403.01373*.
- Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. 2023b. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*.
- Zhuoran Zhang, Tengyue Wang, Xilin Gong, Yang Shi, Haotian Wang, Di Wang, and Lijie Hu. 2025. When modalities conflict: How unimodal reasoning uncertainty governs preference dynamics in mllms. *arXiv preprint arXiv:2511.02243*.
- Jinkun Zhao, Lei Huang, and Wenjun Wu. 2025. What color is it? a text-interference multimodal hallucination benchmark. *arXiv preprint arXiv:2511.13400*.
- Tinghui Zhu, Qin Liu, Fei Wang, Zhengzhong Tu, and Muhao Chen. 2024. Unraveling cross-modality knowledge conflicts in large vision-language models. *arXiv preprint arXiv:2410.03659*.

A Experimental Details

A.1 Computational Requirements

All experiments in this work were conducted using a single NVIDIA RTX 3090 GPU with 24 GB of memory. Our analysis does not involve model training or fine-tuning, and instead relies on forward inference, attention-head ablation, and probability extraction. The complete set of experiments, including exploratory analyses and ablations not included in the final paper, consumed approximately 200-300 GPU hours. Peak memory usage occurred during attention output analysis, but all experiments fit within a single GPU.

A.2 Licenses and Third-Party Usage

This work is implemented using PyTorch (Paszke et al., 2019), an open-source deep learning framework licensed under the BSD license, and the Hugging Face Transformers library (Wolf et al., 2019), licensed under Apache 2.0. All software usage complies with their respective license terms. For evaluation, we use CountBench (Paiss et al., 2023) and Visual CounterFact (Golovanevsky et al., 2025a) licensed under the Apache 2.0 license.

A.3 Models

We evaluate a three diverse VLMs. Table 7 provides details on the open-source models used in our experiments. These models are sourced from Hugging Face, with their specific repository paths listed for reproducibility.

A.4 Datasets

We use CountBench (Paiss et al., 2023) for our counting task and Visual CounterFact (Golovanevsky et al., 2025a) for our color identification task. CountBench consists of 491 text-image pairs. We adapt CountBench using our novel misaligned prompts, yielding a total of 3,437 image-prompt pairs. We use the “color” split of the Visual CounterFact dataset, which comprises 493 base-image-label pairs. We adapt Visual CounterFact with our base and misaligned prompts, yielding 2,465 image-prompt pairs.

A.5 Generation Details

For each image with ground-truth object count N , we first present the model with a *baseline prompt*, $P_B =$ “How many [objects] are there in the image?”, and retain only samples for which the model correctly predicts N . We then introduce *misaligned*

prompts asking the model to “Describe the $N + k$ [objects] in the image,” where $k \in \{1, 2, \dots, 5\}$. Here, the offset k serves as a *discrepancy distance*, measuring the magnitude of misalignment between the prompt and the visual input.

We allow the model to generate long-form responses up to 75 tokens and use a rule-based algorithm to determine accuracy by extracting the first non-negated numerical reference in the response (e.g., excluding constructions such as “there are not eight waterlilies”). A PIH occurs when the model proceeds to describe $N + k$ objects without correcting the count, despite contradictory visual evidence.

A.6 Accuracy on CountBench

We first evaluate the baseline accuracy of LLaVA-OneVision, Qwen-VL and Janus-Pro on CountBench (Paiss et al., 2023). Table 8 breaks performance down by the number of objects. All models perform reliably at very small counts (2–4 objects), but accuracy declines as counts increase. Notably, there are sharp dips at **5 objects** (e.g., LLaVA-1.6: 87.27% to 65.91%, QwenVL: 90.91% to 75.00%) and again at **7 objects** (Janus-Pro: 87.72% to 55.56%, LLaVA-OneVision: 89.47% to 48.15%). These discontinuities suggest the presence of training biases that systematically disadvantage certain numbers, a phenomenon supported by prior work linking VLM counting performance to biases in object frequency distributions seen during pretraining (Rudman et al., 2025; Vo et al., 2025).

B PIH with Large Offsets

Figure 6 extends the analysis in Figure 2 to large discrepancy offsets ($k \in \{10, 20, 50\}$), where the prompted count substantially exceeds the number of objects present in the image. Across all three models, we observe that once the number of ground-truth objects exceeds approximately four, prompt-induced hallucinations persist even under extreme misalignment: models frequently produce responses matching the prompted count, while corrections to the true count become rare.

However, LLaVA-OneVision exhibits a notable deviation from this pattern. While Qwen-VL and Janus-Pro show near complete convergence toward prompt matching at higher object counts, regardless of the offset magnitude, LLaVA-OneVision maintains a wider separation between prompt-aligned and ground-truth-aligned responses. In particu-

Model Name	Underlying LM	Vision Encoder	Size	HuggingFace Path
LLaVA-OneVision	Qwen2	SigLIP	7B	llava-hf/llava-ov-hf
Qwen-VL	Qwen2	DFN-ViT w/ RoPE-2D	7B	Qwen/Qwen2-VL-7B-Instruct
Janus Pro	DeepSeek-LLM	SigLIP-Large-Patch16-384	7B	deepseek-ai/Janus-Pro-7B

Table 7: Details of evaluated models, including underlying language models, vision encoders, sizes, and Hugging Face model paths.

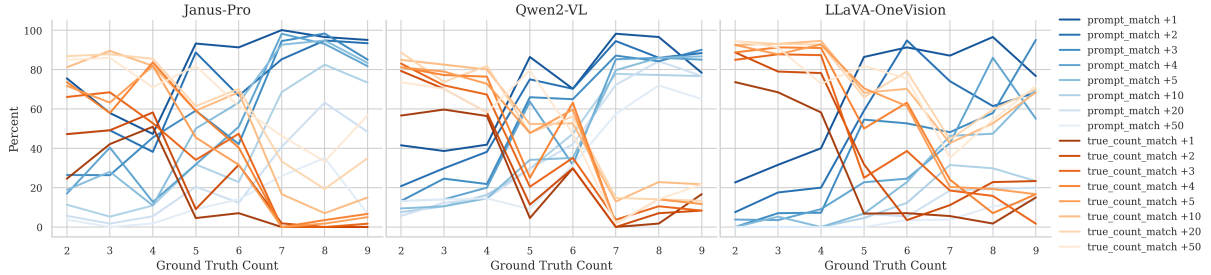


Figure 6: **PIH rates across different number of ground-truth objects.** Blue lines show the percentage of outputs matching the prompted number for different offsets k , while orange lines show matches to the true count. Bias grows with object count, with four objects marking the approximate threshold at which models stop correcting the prompt. Prompt match and true count match do not sum to 100% because some model responses do not contain a number.

Num	LLaVA-OV	Janus-Pro	QwenVL
2	90.57	84.91	90.57
3	91.23	87.72	85.96
4	87.27	85.45	90.91
5	65.91	70.45	75.00
6	89.47	87.72	85.96
7	48.15	55.56	51.85
8	71.93	80.70	75.44
9	68.33	86.67	71.67

Table 8: Performance across different numbers of objects on CountBench.

lar, for offsets of +10, +20, and +50, LLaVA-OneVision continues to produce a non-trivial fraction of corrections to the true count, even when the discrepancy is large.

This behavior indicates that LLaVA-OneVision is more resilient to extreme prompt-image mismatches, retaining some capacity to reject implausible prompt specifications at higher object counts. The persistence of ground-truth corrections under large offsets is consistent with our broader findings that LLaVA-OneVision exhibits stronger visual grounding and reduced reliance on prompt-based copying compared to Qwen-VL and Janus-Pro. Nevertheless, even for LLaVA-OneVision, prompt-induced hallucinations dominate once object counts are sufficiently high, underscoring the strength of prompt-driven biases in this regime.

C Base Probabilities for All Models

Figure 7 shows the probabilities for the true count N and neighboring counts $N+k$ under the baseline prompt (P_B) for all models.

D Attention Knockout Details

To analyze the causal role of individual attention heads in PIH, we perform attention head knockouts using mean ablation (Wang et al., 2022; Mueller et al., 2025). Let $H^{(l,h)} \in \mathbb{R}^{T \times d}$ denote the output of head h at layer l across T tokens. We compute the head-wise mean

$$\mu^{(l,h)} = \frac{1}{T} \sum_{t=1}^T H_t^{(l,h)},$$

and replace the head output at every token position with this mean, i.e.,

$$\tilde{H}_t^{(l,h)} = \mu^{(l,h)} \quad \forall t \in \{1, \dots, T\}.$$

This intervention removes token-specific information carried by the head while preserving its overall activation magnitude.

We apply this procedure head-by-head and quantify knockout success as the fraction of samples that switch from the prompted (hallucinated) count to the true image-grounded count under the misaligned prompt. We rank heads by their corrective effect and select the top- m heads per model. We

Model	Rank	Head	Success (%)
Qwen-VL			
	1	L0H3	38.90
	2	L0H6	26.89
	3	L11H15	16.71
	4	L0H11	8.09
	5	L15H8	7.57
	6	L14H13	7.05
	7	L0H10	6.79
	8	L7H3	6.53
	9	L0H24	6.01
	10	L8H13	5.48
Janus-Pro			
	1	L0H20	31.18
	2	L1H7	27.96
	3	L14H27	17.74
	4	L12H9	17.47
	5	L11H18	15.59
	6	L13H2	13.98
	7	L0H24	13.44
	8	L14H28	13.17
	9	L8H11	12.90
	10	L8H3	12.63
LLaVA-OneVision			
	1	L0H3	48.99
	2	L0H6	37.92
	3	L0H26	30.20
	4	L11H15	27.85
	5	L0H23	27.52
	6	L0H24	21.48
	7	L0H15	20.81
	8	L14H9	17.79
	9	L0H11	16.78
	10	L17H22	15.10

Table 9: Top-10 attention heads per model ranked by individual knockout success. Success is measured as the percentage of samples that switch from PIH to the correct image-grounded count when the head is ablated.

then perform a second-stage grouped knockout in which all top- m heads are ablated simultaneously for $m \in \{1, 3, 5, 10\}$, selecting m as a hyperparameter per-model.

E Full List of PIH-heads

Table 9 reports the top-ranked attention heads for each model, ranked by their individual knockout success. Knockout success is defined as the fraction of misaligned-prompt samples that switch from matching the prompted (hallucinated) count to matching the true image-grounded count when a single attention head is ablated via mean ablation. Higher values therefore indicate a stronger causal role in propagating prompt-induced numerical errors.

For both Janus-Pro and LLaVA-OneVision, performance improves monotonically with larger head sets, with the best results achieved when ablating

Model	Best-1	Best-3	Best-5	Best-10
Janus-Pro	31.18	65.06	73.19	77.91
LLaVA-OV	48.99	72.35	73.68	83.08
Qwen-VL	38.90	67.14	66.86	3.43

Table 10: Global mean success rates (%) under different PIH head selection strategies. Best- m denotes ablating the top- m PIH heads ranked by individual knockout success.

the top 10 PIH heads. In contrast, Qwen-VL exhibits its highest success rate when ablating only the top 3 heads, with performance degrading substantially when larger groups are removed. This divergence suggests that while PIH in Janus-Pro and LLaVA-OneVision is distributed across a broader set of attention heads, Qwen-VL relies on a smaller subset of highly specialized heads, and removing additional heads begins to interfere with general generation behavior. Accordingly, we select the top-10 heads as the final PIH head sets for Janus-Pro and LLaVA-OneVision, and the top-3 heads for Qwen-VL in all subsequent experiments.

E.1 Attention Patterns of PIH-heads

To gain qualitative insight into the behavior of PIH attention heads across models, we visualize the attention patterns of the highest-ranked PIH head in each model, Qwen-VL (L0H3), LLaVA-OneVision (L0H3), and Janus-Pro (L0H20), over the full input and generation sequences, averaged across all samples. Representative examples are shown in Figure 8.

In all visualizations, query tokens are shown on the y-axis and key tokens on the x-axis, such that each row of attention scores sums to 1. Since images are represented as long sequences of image pad tokens (often several hundred), we aggregate the majority of these tokens into a single row and column, denoted as *IMAGE_PADS*, by summing their attention scores. We retain the first three and last three image pads as individual entries to preserve boundary structure and improve interpretability.

The PIH head in Qwen-VL (L0H3) exhibits a mixed attention pattern. A substantial portion of attention lies along the diagonal over the text sequence, indicating behavior similar to local self-attention. At the same time, we observe consistent and structured attention within the image-token region, including attention to the aggregated

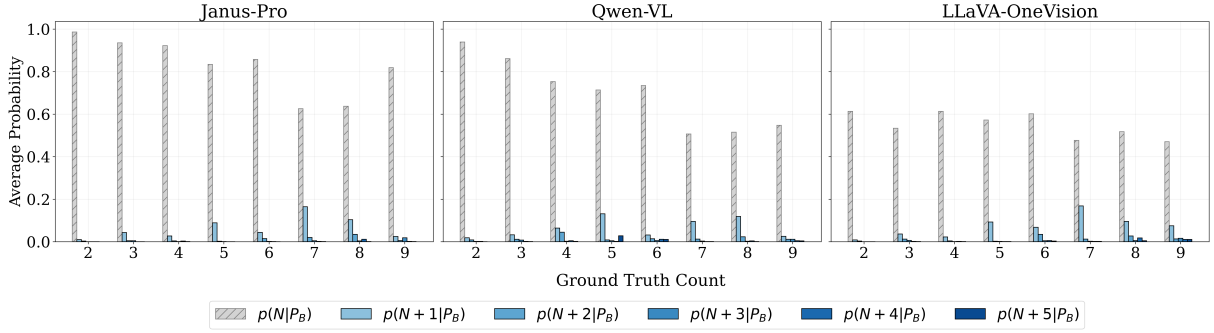


Figure 7: **Base probabilities on the object counting task.** We plot the probability assigned to the true count N and neighboring counts $N + k$ ($k \in \{1, \dots, 5\}$) given the prompt “How many [objects] are in the image?”. As object counts increase, confidence in N decreases and probability mass spreads to neighboring counts, matching the increased susceptibility to PIH.

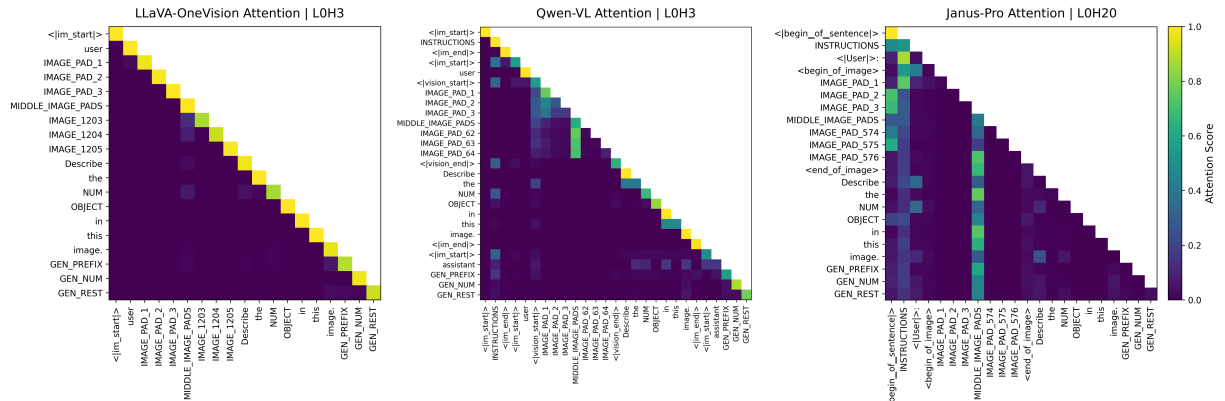


Figure 8: Attention patterns for the top-ranked PIH attention head in three multimodal models, averaged over all samples and computed over the full input–generation sequence. **Left:** LLaVA-OneVision (L0H3), **Center:** Qwen-VL (L0H3), **Right:** Janus-Pro (L0H20). Query tokens are shown on the y-axis and key tokens on the x-axis, with each row summing to 1. Image pad tokens are largely aggregated into a single row and column (*IMAGE_PADS*), with boundary tokens retained.

IMAGE_PADS and nearby vision boundary tokens. This combination places Qwen-VL L0H3 between purely textual self-attention and fully image-focused attention.

In contrast, the PIH head identified in Janus-Pro (L0H20) shows attention patterns that are predominantly concentrated on image tokens. The diagonal structure characteristic of local textual self-attention is largely absent, and attention is instead focused almost entirely within the image-token region across many query positions. This behavior suggests that the PIH head in Janus-Pro functions primarily as an image-focused attention head.

The PIH head in LLaVA-OneVision (L0H3) behaves almost entirely as a local self-attention head over text. Attention is concentrated along the diagonal, with little to no direct attention to image tokens in the unablated model. Ablation of the PIH head in LLaVA-OneVision leads to the strongest

performance improvements among the three models. One possible explanation is that the targeted PIH head in LLaVA-OneVision is not primarily focused on image tokens, but instead behaves as a largely textual self-attention head. As a result, ablating this head may predominantly disrupt copying or local textual interactions, without directly impairing image processing. More generally, ablation may induce second-order effects across all models, whereby other heads adjust their attention patterns and potentially increase attention to image tokens. In the case of LLaVA-OneVision, we hypothesize that the combination of not removing an image-focused head and benefiting from such second-order redistribution effects contributes to its robustness and improved performance under PIH ablation.

For all three models, we do not observe a clear, isolated peak of attention on the generated number

token, so while some of the heatmap are suggestive of modality-specific routing, none are conclusive about how the numeric value is produced. This inconclusiveness is reasonable, as interpreting the internal function of attention heads from visualization alone remains a well-known challenge despite their demonstrated causal effects (Jain and Wallace, 2019; Serrano and Smith, 2019; Brunner et al., 2019; Grimsley et al., 2020).

F Output Examples After Knockouts

Table 11 shows examples of generated texts after PIH-head ablations.

G Color Task

To investigate whether PIH heads generalize beyond the counting task, we extend our framework to a color prediction task using the Visual CounterFact dataset (Golovanovsky et al., 2025a). This setting replaces numerical offsets with color offsets while preserving the overall experimental structure used in counting. We use "What color is the [object] in the image?" as the baseline prompt, and "Describe the $C + k$ [object]" as the misaligned prompt, where C denotes the ground-truth color of the object in the image and $k \in \{1, 2, 3\}$ denotes the perceptual "difference" from the C on the color wheel. In particular, $k = 1$ indicates a highly similar color and $k = 3$ indicates the *contrasting color* which is the further distance on the color wheel from C . For example, given $C = \text{red}$, we have orange, yellow, and green for $k = 1, 2, 3$, respectively.

H Layer-wise Attention Mass Shifts After PIH Ablation

While Figure 4 focuses on the single layer exhibiting the largest intervention-induced shift in attention mass, we additionally examine how PIH head ablation affects attention allocation *across all transformer layers*. Specifically, for each layer we compute the change in attention mass to image tokens relative to the baseline, with the complementary change applied to text tokens (since attention mass sums to one).

Figure 10 shows the layer-wise change in image and text attention mass for LLaVA-OneVision, Qwen-VL, and Janus-pro. Across all models, PIH ablation consistently increases attention to image tokens while reducing attention to text tokens, in-

dicating a systematic rebalancing of multimodal attention rather than a highly localized effect.

Notably, the magnitude and depth of this shift varies across architectures. LLaVA-OneVision exhibits a pronounced increase in image attention in early layers, whereas Qwen-VL and Janus-pro show smaller but more distributed changes concentrated in later layers. These trends suggest that PIH head ablation alters the global flow of multimodal information throughout the network.

Before PIH Knockout	After PIH Knockout
Qwen-VL	
<ul style="list-style-type: none"> The image shows a set of 52 glasses, which appear to be stemless wine glasses. They are filled with red wine, and the glasses are arranged in a grid pattern, with 13 rows and 4 columns. The image shows three pairs of leggings. The first pair is light pink, the second pair is white with a small pattern, and the third pair is also white but without any pattern. The image shows a total of four wineglasses. Two of them are filled with red wine and are placed side by side on a wooden table. The other two wineglasses are empty and are also placed on the table. 	<ul style="list-style-type: none"> The image shows two wine glasses. The glasses are stemless and have a modern, minimalist design. They are filled with red wine, and the glasses appear to be made of clear glass. The image shows two pairs of leggings. The pair on the left is light pink, and the pair on the right is white with a subtle pattern. The image shows two wine glasses filled with red wine. The glasses are placed on a wooden table, and there is a bottle of red wine and a basket of grapes in the background.
Janus-Pro	
<ul style="list-style-type: none"> The image shows four kids standing together outdoors. They are all wearing casual clothing, with one of them in a plaid shirt over a yellow t-shirt. The other three are in solid-colored shirts. The image shows four croissants, which are golden brown in color and have a flaky, layered texture. They are arranged on a wooden cutting board, with two croissants placed side by side and the other two slightly overlapping each other. The image shows 9 planes flying in formation. They are all blue with white markings and have the word "Canada" written on them. The planes are arranged in a tight, symmetrical pattern, with four planes on the top row and five on the bottom. 	<ul style="list-style-type: none"> The image shows two children standing close together outdoors. The child on the left has curly hair and is wearing a bright orange shirt with a red and blue checkered shirt over it. The child on the right has short hair and is wearing a dark shirt. The image shows two croissants placed on a wooden cutting board. They are golden brown, flaky, and have a slightly twisted shape, typical of traditional croissants. The image shows four blue aerobatic planes flying in formation. The planes are marked with the letters "AC" and have a distinctive design with white accents.
LLaVA-OneVision	
<ul style="list-style-type: none"> The image displays five different illustrations of classic convertible cars arranged in a two-by-three grid. Each car is depicted in a different color and style, showcasing a variety of vintage designs. The image shows a set of eight forks arranged in two rows against a blue background. Each fork has a polished, reflective surface, indicating they are made of a shiny metal, likely silver or a silver-plated material. The image displays a collection of 17 plastic water bottles arranged in a row against a white background. Each bottle appears to be filled with water, and they vary in shape and size. 	<ul style="list-style-type: none"> The image presents a collection of four vintage cars, each distinct in color and model, arranged in a two-by-two grid. The image shows a set of six ornate silver forks arranged in a row against a blue background. Each fork has a polished finish and features intricate designs on the handles. The image displays a collection of seven plastic water bottles arranged in a row against a white background, with cap colors pink, white, blue, white, blue, white, and white.

Table 11: Qualitative examples of generations before and after PIH attention head knockout. Before intervention, models not only copy the incorrect count implied by the misaligned prompt, but also hallucinate non-existing details. After ablating PIH heads, generations more consistently reflect the true number of objects present in the image across all three models.

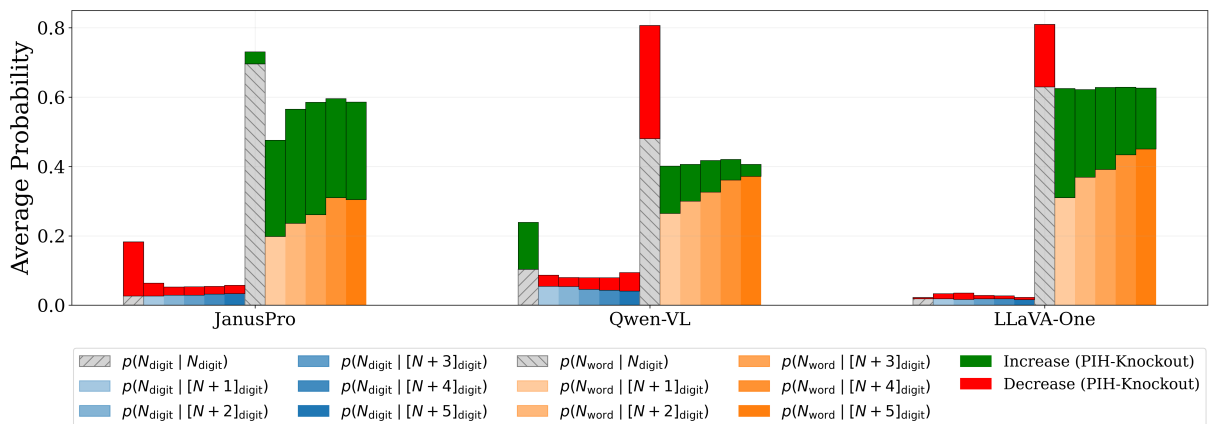


Figure 9: Impact of knockout on the probability of the ground-truth answer in both digit (blue) and word form (orange) given the prompt "Describe the $[N + k]_{digit}$ [objects]" for count. Note that when $k = 0$, there is no modality conflict in the prompt.

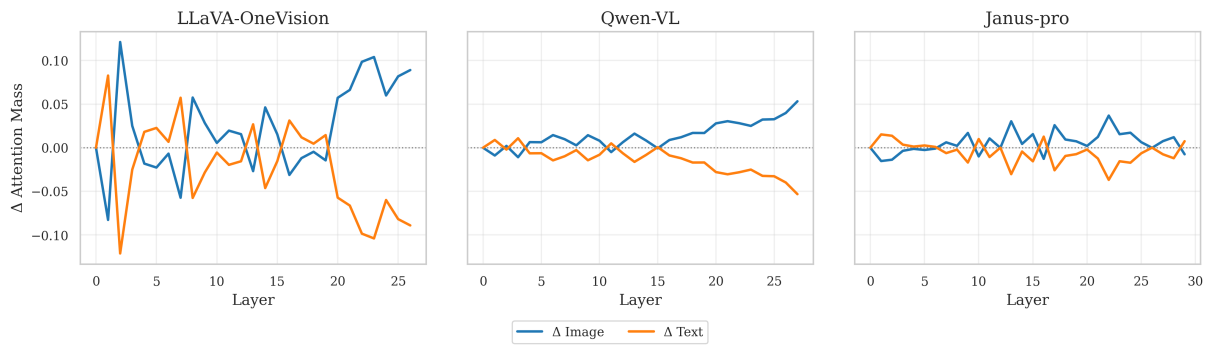


Figure 10: **Layer-wise change in attention mass after PIH head ablation.** For each model, we plot the change in attention mass to image tokens (blue) and text tokens (orange) as a function of transformer layer. Positive values indicate increased attention allocation relative to the baseline. Across all models, PIH ablation consistently shifts attention from text to image tokens, with LLaVA-OneVision exhibiting the strongest and earliest effect.