HIVE: Understanding Post-Hallucination Reasoning in Vision Language Models

Feng He[1*], Zhenting Wang[2], Qifan Wang[3], Qiang Guan[4], Dongfang Liu[1†], Ruixiang Tang[2], and Qiankun Li[5†]

[1] Purdue University
[2] Rutgers University
[3] Meta AI
[4] Kent State University
[5] Imperial College London

Abstract. Hallucinations in vision–language models (VLMs) are commonly treated as semantic errors, yet they often arise from partial or ambiguous visual evidence. Prior work mainly focuses on detecting or suppressing hallucinations at generation time, leaving the subsequent reasoning stage largely unexplored. In this work, we study *Post-Hallucination Reasoning* (PHR), the stage in which hallucinated semantics enter the model's inference context and influence downstream predictions. To systematically investigate PHR, we introduce the **HIVE** (**H**allucination **I**nference and **V**erification **E**ngine), an evaluation infrastructure that enables controlled comparisons between faithful and hallucinated captions. Across nine tasks and nine models, we observe structured modality-dependent patterns: hallucinated captions often improve accuracy on vision–language tasks, while text-only tasks exhibit limited or unstable effects. Further analyses show that hallucinated cues broaden semantic coverage and reshape reasoning dynamics while preserving stable inference. These findings highlight that hallucinated semantics may influence downstream reasoning once they enter the model's inference context. Understanding this post-hallucination stage is important for improving the reliability and interpretability of multimodal reasoning systems.

Keywords: VLMs · Hallucination Analysis · Multimodal Reasoning

1 Introduction

Vision–language models (VLMs) often operate under partial observability [11,23,24,31,40,56]. Occlusion, low resolution, and motion blur frequently lead them to produce speculative semantic completions [16,46,55]. According to community standards, such unverifiable content is considered *hallucination*. However, hallucinations arise naturally from incomplete visual evidence rather than isolated system failures [7, 9, 10, 14, 59]. Consequently, most existing studies focus

[*] Work done while visiting Purdue University.
[†] Corresponding authors.

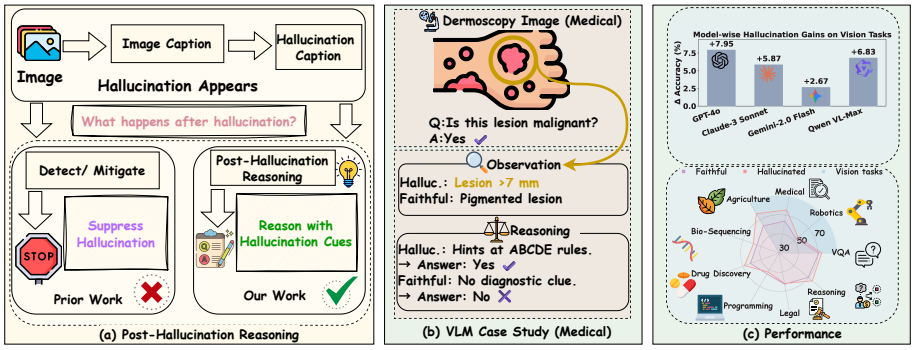

Fig. 1: Post–hallucination reasoning in VLMs. (a) Prior work mainly focuses on detecting or suppressing hallucinations, whereas we study reasoning that occurs after hallucinations appear. **(b)** A medical example illustrates how hallucinated cues (e.g., lesion size) can alter the model's observation and reasoning process. **(c)** Across models and vision tasks, hallucinations can systematically change task performance, revealing structured post–hallucination dynamics, especially in vision–language tasks.

on detecting [13,23,27,28,32,43,58] or mitigating hallucinations [37,38,47,53,61] at the moment they appear. Far less attention has been paid to what happens after hallucinated semantics enter the reasoning process of a VLM, leaving this post-hallucination stage largely unexplored.

Meanwhile, studies of LLMs show that chain-of-thought reasoning [4, 22, 29, 36, 37, 50, 51] does not strictly follow human causal logic: incorrect intermediate steps can still lead to correct conclusions. This observation suggests that reasoning may continue to evolve in latent space even when intermediate representations are imperfect or partially incorrect. By analogy, hallucinations in VLMs may also interact with downstream reasoning processes. However, whether such speculative cues help, hinder, or simply perturb reasoning remains unclear.

We refer to this phenomenon as **Post-Hallucination Reasoning (PHR)**, which describes the reasoning behavior that emerges after hallucinated semantics become part of the model's inference context and influence subsequent predictions. Despite the prevalence of hallucinations in multimodal systems, the structure and impact of PHR remain poorly understood. As illustrated in Fig. 1, once hallucinated semantics enter the model's inference context, they may reshape the observation and reasoning process, ultimately influencing downstream predictions. Across diverse benchmarks, we observe a consistent pattern: hallucinated semantic cues can sometimes *improve* downstream reasoning in vision-language tasks. This suggests that hallucinations may occasionally provide speculative semantic anchors that guide inference rather than merely degrading predictions. However, this effect is neither universal nor well understood. These observations raise two fundamental questions: **(1)** *When does PHR improve downstream reasoning?* **(2)** *Why does PHR lead to improved reasoning?*

To systematically study these questions, we introduce **HIVE** (**H**allucination **I**nference and **V**erification **E**ngine), an evaluation infrastructure for studying PHR. HIVE constructs paired faithful and hallucinated captions, verifies their validity, and enables controlled measurement of their influence on downstream predictions. By comparing raw inputs, faithful, and hallucinated captions under matched conditions, HIVE allows us to isolate the reasoning effects induced by hallucinated semantics and analyze the structure of PHR. Across 9 tasks and 9 models, we observe structured patterns: hallucinated semantic cues often improve performance on vision-language benchmarks, with gains varying systematically across model families and decoding settings.

To understand the mechanisms underlying PHR, we analyze hallucination-induced reasoning dynamics at multiple levels. **(I)** Hallucinated captions broaden semantic coverage and widen embedding distributions. **(II)** They modulate reasoning entropy, with correct predictions correlating with higher entropy.

These findings reveal **PHR** as an overlooked stage of multimodal inference, where hallucinated semantics can reshape and sometimes enhance downstream reasoning rather than merely introducing noise. Our contributions are:

❶ **Post-Hallucination Reasoning Phenomenon.** We identify and characterize **PHR**, a previously overlooked stage in which hallucinated semantics influence downstream reasoning in VLMs across diverse task settings. We observe structured patterns showing that hallucinated semantic cues can systematically reshape model predictions rather than merely introducing noise.

❷ **Evaluation Infrastructure for PHR.** We present HIVE, an evaluation infrastructure designed to study PHR through controlled semantic comparisons. HIVE organizes caption generation, hallucination discrimination, and downstream evaluation under matched configurations, enabling paired comparisons between faithful and hallucinated captions and allowing us to isolate and quantify the downstream impact across multiple models and task domains.

❸ **Mechanistic Analysis of PHR.** We analyze PHR across the input, reasoning-process, and output levels. Our results show that hallucinated captions broaden semantic coverage, induce distinctive distributional shifts, and modulate reasoning dynamics associated with successful predictions. These findings reveal how hallucinated semantics reshape reasoning trajectories in multimodal inference.

2 Related Work

Hallucinations in VLMs: Detection and Mitigation. Hallucination denotes content inconsistent with the given input. Hallucinations are widely observed in LLM and VLM outputs, appearing in tasks such as summarization [60] and open-domain QA [41]. Their presence undermines reliability in high-stakes domains, including healthcare [20,35] and legal decision support [1,8]. Two major threads dominate: detection and mitigation. For detection, studies propose factuality metrics and benchmarks such as FactCC [17], QAGS [45], TruthfulQA [25], and Q2 [12], along with agreement- or internals-based judging [6, 32, 43]. For mitigation, approaches include (I) instruction-level tuning [26, 54, 57], (II) constrained decoding [3, 19, 34, 43], and (III) training or retrieval augmentation

[21, 33, 42]. These lines of work largely treat hallucinations as undesirable errors and focus on detecting or suppressing them, leaving the role of hallucinated semantics in subsequent reasoning largely unexplored.

Intermediate Reasoning and Latent Dynamics in LLMs. Recent studies on language-only models reveal that a model's reasoning process does not necessarily align with its explicit chain-of-thought text. Multiple works show that LLMs may generate incorrect, inconsistent, or partially fabricated intermediate steps while still producing correct final answers [39, 49]. This phenomenon suggests that reasoning continues to evolve in the latent space and is not strictly determined by the faithfulness of intermediate text. Further analyses indicate that LLMs can maintain stable internal inference even when intermediate steps contain hallucinated elements [15, 18, 48]. These findings imply that unfaithful or speculative content does not terminate reasoning; instead, the model proceeds to a subsequent reasoning stage that may still yield coherent conclusions.

Hallucinations in Multimodal Reasoning Models. Recent work has begun to study hallucination effects in multimodal reasoning, showing that imperfect perceptual cues can influence reasoning outcomes in VLMs. Studies such as More Thinking, Less Seeing [52], More Thought, Less Accuracy [44], MIRAGE [5], and multimodal CoT analyses [30] indicate that hallucinations interact with reasoning dynamics beyond perception-level errors. A recent survey further summarizes evaluation and detection efforts in this area [2]. However, these methods mainly analyze either intermediate reasoning traces or benchmark-level behavior, without explicitly studying how hallucinated semantics, once used as input context, affect downstream reasoning.

Different from existing methods, we focus on a largely overlooked stage in VLM inference: **Post-Hallucination Reasoning (PHR)**, where hallucinated semantics, once generated and adopted by the model, may influence subsequent reasoning and task predictions. Prior work on VLM hallucinations focuses almost entirely on whether hallucinations occur, why they arise, and how to reduce them. Yet little attention has been paid to how hallucinated semantics affect the reasoning processes that follow. Consequently, the structure and impact of *PHR* remain poorly understood, and the field lacks frameworks for controlled semantic interventions to study it.

3 Evaluation Pipeline for PHR

3.1 Problem Setup

HIVE is designed as an evaluation infrastructure for studying PHR in VLMs. It examines how hallucinated semantics, once generated and adopted by the model, influence subsequent reasoning and task predictions compared with faithful semantics. Rather than introducing a new model or training method, HIVE follows established community practices, including prompting-based caption generation and hallucination detection, and organizes them into a controlled semantic intervention framework for analyzing the causal effects of hallucinated semantics. To examine this question, we organize HIVE around three principles.

① **Fair comparison.** We must ensure that the only difference between faithful and hallucinatory captions lies in the presence of hallucination itself. Thus,

captions are generated from a unified source with identical prompts, temperature, and token budget. This design rules out confounds and allows us to focus on the difference in downstream performance across tasks and models, measured as the accuracy gap $\Delta(H-F)$ between hallucinated and faithful augmentations.

② **Task-agnostic hallucination generation.** Hallucinations naturally arise when an LLM is asked to produce a free-form caption of any input, regardless of modality. Some captions remain faithful, while others introduce unverifiable or speculative elements. This inherent property enables us to adapt the paradigm seamlessly across diverse textual and multimodal tasks.

③ **Reliable discrimination.** Hallucination detection is inherently imperfect, and even human annotators may disagree. To enhance robustness, we adopt an ensemble of detectors that independently judge each caption, and apply majority voting to obtain the final label. This ensemble strategy ensures that the framework remains reliable under noisy classifiers. Formally, let $x \in \mathcal{X}$ denote an input with label $y \in \mathcal{Y}$, and $f : \mathcal{X} \times \mathcal{C} \to \mathcal{Y}$ a task-specific model. Each input can be paired with a faithful caption $\boldsymbol{C}_F$ or a hallucinatory caption $\boldsymbol{C}_H$. Such a triplet design ensures that hallucination is the only varying factor across conditions, thus enabling a controlled and interpretable evaluation. Formally, we define these conditions:

$$y_{\text{RAW}} = \underbrace{\boldsymbol{f}(x)}_{\text{Raw}}, \quad y_F = \underbrace{\boldsymbol{f}(x \,\|\, \boldsymbol{C}_F)}_{+\text{ Faithful}}, \quad y_H = \underbrace{\boldsymbol{f}(x \,\|\, \boldsymbol{C}_H)}_{+\text{ Hallucinatory}}, \tag{1}$$

where $\|$ denotes concatenation with the task instruction. Given an evaluation metric $\mathcal{L}(\hat{y}, y)$ (instantiated as accuracy in our experiments), we quantify the hallucination effect as

$$\boldsymbol{\Delta}(H-F) \;=\; \mathbb{E}_{(x,y)\sim\boldsymbol{\mathcal{D}}}\Big[\, \mathcal{L}\big(\boldsymbol{f}(x \,\|\, \boldsymbol{C}_H),\, y\big) - \mathcal{L}\big(\boldsymbol{f}(x \,\|\, \boldsymbol{C}_F),\, y\big) \,\Big]. \tag{2}$$

This paired comparison isolates hallucination as a controlled experimental variable and enables apples-to-apples analysis of **PHR** across models and tasks.

3.2 HIVE Evaluation Pipeline

The HIVE evaluation pipeline consists of three modules: (I) **Caption Generator**, (II) **Caption Discriminator**, and (III) **Task Solver**. Given an input instance, the pipeline proceeds through these components to ensure consistent and controlled evaluation across tasks. The Caption Generator first takes the raw input (text, image, or structured record) and produces a set of candidate captions under a unified, task-agnostic prompt, which may include both faithful and hallucinatory semantics. The Caption Discriminator then evaluates these candidates and classifies each as either faithful ($\boldsymbol{C}_F$) or hallucinatory ($\boldsymbol{C}_H$); only contrasted pairs $\langle \boldsymbol{C}_F, \boldsymbol{C}_H \rangle$ with majority agreement among detectors are retained. Finally, the Task Solver integrates the original input x, one of the paired captions, and a task-specific instruction to produce the final prediction y. This design yields three controlled conditions Raw ($y_{\text{RAW}} = \boldsymbol{f}(x)$), +Faithful ($y_F = \boldsymbol{f}(x \,\|\, \boldsymbol{C}_F)$), and +Hallucinatory ($y_H = \boldsymbol{f}(x \,\|\, \boldsymbol{C}_H)$).

Caption Generator. The Caption Generator aims to produce diverse semantic candidates that may include both faithful and hallucinatory variants. All captions are generated from a unified source using the same prompt, temperature, and token budget, ensuring that decoding hyper-parameters cannot confound attribution. Given an input x, the generator outputs N natural-language captions describing it. Due to the inherent stochasticity of LLMs, some captions remain faithful while others introduce speculative content, which later enables the construction of contrasted F/H pairs by the discriminator. This design guarantees that both F and H captions originate from an identical generation process, providing a controlled entry point for subsequent evaluation.

Caption Discriminator. Given the candidate captions, the Caption Discriminator determines whether each caption is faithful (C_F) or hallucinatory (C_H). Since hallucination detection is inherently noisy, we adopt an ensemble of multiple detectors, each providing an independent judgment. Final labels are decided via majority voting, which significantly improves robustness under noisy or imperfect classifiers. We further verify the reliability of the hallucination discriminator through dedicated control experiments (Section 4.2). Detailed implementation specifics of the individual detectors and ensemble configuration are provided in Appendix S7 for full clarity and reproducibility.

Task Solver. The solver is not a novel model but a controlled interface to isolate the effect of captions on downstream predictions under identical model and decoding settings. We measure the isolated impact of caption faithfulness by contrasting predictions under three conditions $C \in \{\textsc{raw}, \boldsymbol{C}_F, \boldsymbol{C}_H\}$. We use a unified prompt builder that concatenates three parts:

$$\Phi(\mathcal{I}_{\text{task}}, x, C) = \textsc{Instruction}\, \mathcal{I}_{\text{task}} \parallel \textsc{Serialized Input}\, \sigma(x) \parallel \textsc{Seed}\, s(C) \quad (3)$$

HIVE frames hallucination as a controlled semantic intervention, enabling systematic study of how speculative cues influence downstream reasoning in multimodal models. Prompt templates are listed in Appendix S2.

4 Experiments

4.1 When Does PHR Help?

We begin by characterizing how hallucinated semantics affect downstream predictions across a diverse set of models and tasks. Using the HIVE evaluation infrastructure, we systematically analyze PHR across 9 tasks and 9 models spanning both textual and multimodal scenarios. Each model is evaluated under two input conditions: faithful and hallucinated. Table 1 reports results on text-only benchmarks, while Table 2 summarizes vision–language tasks. Further dataset and model details are provided in Appendix S6.

❶ **LLMs on text-only tasks: limited or unstable benefits.** On textual tasks, hallucinations provide little or no benefit. Across multiple LLM families, performance under hallucinated inputs is often similar to or lower than the faithful setting, with only occasional isolated improvements (see Table 1). The overall trends are small and inconsistent, indicating that hallucinated semantics are hard to exploit in rule-based or symbolically constrained tasks.

Table 1: Faithful (F) vs. Hallucinated (H) path accuracy on text-only tasks. Cells show accuracy (%). $\Delta(H - F)$ denotes the percentage-point difference between the hallucinated and faithful paths (↑ gain, ↓ drop).

Dataset	P.	GPT-4o	GPT-3.5	Claude-3 Sonnet	DeepSeek v3	Mistral Large	O3	DeepSeek R1
AntiCP2	F	54.59 (–)	43.63 (–)	46.84 (–)	52.19 (–)	56.69 (–)	53.95 (–)	49.87 (–)
	H	58.35 ↑+3.76	47.76 ↑+4.13	48.21 ↑+1.37	45.01 ↓-7.18	57.84 ↑+1.15	50.17 ↓-3.78	46.42 ↓-3.45
BBBP	F	61.67 (–)	60.75 (–)	64.07 (–)	61.60 (–)	59.41 (–)	73.27 (–)	70.53 (–)
	H	68.33 ↑+6.66	57.53 ↓-3.22	64.95 ↑+0.88	55.56 ↓-6.04	59.65 ↑+0.24	59.47 ↓-13.80	58.88 ↓-11.65
CodeXGLUE	F	55.15 (–)	58.90 (–)	49.25 (–)	49.46 (–)	50.11 (–)	45.10 (–)	51.54 (–)
	H	52.75 ↓-2.40	57.40 ↓-1.50	53.40 ↑+4.15	50.13 ↑+0.67	56.40 ↑+6.29	46.06 ↑+0.96	48.95 ↓-2.59
SARA v3	F	62.93 (–)	52.28 (–)	62.07 (–)	58.10 (–)	59.14 (–)	65.52 (–)	58.62 (–)
	H	62.24 ↓-0.69	54.83 ↑+2.55	63.97 ↑+1.90	58.96 ↑+0.86	60.34 ↑+1.20	62.07 ↓-3.45	54.48 ↓-4.14
ProofWriter	F	69.49 (–)	66.55 (–)	76.03 (–)	85.17 (–)	70.86 (–)	97.76 (–)	89.31 (–)
	H	75.00 ↑+5.51	66.21 ↓-0.34	76.73 ↑+0.70	85.86 ↑+0.69	68.62 ↓-2.24	98.45 ↑+0.69	92.93 ↑+3.62

❷ VLMs on vision–language tasks: strong and consistent gains. On vision–language benchmarks, hallucinated inputs lead to clear and substantial improvements across models (see Table 2). GPT-4o, Gemini-2.0-Flash, and Qwen-VL-Max generally show gains, with occasional drops on specific datasets such as GQA in certain settings. Unlike the text-only setting, where hallucination effects are small and unstable, VLMs reliably benefit from hallucinated semantics across multiple tasks and model families.

These results suggest that hallucinated captions can enrich visual grounding by introducing additional perceptual cues that support downstream reasoning. Vision–language tasks often involve partial observability, where important cues may be missing or ambiguous. In such settings, hallucinated captions can serve as speculative hypotheses that expand the model's hypothesis space and provide additional semantic anchors for reasoning.

Model-wise variation. While this overall pattern holds across VLMs, the magnitude of PHR gains varies by model. For example, GPT-4o exhibits moderate but stable improvements, whereas Gemini-2.0-Flash and Qwen-VL-Max often show larger gains on perception-heavy tasks. This variation likely reflects differences in visual grounding strength and reasoning strategies across models: models with stronger perceptual grounding can better exploit speculative semantic cues introduced by hallucinated captions, while others benefit less.

4.2 Validating the Hallucination Discriminator

Reliability on hallucination benchmarks. To assess the reliability of HIVE's hallucination discriminator, we evaluate it in two complementary settings. First, we test the discriminator on the TruthfulQA benchmark, which is widely used to probe hallucination behavior in language models. The discriminator achieves an accuracy of **81.76%** on this benchmark. To approximate real-world cross-domain usage, we construct a curated dataset of 180 captions by sampling 20 captions from each of nine tasks. These captions are manually

Table 2: Faithful (F) vs. Hallucinated (H) accuracy on vision-language tasks. Cells show mean accuracy (%). Δ denotes $H - F$ in percentage points and is shown inline at bottom-right.

Dataset	P.	GPT-4o	Claude 3 Sonnet	Gemini 2.0 Flash	Qwen VL-Max
GQA	F	71.36$_{(-)}$	62.02$_{(-)}$	75.23$_{(-)}$	69.88$_{(-)}$
	H	74.22$_{\uparrow+2.86}$	61.03$_{\downarrow-0.99}$	75.69$_{\uparrow+0.46}$	67.90$_{\downarrow-1.98}$
Dex-Net	F	53.25$_{(-)}$	50.29$_{(-)}$	49.88$_{(-)}$	51.54$_{(-)}$
	H	55.76$_{\uparrow+2.51}$	50.71$_{\uparrow+0.42}$	51.15$_{\uparrow+1.27}$	54.12$_{\uparrow+2.58}$
ISIC	F	63.88$_{(-)}$	54.19$_{(-)}$	67.23$_{(-)}$	58.71$_{(-)}$
	H	75.64$_{\uparrow+11.76}$	61.02$_{\uparrow+6.83}$	67.90$_{\uparrow+0.67}$	75.62$_{\uparrow+16.91}$
PlantVillage	F	62.73$_{(-)}$	55.28$_{(-)}$	62.71$_{(-)}$	67.84$_{(-)}$
	H	77.41$_{\uparrow+14.68}$	72.50$_{\uparrow+17.22}$	70.98$_{\uparrow+8.27}$	77.66$_{\uparrow+9.82}$

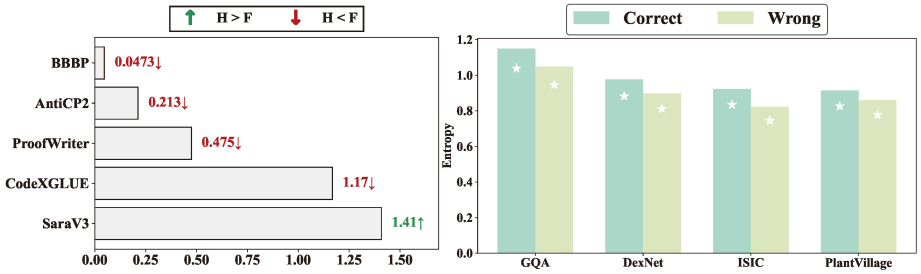

Fig. 2: Process- and output-level analysis of hallucination effects. Left: reasoning chain embeddings show that hallucinated prompts reshape inference trajectories with task-dependent differences ($p < 0.05$), diversifying reasoning. **Right**: caption entropy analysis shows that correct predictions exhibit higher entropy than incorrect ones, indicating expanded semantic coverage. Stars indicate $p < 0.05$.

annotated as either *hallucinated* or *faithful*. Evaluating the discriminator on this human-labeled dataset yields an accuracy of **83.72%**. These results indicate that the discriminator generalizes beyond a single benchmark and can reliably separate hallucinated from faithful captions, providing a stable basis for the paired comparisons used throughout our experiments.

Random-checker control. To rule out improvements arising from arbitrary filtering, we replace the hallucination discriminator with a *random* checker that accepts captions without factual assessment. All decoding controls remain fixed to ensure a fair comparison. As shown in Table 3, random filtering yields only negligible performance changes (0.17–1.23%) and no statistically significant gains across datasets (all $p > 0.14$). This result indicates that the improvements reported in Section 4.1 are unlikely to arise from chance filtering effects.

Prompt robustness. To rule out the possibility that the observed gains arise from prompt sensitivity, we repeat the experiments using semantically equivalent prompt paraphrases. Specifically, we evaluate three prompts: the original prompt (P0) and two paraphrased variants (P1 and P2). The results are summarized in Table 4. Across datasets and models, hallucination-augmented

Table 3: Random checker ablation. Δ denotes the absolute accuracy difference (H–F). p from two-sided paired t-tests; (n.s.) = not significant at $p<0.05$.

Dataset	Faithful (F)	H + Random (H)	Δ (H–F)	p
AntiCP2	0.5015 ± 0.0124	0.5114 ± 0.0200	+0.0100	0.526 (n.s.)
PlantVillage	0.7358 ± 0.0189	0.7481 ± 0.0205	+0.0123	0.354 (n.s.)
Dex-Net	0.4950 ± 0.0068	0.4967 ± 0.0062	+0.0017	0.757 (n.s.)
ISIC	0.5763 ± 0.0067	0.5805 ± 0.0042	+0.0043	0.142 (n.s.)

inputs consistently outperform faithful inputs under all prompt variants. Although absolute accuracies vary slightly due to known prompt sensitivity in LLMs, the direction of the effect remains stable. This suggests that the observed improvements are not tied to a specific prompt formulation.

Table 4: Prompt robustness across paraphrased prompts. P0 denotes the original prompt, while P1 and P2 are semantically equivalent paraphrases. Δ reports the accuracy improvement from hallucination injection (H–F).

Dataset / Model	P0 Δ (%)	P1 Δ (%)	P2 Δ (%)	Consistency
PlantVillage / GPT-4o	+14.68	+28.00	+9.89	✓
PlantVillage / Claude 3 Sonnet	+17.22	+16.42	+12.84	✓
ISIC / GPT-4o	+11.76	+4.55	+5.00	✓
ISIC / Claude 3 Sonnet	+6.83	+4.54	+4.77	✓

Token-level ablation. We conduct a token-level ablation study to test whether hallucinated content is *necessary* for successful predictions. Specifically, we first identify the subset of samples that are solved only when hallucinated captions are provided, where the hallucinated path succeeds while the faithful path fails. Within these captions, we locate hallucinated tokens that serve as *core evidence* in the model's reasoning process. We then mask these tokens using a neutral placeholder and re-evaluate the model on the same samples. As reported in Table 5, we evaluate a subset of samples where the hallucinated path succeeds while the faithful path fails. For this subset, *H-before* is near-perfect by construction, and post-ablation accuracy (*H-after*) drops substantially across all four datasets after masking hallucinated evidence tokens. This result indicates that key hallucinated tokens act as informative cues rather than redundant noise, and that the model relies on them to reach correct predictions.

4.3 Reasoning Convergence Analysis

To assess whether the semantic shifts introduced by hallucinated captions affect the stability of model inference, we analyze reasoning convergence at two complementary levels: (I) **Intra-chain convergence**, which measures whether intermediate reasoning steps progressively align with the final conclusion under hallucinated inputs (Fig. 3 left); and (II) **Inter-chain consistency**, which evaluates whether multiple reasoning paths sampled from the same input converge to semantically similar trajectories across different sampling seeds (Fig. 3

Table 5: Accuracy after hallucination removal across domains. H-after Acc. denotes post-ablation accuracy after masking hallucinated evidence tokens.

Dataset	Domain	Task	H-after Acc.
AntiCP2	Biomedicine	Peptide cls.	0.244 ± 0.085
PlantVillage	Agriculture	Disease recog.	0.700 ± 0.111
Dex-Net	Robotics	Grasping pred.	0.364 ± 0.061
ISIC	Dermatology	Lesion diagnosis	0.380 ± 0.062

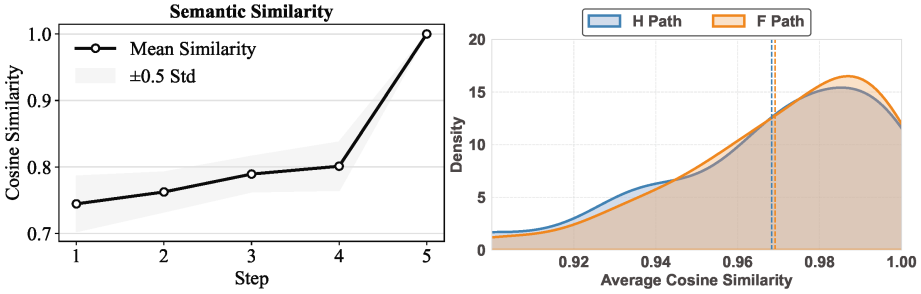

Fig. 3: Inter-chain stability on the PlantVillage dataset. Left: Step-wise cosine similarity shows reasoning chains progressively converge during inference. **Right**: Hallucinated (H) and faithful (F) captions exhibit highly overlapping similarity distributions, indicating that hallucinations preserve stability across sampling runs.

right). These analyses provide a fine-grained view of how hallucinations affect convergence both within and across reasoning chains, enabling us to determine whether they destabilize or preserve the model's inference process. Further implementation details are provided in Appendix S9.

❸ **Hallucinations do not disrupt intra-chain convergence.** From Fig. 3 (left), step-to-final semantic similarity steadily increases, with intermediate steps progressively aligning with the final conclusion. This trajectory indicates that reasoning chains naturally converge as inference unfolds rather than drifting from the target answer. Moreover, the narrowing variance band suggests that this convergence pattern remains stable across runs and datasets.

❹ **Reasoning chains exhibit strong inter-chain consistency.** From Fig. 3 (right), we observe that reasoning paths generated under hallucinated (H) and faithful (F) captions both achieve very high pairwise similarity across multiple sampling runs (means ≈ 0.97). The two distributions nearly overlap, and statistical tests confirm no significant difference between them ($p > 0.6$). This indicates that, regardless of whether captions contain hallucinations, the model converges to consistent reasoning trajectories across chains. Rather than diverging into unstable alternatives, multiple sampled paths remain semantically aligned, underscoring the robustness of the model's inference.

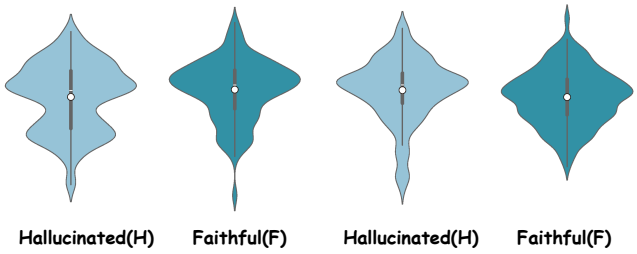

Fig. 4: Distribution of caption embeddings (Faithful (F) vs. Hallucinated (H)). Hallucinated inputs exhibit wider semantic spread and longer tails.

Takeaway ❶. Reasoning chains with hallucinated captions remain stable: intermediate steps consistently converge to the final answer and multiple paths sampled stay highly aligned. This stability highlights that hallucinations can support reliable and reproducible inference.

4.4 Why PHR Helps

To explain the strong and consistent improvements observed on vision–language tasks (see Section 4.1), we analyze why hallucinated captions provide useful semantic signals for VLMs. Our analysis focuses on the beneficial case and examines hallucination effects at three complementary levels: **(I) Input-level shifts**. Hallucinated captions substantially reshape the semantic inputs provided to the model, exhibiting broader distributional spread and lower similarity to faithful captions (Fig. 4). This indicates that hallucinations introduce meaningful semantic variation rather than redundant noise, enriching the model's visual grounding. **(II) Process-level modulation**. Reasoning-chain entropy analysis (Fig. 2 left) shows that hallucinations modulate inference dynamics in a task-dependent manner. They reduce trajectory entropy on reasoning-heavy tasks (promoting convergence and stability), while increasing entropy on structurally open-ended tasks (supporting exploration), indicating active reshaping of the model's reasoning process. **(III) Output-level diversity**. Correct predictions consistently exhibit higher caption entropy than incorrect ones (Fig. 2 right), suggesting that broader semantic coverage induced by hallucinations is positively associated with successful reasoning. Rather than mere lexical variety, this diversity reflects expanded semantic grounding that supports downstream task performance. Further implementation details of similarity computation, entropy estimation, and statistical testing are provided in Appendix S8. We give the following observations:

❺ Hallucinations reshape semantic inputs. From Fig. 4, we observe that hallucinated captions differ systematically from faithful ones in both mean similarity and distributional spread. Hallucinated inputs exhibit wider variance and heavier tails in the embedding space, a difference that is statistically significant under paired t-tests ($p < 0.01$). This confirms that they introduce genuine

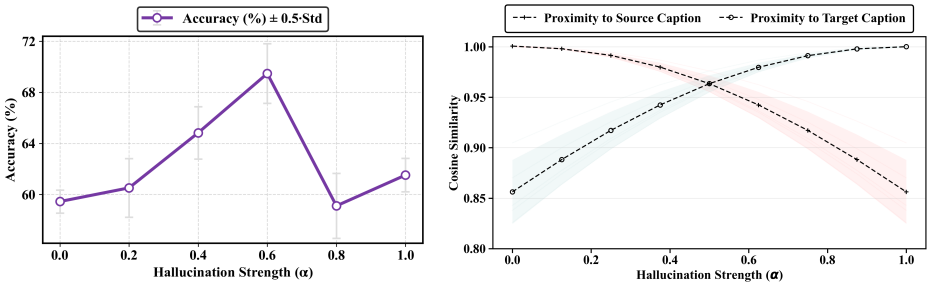

Fig. 5: Downstream task accuracy as a function of hallucination strength. We gradually interpolate between faithful and hallucinated captions and evaluate downstream performance. The results exhibit an inverted-U pattern: introducing moderate hallucination improves accuracy, while excessive hallucination reduces it.

input-level shifts rather than acting as redundant noise, providing models with additional anchors to explore alternative reasoning paths. Importantly, this shift is consistently observed across multiple datasets, underscoring that input-level semantic reshaping is a general property of hallucinations rather than a dataset-specific artifact, and holds robustly across diverse modalities and domains.

❻ **Hallucinations modulate reasoning dynamics.** As shown in Fig. 2 (left), hallucinated prompts alter the entropy of reasoning trajectories in a task-dependent manner. On reasoning-heavy tasks such as BBBP, AntiCP2, and ProofWriter, hallucinations *reduce* movement entropy, suggesting that they promote more convergent and stable inference. Conversely, on structurally open-ended tasks such as SARA v3, hallucinations *increase* entropy, enabling the model to explore a broader range of reasoning paths. These differences are statistically significant (paired t-tests, $p < 0.05$), indicating that hallucinations do not merely inject noise but actively reshape inference dynamics in ways that can either encourage convergence or support exploration, depending on the task.

❼ **Correct predictions align with higher caption entropy.** As shown in Fig. 2 (right), hallucinated captions that lead to correct predictions consistently exhibit higher semantic entropy than those leading to incorrect predictions, across datasets such as GQA, Dex-Net, ISIC, and PlantVillage. These differences are statistically significant ($p < 0.05$), and the effect holds consistently across all evaluated datasets, underscoring that higher semantic diversity is a general marker of successful reasoning rather than a dataset-specific artifact. Rather than mere lexical variety, this result highlights that semantic diversity in the latent space is a useful signal that supports accurate task performance.

Takeaway ❷. Hallucinations consistently reshape inputs, modulate reasoning trajectories, and higher semantic diversity correlates with correct outcomes, indicating that their utility arises from broadening the semantic space rather than adding redundant noise.

We further analyze hallucinations that arise within the reasoning chain itself. By detecting the position of the first hallucinated token (early, middle, late), we observe consistent positional patterns: early hallucinations tend to correlate with stronger downstream improvements. This suggests that speculative cues introduced early in the reasoning process may shape the evolving inference trajectory. Detailed experimental results are provided in Appendix S4.

Intuitively, hallucinated captions expand the semantic hypothesis space available to the model. Under partial observability, faithful captions often describe only visible attributes, which may provide insufficient cues for reasoning. Hallucinated cues introduce speculative but task-relevant semantic anchors that guide the model toward plausible interpretations of the input. When these anchors align with the latent structure of the task, they help the model explore useful reasoning trajectories rather than remaining confined to incomplete observations.

4.5 Case Study

Beyond aggregate results, we present a case study to illustrate how hallucinated captions reshape reasoning chains in practice. We select a sample from the ISIC dataset, where the task is to determine whether a skin lesion is benign or malignant. Given the same input, the faithful (F) caption focuses on superficial attributes (e.g., asymmetry, irregular borders), constraining the reasoning to melanoma-like criteria and resulting in an incorrect malignant prediction. In contrast, the hallucinated (H) caption introduces a vascular cue suggestive of a seborrheic keratosis (SK) frame. Although this cue is not strictly faithful to the image, it provides a task-aligned semantic trigger that reshapes the reasoning trajectory: intermediate steps increasingly anchor the SK frame, ultimately leading to the correct benign diagnosis. This example demonstrates how hallucinations can supply additional semantic anchors that steer the reasoning process toward more effective diagnostic paths, rather than merely injecting noise. As illustrated in Fig. 6, hallucinated captions can introduce auxiliary but task-relevant cues that guide the reasoning chain toward the correct outcome. More case studies across additional datasets are provided in Appendix S5.

4.6 Ablation Study

To identify the conditions under which hallucinations provide the strongest utility, we conduct a set of ablation studies examining three key generation factors: sampling temperature, token budget, and hallucination intensity. These analyses reveal how different generation configurations shape the semantic expansion introduced by hallucinated captions and how such expansion translates into downstream performance gains.

Temperature. We analyze how sampling temperature influences the usefulness of hallucinated captions. As shown in Table 6, all four datasets peak at $T = 0.6$, yielding the strongest and most consistent gains (e.g., $+11.76\%$ on ISIC, $+14.68\%$ on PlantVillage, $+2.51\%$ on Dex-Net, $+3.76\%$ on AnticP2). Lower temperatures ($T = 0.0/0.3$) generate conservative captions with limited semantic expansion (e.g., -4.27% on AnticP2, -5.05% on ISIC), while higher temperature ($T = 0.9$) increases variance and reduces controllability (e.g.,

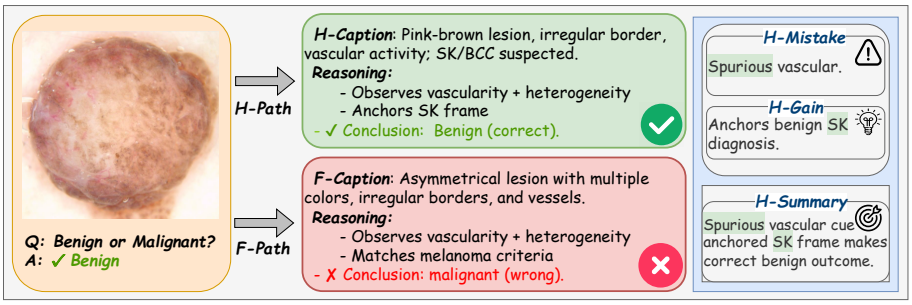

Fig. 6: Case study on ISIC. A hallucinated (H) caption introduces a spurious vascular cue that anchors the reasoning toward a seborrheic keratosis (SK) frame, ultimately yielding the correct benign diagnosis. In contrast, the faithful (F) caption confines reasoning to superficial features, leading to a malignant misclassification, highlighting hallucination's potential as constructive guidance.

−5.00% on AntiCP2). These observations indicate that a moderate temperature around $T = 0.6$ achieves the most effective balance, providing useful semantic enrichment without introducing excessive variability.

Table 6: Hallucination-induced gain (Δ) across temperature and token conditions. We report relative gain $\Delta = H - F$, isolating hallucination effects by eliminating baseline accuracy differences across datasets. **Bold** values mark the strongest gain and underlined values the second-best.

Dataset	Temperature (T)				Token Length			
	0.0	0.3	0.6	0.9	128	256	512	1024
AntiCP2	-4.27	+0.10	**+3.76**	-5.00	+0.15	+3.76	**+4.48**	-0.14
PlantVillage	+2.30	-4.46	**+14.68**	+1.51	-4.66	**+14.68**	+7.86	+9.49
Dex-Net	+0.07	+1.88	**+2.51**	+0.14	+1.56	**+2.51**	-2.04	+1.93
ISIC	+9.26	-5.05	**+11.76**	+3.70	-2.10	**+11.76**	+1.33	-0.48

Maximum token budget. We examine the impact of token length using temperature $T = 0.6$ (Table 6). Short generations (128 tokens) limit semantic coverage and result in weak or inconsistent gains. A budget of 256 tokens produces strong and stable improvements across all datasets. Larger budgets (512 or 1024 tokens) can yield higher peaks but also introduce greater variance. Ablating key hallucinated tokens results in a substantial accuracy drop, confirming the benefits of hallucinations rely on sufficient and well-structured semantic cues.

Hallucination intensity. To assess how hallucination strength affects downstream performance, we generate captions with different hallucination levels using GPT-4o and group them into strong and weak categories. We interpolate between faithful and hallucinated captions, and between strong and weak hallucinations, then re-project them into SBERT space for alignment. Fig. 5 shows smooth semantic transitions (right) and an inverted U-shaped pattern (left),

where moderate hallucination intensity provides the most reliable accuracy improvements, while excessive intensity reduces controllability.

Takeaway ❸. Moderate hallucination levels provide the strongest gains, indicating that the benefits of hallucinated cues depend on controlled semantic expansion rather than excessive or weak generation.

5 Discussion and Conclusion

Our findings point to a dual characterization of post-hallucination reasoning. The *faithful path* encourages exploitation, grounding the model in verified evidence and producing precise but narrow predictions. The *hallucinated path* promotes exploration, expanding the hypothesis space through speculative but task-relevant cues that occasionally unlock shortcuts unavailable to faithful inputs. Hallucinations are thus not merely errors but alternative signals that broaden the model's inference landscape. Their benefits, however, depend critically on control: moderate hallucination intensity enriches semantics without destabilizing inference, whereas excessive or misaligned hallucinations degrade reliability. Temperature, token budget, and interpolation strength provide practical levers for tuning this balance across datasets, models, and decoding settings, and fallback mechanisms to the faithful path help manage risk.

Limitations. Although our study reveals consistent patterns of PHR across tasks and models, several limitations remain. First, our evaluation is conducted on a fixed set of benchmarks and may not fully capture the diversity of real-world multimodal reasoning scenarios. Second, our interventions operate primarily at the caption level, whereas hallucinations may also arise within intermediate reasoning steps or latent representations. Third, while hallucinated cues can sometimes assist reasoning under partial observability, they may also introduce spurious signals in other settings. Future work should investigate broader task domains and develop principled mechanisms to better control hallucination induced semantic expansion in multimodal reasoning systems.

Broader Impact. This work contributes to a deeper understanding of hallucination behavior in multimodal models by examining how hallucinated semantics interact with downstream reasoning. Importantly, our findings should not be interpreted as advocating hallucination as a deliberate inference strategy. Rather, PHR appears to arise as a by-product of reasoning under incomplete evidence. Understanding this phenomenon may inform future methods that balance exploratory semantic cues with reliable grounding and verification.

We hope this work encourages the community to further investigate the role of hallucinated semantics in model reasoning and develop principled methods to understand and control PHR.

Acknowledgements

We thank the reviewers and area chair for their constructive feedback.

References

1. Bendahman, N., Pinel-Sauvagnat, K., Hubert, G., Billami, M.B.: Not all hallucinations are good to throw away when it comes to legal abstractive summarization. In: NAACL (2025) 3
2. Chen, Z., Min, Y., Zhang, J., Yan, B., Wang, J., Wang, X., Shan, S.: A survey of multimodal hallucination evaluation and detection. IJCV **134**(3), 131 (2026) 4
3. Choi, S., Fang, T., Wang, Z., Song, Y.: Kcts: Knowledge-constrained tree search decoding with token-level hallucination detection. In: EMNLP (2023) 3
4. Creswell, A., Shanahan, M.: Faithful reasoning using large language models. arXiv preprint arXiv:2208.14271 (2022) 2
5. Dong, B., Ni, M., Huang, Z., Yang, G., Zuo, W., Zhang, L.: Mirage: Assessing hallucination in multimodal reasoning chains of mllm. In: NeurIPS. vol. 38, pp. 122910–122955 (2026) 4
6. Du, X., Xiao, C., Li, Y.: Haloscope: harnessing unlabeled llm generations for hallucination detection. In: NeurIPS (2024) 3
7. Gong, X., Ming, T., Wang, X., Wei, Z.: Damro: Dive into the attention mechanism of lvlm to reduce object hallucination. In: EMNLP. pp. 7696–7712 (2024) 1
8. Guha, N., Nyarko, J., Ho, D.E., Ré, C., Chilton, A., Narayana, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D.N., et al.: Legalbench: a collaboratively built benchmark for measuring legal reasoning in large language models. In: NeurIPS (2023) 3
9. Han, G., Lim, S.N.: Few-shot object detection with foundation models. In: CVPR. pp. 28608–28618 (2024) 1
10. Han, J., Ren, Y., Ding, J., Yan, K., Xia, G.S.: Few-shot object detection via variational feature aggregation. In: AAAI. vol. 37, pp. 755–763 (2023) 1
11. Hong, Y., Wu, Q., Qi, Y., Rodriguez-Opazo, C., Gould, S.: Vln bert: A recurrent vision-and-language bert for navigation. In: CVPR. pp. 1643–1653 (2021) 1
12. Honovich, O., Choshen, L., Aharoni, R., Neeman, E., Szpektor, I., Abend, O.: Q2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In: EMNLP (2021) 3
13. Hu, X., Ru, D., Qiu, L., Guo, Q., Zhang, T., Xu, Y., Luo, Y., Liu, P., Zhang, Y., Zhang, Z.: Knowledge-centric hallucination detection. In: EMNLP. pp. 6953–6975 (2024) 2
14. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P.: Survey of hallucination in natural language generation. ACM computing surveys **55**(12), 1–38 (2023) 1
15. Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., et al.: Language models (mostly) know what they know. arXiv preprint arXiv:2207.05221 (2022) 4
16. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NeurIPS (2012) 1
17. Kryściński, W., McCann, B., Xiong, C., Socher, R.: Evaluating the factual consistency of abstractive text summarization. In: EMNLP (2020) 3
18. Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J., et al.: Measuring faithfulness in chain-of-thought reasoning. arXiv preprint arXiv:2307.13702 (2023) 4
19. Lee, Y.L., Tsai, Y.H., Chiu, W.C.: Delve into visual contrastive decoding for hallucination mitigation of large vision-language models. arXiv preprint arXiv:2412.06775 (2024) 3

20. Lehman, E., Jain, S., Pichotta, K., Goldberg, Y., Wallace, B.C.: Does bert pre-trained on clinical notes reveal sensitive data? In: NAACL (2021) 3
21. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. In: NeurIPS (2020) 4
22. Li, J., Cao, P., Chen, Y., Xu, J., Li, H., Jiang, X., Liu, K., Zhao, J.: Towards better chain-of-thought: A reflection on effectiveness and faithfulness. In: ACL. pp. 10747–10765 (2025) 2
23. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: EMNLP (2023) 1, 2
24. Li, Z., Wu, X., Du, H., Liu, F., Nghiem, H., Shi, G.: A survey of state of the art large vision language models: Benchmark evaluations and challenges. In: CVPR. pp. 1587–1606 (2025) 1
25. Lin, S., Hilton, J., Evans, O.: Truthfulqa: Measuring how models mimic human falsehoods. In: ACL (2022) 3
26. Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., Wang, L.: Mitigating hallucination in large multi-modal models via robust instruction tuning. In: ICLR (2024) 3
27. Liu, S., Halder, K., Qi, Z., Xiao, W., Pappas, N., Htut, P.M., John, N.A., Benajiba, Y., Roth, D.: Towards long context hallucination detection. In: NAACL. pp. 7827–7835 (2025) 2
28. Luo, J., Li, T., Wu, D., Jenkin, M., Liu, S., Dudek, G.: Hallucination detection and hallucination mitigation: An investigation. arXiv preprint arXiv:2401.08358 (2024) 2
29. Lyu, Q., Havaldar, S., Stein, A., Zhang, L., Rao, D., Wong, E., Apidianaki, M., Callison-Burch, C.: Faithful chain-of-thought reasoning. In: IJCNLP. pp. 305–329 (2023) 2
30. Ma, J., Suo, W., Wang, P., Zhang, Y.: Understanding and mitigating hallucinations in multimodal chain-of-thought models. In: CVPR. pp. 40224–40234 (2026) 4
31. Ma, Y.J., Hejna, J., Fu, C., Shah, D., Liang, J., Xu, Z., Kirmani, S., Xu, P., Driess, D., Xiao, T., et al.: Vision language models are in-context value learners. In: ICLR (2024) 1
32. Manakul, P., Liusie, A., Gales, M.: Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In: EMNLP (2023) 2, 3
33. Manevich, A., Tsarfaty, R.: Mitigating hallucinations in large vision-language models (lvlms) via language-contrastive decoding (lcd). In: ACL (2024) 4
34. Mudgal, S., Lee, J., Ganapathy, H., Li, Y., Wang, T., Huang, Y., Chen, Z., Cheng, H.T., Collins, M., Strohman, T., et al.: Controlled decoding from language models. In: ICML (2024) 3
35. Nori, H., King, N., McKinney, S.M., Carignan, D., Horvitz, E.: Capabilities of gpt-4 on medical challenge problems. arXiv preprint arXiv:2303.13375 (2023) 3
36. Parcalabescu, L., Frank, A.: On measuring faithfulness or self-consistency of natural language explanations. In: ACL (2024) 2
37. Peng, S., Yang, S., Jiang, L., Tian, Z.: Mitigating object hallucinations via sentence-level early intervention. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 635–646 (2025) 2
38. Qiu, Y., Ziser, Y., Korhonen, A., Ponti, E., Cohen, S.B.: Detecting and mitigating hallucinations in multilingual summarisation. In: EMNLP (2023) 2
39. Rajpurkar, P., Jia, R., Liang, P.: Know what you don't know: Unanswerable questions for squad. In: ACL (2018) 4

40. Rohrbach, A., Hendricks, L.A., Burns, K., Darrell, T., Saenko, K.: Object hallucination in image captioning. In: EMNLP (2018) 1
41. Sadat, M., Zhou, Z., Lange, L., Araki, J., Gundroo, A., Wang, B., Menon, R.R., Parvez, M., Feng, Z.: Delucionqa: Detecting hallucinations in domain-specific question answering. In: EMNLP (2023) 3
42. Sennrich, R., Vamvas, J., Mohammadshahi, A.: Mitigating hallucinations and off-target machine translation with source-contrastive and language-contrastive decoding. In: EACL (2024) 4
43. Su, W., Wang, C., Ai, Q., Hu, Y., Wu, Z., Zhou, Y., Liu, Y.: Unsupervised real-time hallucination detection based on the internal states of large language models. In: ACL (2024) 2, 3
44. Tian, X., Zou, S., Yang, Z., He, M., Waschkowski, F., Wesemann, L., Tu, P.H., Zhang, J.: More thought, less accuracy? on the dual nature of reasoning in vision-language models. In: ICLR (2026), https://openreview.net/forum?id=XpL5eqjCjF 4
45. Wang, A., Cho, K., Lewis, M.: Asking and answering questions to evaluate the factual consistency of summaries. In: ACL (2020) 3
46. Wang, F., Zhou, W., Huang, J.Y., Xu, N., Zhang, S., Poon, H., Chen, M.: mdpo: Conditional preference optimization for multimodal large language models. In: EMNLP. pp. 8078–8088 (2024) 1
47. Wang, X., Pan, J., Ding, L., Biemann, C.: Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In: ACL (2024) 2
48. Wang, X., Zhou, D.: Chain-of-thought reasoning without prompting. In: NeurIPS (2024) 4
49. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al.: Emergent abilities of large language models. TMLR (2022) 4
50. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E.H., Le, Q.V., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models. In: NeurIPS (2022) 2
51. Xu, J., Fei, H., Pan, L., Liu, Q., Lee, M.L., Hsu, W.: Faithful logical reasoning via symbolic chain-of-thought. In: ACL (2024) 2
52. Xu, Z., Liu, C., Wei, Q., Wu, J., Zou, J., Wang, X., Zhou, Y., Liu, S.: More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. In: NeurIPS. vol. 38, pp. 82878–82905 (2026) 4
53. Yin, H., Si, G., Wang, Z.: Clearsight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14625–14634 (2025) 2
54. Yu, Q., Li, J., Wei, L., Pang, L., Ye, W., Qin, B., Tang, S., Tian, Q., Zhuang, Y.: Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In: CVPR (2024) 3
55. Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: Visual commonsense reasoning. In: CVPR (2019) 1
56. Zhang, J., Huang, J., Jin, S., Lu, S.: Vision-language models for vision tasks: A survey. TPAMI **46**(8), 5625–5644 (2024) 1
57. Zhang, J., Wang, T., Zhang, H., Lu, P., Zheng, F.: Reflective instruction tuning: Mitigating hallucinations in large vision-language models. In: ECCV (2024) 3
58. Zhang, T., Qiu, L., Guo, Q., Deng, C., Zhang, Y., Zhang, Z., Zhou, C., Wang, X., Fu, L.: Enhancing uncertainty-based hallucination detection with stronger focus. In: EMNLP. pp. 915–932 (2023) 2

59. Zhang, W., Wang, Y.X.: Hallucination improves few-shot object detection. In: CVPR. pp. 13008–13017 (2021) 1
60. Zhao, Z., Cohen, S.B., Webber, B.: Reducing quantity hallucinations in abstractive summarization. In: EMNLP (2020) 3
61. Zhou, X., Zhang, M., Lee, Z., Ye, W., Zhang, S.: Hademif: Hallucination detection and mitigation in large language models. In: ICLR (2025) 2