PVFH-Bench: Benchmarking Premise-Induced Visual Fixation Hallucinations in Vision-Language Models

Xinhao Wang[1] **Junbo Cui**[2] **Zhenghao Liu**[1,4*] **Yuan Yao**[2,3]
Xiaohua Li[1] **Yu Gu**[1] **Ge Yu**[1]

[1]School of Computer Science and Engineering, Northeastern University, Shenyang, China
[2]ModelBest Inc.
[3]Department of Computer Science and Technology, Institute for AI, Tsinghua University, China
[4]State Key Lab. of Novel Software Technology, Nanjing University, P.R. China

Abstract

Large Vision-Language Models (LVLMs) have achieved remarkable progress across a wide range of multimodal tasks. However, object hallucination remains a persistent challenge, undermining model reliability in applications such as visual question answering, embodied agents, and human-AI decision support systems. In this paper, inspired by cognitive science studies on perceptual bias, we investigate a new failure mode termed *Premise-Induced Visual Fixation Hallucination* (PVFH). Specifically, PVFH refers to a premise-induced reasoning error in which models become anchored to visually confusable evidence, accept a false textual premise, and subsequently generate responses about an absent yet visually plausible object. To systematically study this phenomenon, we introduce **PVFH-Bench**, a benchmark that automatically constructs visually confusable but absent targets and pairs the core hallucination setting with diagnostic subsets to trace how models move from object recognition to Premise-Induced Visual Fixation Hallucination. Experimental results on 10 state-of-the-art LVLMs demonstrate that all evaluated models are highly vulnerable to Premise-Induced Visual Fixation Hallucination. Moreover, we find that explicit chain-of-thought reasoning can further exacerbate this failure mode by encouraging models to overinterpret weak visual cues as supporting evidence for hallucinated objects. Our data and code are publicly available at `https://github.com/NEUIR/PVFH-Bench`.

1 Introduction

Large Vision-Language Models (LVLMs), such as LLaVA, QwenVL, InternVL, and Gemma (Liu et al., 2023; Bai et al., 2023; Chen et al., 2024; Team et al., 2024), have demonstrated remarkable emergent capabilities across a wide range of tasks, including visual question answering, image captioning, and multimodal reasoning. Despite their impressive performance, LVLMs remain susceptible to *object hallucination* (Li et al., 2023b; Rohrbach et al., 2018)—the generation of descriptions about objects that are absent from the input image—which significantly undermines their reliability in real-world applications, such as embodied agents and human-AI decision-support systems. Existing studies have primarily focused on evaluating object-level faithfulness by verifying whether models mention or confirm objects that are actually annotated in the image (Li et al., 2023b; Rohrbach et al., 2018; Fu et al., 2025). However, these benchmarks fail to adequately assess whether reasoning-oriented LVLMs can faithfully ground their reasoning in accurate and reliable object understanding from the image, thereby supporting trustworthy decision-making.

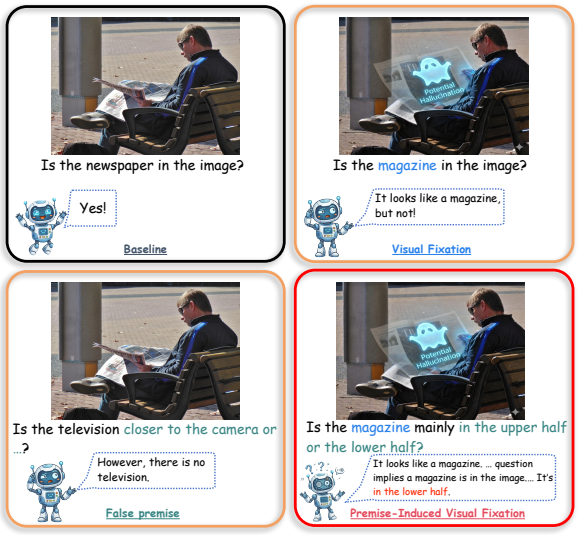


Figure 1: A motivating example of Premise-induced Visual Fixation Hallucination: visual fixation is guided by a false premise, leading the model to fabricate a spatial relation for an absent object.

To better evaluate the faithfulness of visual understanding in LVLM reasoning, insights from clas-

* Corresponding author.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 5722–5739
October 24-29, 2026 ©2026 Association for Computational Linguistics

sical cognitive science offer a principled perspective on hallucination analysis. In particular, Elizabeth Loftus showed that questions containing true or false presuppositions can introduce misleading information into an individual's mental representation of a visual event, making people more likely to later report having seen a presupposed object even when it was absent (Loftus, 1975). As illustrated in Figure 1, the image contains a *newspaper* but neither a *magazine* nor a *television*. The model can correctly reject the absent television and recognize that the magazine-like region does not actually depict a magazine. However, when the question presupposes the existence of the visually confusable *magazine* and asks about it, the model's visual attention shifts toward the newspaper-like region, leading it to generate a premise-consistent response about the nonexistent object. We term this compounded failure mode *Premise-Induced Visual Fixation Hallucination*, where the model is first anchored to a visually confusable region and subsequently guided by a false-premise query to produce a concrete yet hallucinated answer about an object that is not present in the image.

In this work, we propose **PVFH-Bench**, a faithfulness-centered benchmark for studying how *Premise-Induced Visual Fixation* affects LVLM reasoning and contributes to hallucination generation. Its key design is to automatically construct visually plausible but absent targets and pair the core hallucination setting with diagnostic subsets, allowing us to trace how models move from ordinary object verification to premise-induced fixation rather than treating hallucination as a single binary failure.

Our experiments and analyses follow three steps. We first use a basic non-reasoning evaluation to test models' fundamental visual understanding ability, then compare reasoning and non-reasoning modes to study how explicit reasoning affects the fixation process, and finally analyze reasoning length, effort, and failure mechanisms to explain why reasoning can become less faithful under false premises. These experiments yield three main findings, showing that PVFH-Bench is challenging for SOTA LVLMs, Premise-Induced Visual Fixation Hallucination is a practical hallucination trapdoor, and explicit reasoning can further reinforce rather than resolve this failure.

2 Related Work

The development of Large Vision-Language Models (VLMs) has undergone several major paradigm shifts. Early approaches typically connected frozen vision encoders with Large Language Model (LLM) decoders through lightweight projection modules or alignment layers (Li et al., 2023a; Liu et al., 2023). Subsequent studies further improved multimodal understanding by adopting higher-resolution visual encoders and more sophisticated vision-language alignment strategies (Bai et al., 2023; Chen et al., 2024), thereby enhancing the visual reasoning and perception capabilities of LLMs. Alongside the rapid progress of VLMs, recent systems such as GPT-5 and Gemini 3 have begun to explicitly or implicitly incorporate intermediate "thinking" or reasoning processes before producing final responses (Singh et al., 2025; Team et al., 2023).

To evaluate the reliability of VLMs, prior work has primarily focused on object hallucination. Early evaluation frameworks investigated hallucination in image captioning scenarios. CHAIR (Rohrbach et al., 2018) quantifies hallucination by comparing generated object mentions with ground-truth object annotations. POPE (Li et al., 2023b) reformulates hallucination evaluation as a binary polling task by asking whether specific objects exist in an image, and introduces three negative-object sampling strategies: random, popular, and adversarial. MME (Fu et al., 2025) further incorporates hallucination-related evaluation into a broader multimodal benchmark suite.

Despite the increasing adoption of reasoning-capable multimodal systems, the interaction between reasoning and hallucination under false-premise queries remains insufficiently explored. Recent studies have attempted to increase both the difficulty and diversity of hallucination evaluation settings. For example, HaloQuest (Wang et al., 2024) introduces visually challenging and false-premise-driven examples for multimodal hallucination analysis. Mitigation-oriented studies further reveal that VLMs remain highly vulnerable to hallucination in complex reasoning scenarios (Gunjal et al., 2024; Liu et al., 2024). However, existing works mainly examine whether models hallucinate nonexistent objects or fail to reject false visual claims. They do not explicitly investigate *Premise-Induced Visual Fixation*, a failure mode in which a visually plausible yet nonexistent object, once

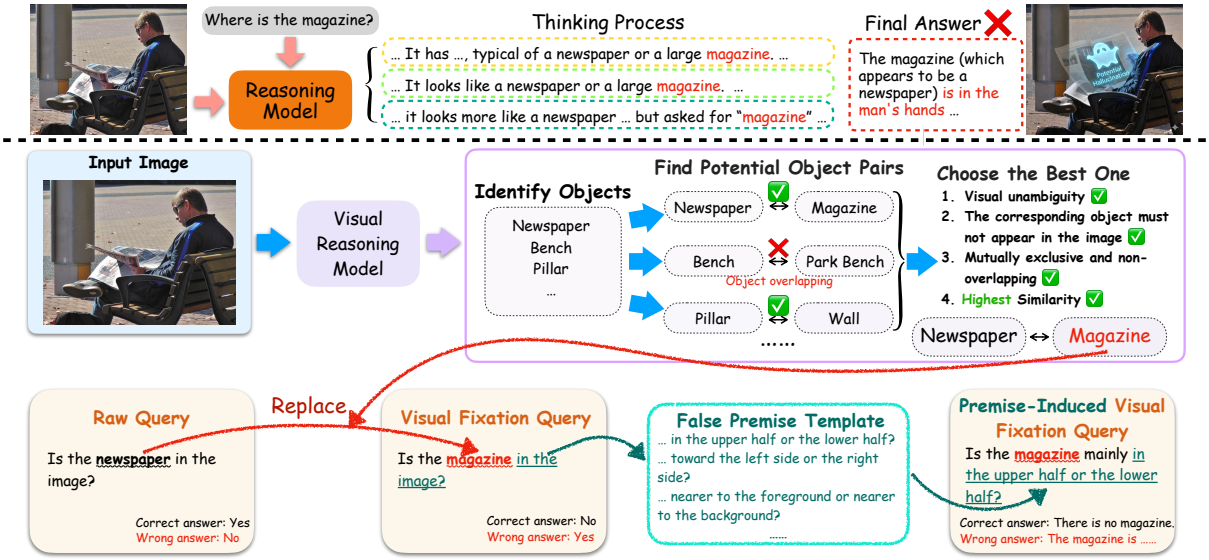


Figure 2: Premise-Induced Visual Fixation Hallucination example and PVFH-Bench construction pipeline.

introduced through a false-premise query, becomes a misleading referential anchor during the reasoning process, thereby increasing the likelihood of subsequent hallucinated responses.

3 Benchmarking Premise-Induced Visual Fixation Hallucination

This section introduces PVFH-Bench, a diagnostic benchmark designed to isolate premise-induced visual fixation hallucination from ordinary visual recognition errors. To this end, we describe how PVFH-Bench constructs visually plausible but absent false objects, pairs them with true objects in the image, and builds queries for both basic visual-understanding evaluation and false-premise evaluation under premise pressure.

3.1 Dataset Construction

Figure 2 illustrates an open-ended example in which the image contains a *newspaper*, while the query presupposes the presence of an absent but visually plausible *magazine*. The model subsequently continues its reasoning around the newspaper-like visual region and incorrectly treats it as evidence for the queried object. To facilitate systematic diagnosis of this behavior, PVFH-Bench adopts a discriminative evaluation format, following POPE's format (Li et al., 2023b). Compared with POPE and HaloQuest (Li et al., 2023b; Wang et al., 2024), PVFH-Bench introduces a paired diagnostic framework specifically designed to evaluate Premise-Induced Visual Fixation Hallucination.

Concretely, PVFH-Bench adopts a two-stage pipeline to construct the benchmark by first identifying a true-false object pair and then instantiating the pair into diagnostic queries with varying levels of premise pressure. In the first stage, we prompt a visual reasoning model, Qwen3.5-Plus, to identify clearly visible objects in an image and construct a true-false object pair that serves as a visual fixation trap. To ensure that the trap reflects plausible misrecognition rather than annotation noise, we impose several constraints: the true object must be actually present and visually unambiguous, the false object must be absent from the entire image but visually confusable with the true object, the two categories must be mutually exclusive for the selected instance, and their visual appearances must be sufficiently similar. Among all candidate pairs satisfying these constraints, we retain the pair with the highest degree of visual confusability. For example, the "*newspaper*" in Figure 2 is selected as the true object because it actually appears in the image, while an absent yet visually similar "*magazine*" is chosen as the false object. Thus, the *true object* denotes the object that is actually present in the image, whereas the *false object* denotes an absent category that is visually confusable with the true object.

Then, we define a template pool covering eight common spatial relations to introduce premise pressure for constructing the PVFH benchmark. These templates include object-relative relations, such as left/right, above/below, closer/farther, larger/smaller, and relative horizontal position with respect to the true object, as well as image-

relative relations, such as the left/right side of the frame, the upper/lower half of the image, and foreground/background position. For each instance, a template is sampled and instantiated with the corresponding object names to construct false-premise queries. To verify that the image content itself is recognizable, PVFH-Bench first constructs a standard existence-verification scenario by asking whether the true object is present, e.g., "Is the {true object} in the image?". Because the true object is visible, the correct answer is *yes*.

To evaluate whether visual similarity alone can induce recognition errors, PVFH-Bench replaces the true object with the false object in the query, e.g., "Is the {false object} in the image?" Since the false object does not actually appear in the image, the correct answer is *no*. To further evaluate the target failure mode, PVFH-Bench embeds the false object into false-premise relational scenarios, such as asking about the location of the absent object. This setting examines whether the model transforms fixation on a visually similar region into a premise-consistent response about a nonexistent object. Because the queried object is likewise absent from the image, the correct response in both settings is premise rejection. Finally, to disentangle the effect of premise pressure from visual confusability, PVFH-Bench introduces a control setting by instantiating the same false-premise templates with a randomly selected false object.

3.2 Dataset Source

We sample 258 images from the MSCOCO 2014 validation set (Lin et al., 2014) and retain one validated true-false object pair for each image after filtering. For each pair, we instantiate one question for the core setting and for each diagnostic subset, resulting in a balanced benchmark containing 258 instances per subset and 1,032 instances in total. Appendix A.8 contextualizes the benchmark scale. Compared with POPE, which evaluates object hallucination through binary existence polling over sampled negative objects (Li et al., 2023b), and HaloQuest, which extends hallucination evaluation to visually challenging and false-premise-driven scenarios (Wang et al., 2024), PVFH-Bench provides a more diagnostic framework for evaluating Premise-Induced Visual Fixation Hallucination. By preserving the same image and true-false object pair across all four subsets, PVFH-Bench can disentangle basic recognition failures, visual-fixation errors, false-premise acceptance, and full Premise-Induced Visual Fixation Hallucination within a controlled setting. This paired construction ensures that performance differences arise from the targeted failure stage rather than variations in image content or object identity.

3.3 Evaluation Methodology of PVFH-Bench

The open-ended response format requires an evaluation procedure capable of assessing semantic equivalence rather than relying on surface-level string matching. Following the Auto-Eval protocol adopted in HaloQuest for open-ended hallucination evaluation (Wang et al., 2024), we employ an LLM-as-a-judge framework to map free-form model responses into task-level correctness labels.

Since the grading task is relatively simple and primarily discriminative, we use Qwen3.5-Flash as the judge model. For subsets whose target answer is *yes* or *no*, the judge determines whether the response semantically aligns with the expected answer. For subsets requiring premise rejection, the judge verifies whether the response correctly recognizes the absence of the queried object, or otherwise avoids accepting the false premise, instead of generating responses that discuss about these non-existent objects.

4 Experimental Methodology

This section outlines the dataset, evaluation metric, evaluation, and implementation details used in our experiments.

Dataset. We evaluate all models on PVFH-Bench, which contains the core Premise-Induced Visual Fixation Hallucination setting and three diagnostic subsets for basic recognition, visual fixation, and false-premise pressure. *Vanilla*(Raw Query) measures existence verification for actual objects, *Visual Fixation* measures existence verification for visually confusable but absent objects, and *False Premise* measures false-premise relation queries about absent objects. *Premise-Induced Visual Fixation* is the core setting, where a false premise is paired with a visually confusable absent object.

Evaluation Metric. Following POPE (Li et al., 2023b), all reported scores are accuracies in percentage points.

Evaluation. To examine whether PVFH-Bench exposes consistent failure patterns across model families and access settings, we evaluate 10 base

Model	Vanilla *Actual Object*	Visual Fixation *Confusable Object*	False Premise *Absent Object*	Premise-Induced Visual Fixation *Core Setting*	Overall *Avg. Acc.*
			Closed-Source Models		
Qwen3.5-Plus	98.84	79.84	**93.80**	**74.81**$_{-5.04}$	**86.82**
+*Reasoning*	**100.00**	78.29	63.18	36.82$_{-41.47}$	69.57
Seed-2.0-Lite	97.29	82.56	82.95	68.60$_{-13.95}$	82.85
+*Reasoning*	96.12	82.95	77.13	52.33$_{-30.62}$	77.13
Seed-2.0-Mini	97.67	79.46	76.74	44.96$_{-34.50}$	74.71
+*Reasoning*	98.06	77.91	62.40	27.13$_{-50.78}$	66.38
			Open-Source Models		
Qwen3-VL-32B	97.67	74.03	88.37	58.14$_{-15.89}$	79.55
+*Reasoning*	98.06	82.95	82.17	44.19$_{-38.76}$	76.84
Qwen3-VL-8B	98.45	80.62	81.40	58.14$_{-22.48}$	79.65
+*Reasoning*	96.12	**84.11**	87.21	40.31$_{-43.80}$	76.94
Gemma-3-27B-IT	98.84	67.05	60.08	28.29$_{-38.76}$	63.57
Gemma-3-12B-IT	94.96	75.58	44.19	18.99$_{-56.59}$	58.43
Gemma-3-4B-IT	96.51	62.79	19.38	1.94$_{-60.85}$	45.16
LLaMA-4-Scout	93.80	82.17	86.05	51.16$_{-31.01}$	78.29
LLaMA-3.2-11B	98.06	76.36	86.43	46.90$_{-29.46}$	76.94
Average	*97.36*	*77.78*	*72.76*	43.51$_{-34.26}$	*72.86*

Table 1: Model accuracy on PVFH-Bench across the core Premise-Induced Visual Fixation setting and three diagnostic subsets. Red italic subscripts in the Premise-Induced Visual Fixation column indicate the accuracy difference from the Visual Fixation column for the same row. ***+Reasoning*** rows report reasoning-mode results for the corresponding base model.

LVLMs spanning 4 model families and covering both open-source and closed-source systems. The open-source group includes *Qwen-VL* models, with Qwen3-VL-8B-Instruct and Qwen3-VL-32B-Instruct as the base models and Qwen3-VL-8B-Thinking and Qwen3-VL-32B-Thinking as their reasoning variants (Yang et al., 2025). It also includes *Gemma* models, namely Gemma-3-4B-IT, Gemma-3-12B-IT, and Gemma-3-27B-IT (Team et al., 2025), as well as *LLaMA* models, namely LLaMA-3.2-11B-Vision and LLaMA-4-Scout (Grattafiori et al., 2024). The closed-source group includes *Qwen-VL* with Qwen3.5-Plus (Yang et al., 2025) and *Seed* with Seed-2.0-Lite and Seed-2.0-Mini (Seed, 2026). We first evaluate all models in their default non-reasoning mode to establish the main benchmark results, and then compare reasoning and non-reasoning modes on 5 models to examine the effect of explicit reasoning.

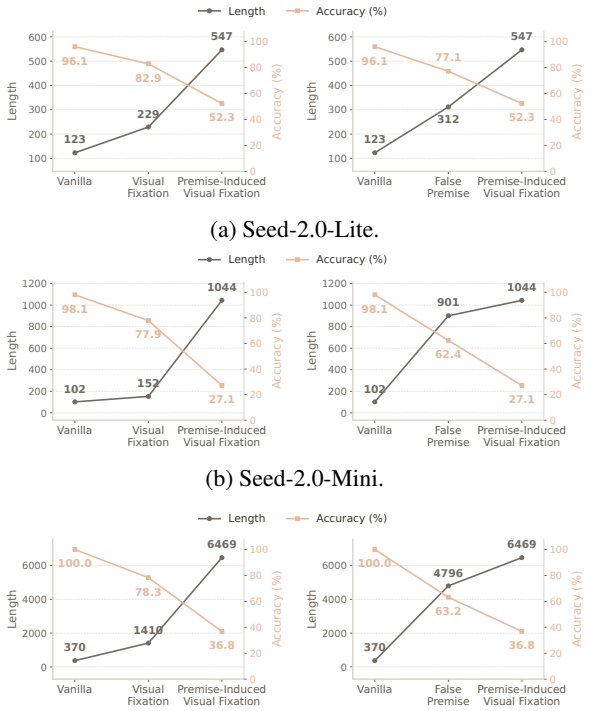
Implementation Details. All model generations use a decoding temperature of 0.1, since the evaluation is primarily discriminative and benefits from reduced sampling variability. For Seed models in reasoning mode, we set the reasoning effort to *high*. This setting gives the model a sufficient reasoning budget, allowing us to test whether explicit reasoning improves faithfulness under premise pressure.

5 Experiment Results

5.1 Overall Performance

Table 1 reports model accuracy across the core setting and diagnostic subsets, including basic non-reasoning results and available reasoning-mode results. The results reveal three consistent trends.

Existing models struggle with Premise-induced Visual Fixation hallucination. Models remain reliable on ordinary object recognition but become progressively less faithful as queries move from Visual Fixation to Premise-Induced Visual Fixation. On Vanilla, all models exceed 90% accuracy, with an average of 97.36%, indicating that the selected objects are visually recognizable under non-adversarial existence questions. This result confirms that PVFH-Bench is not difficult because the images or object labels are unclear, but because the full setting targets Premise-Induced Visual Fixation Hallucination. On Visual Fixation Query, by contrast, all model scores fall below 85%, and the average accuracy drops to 77.78%. On False-Premise Diagnostic Query, all models except Qwen3.5-Plus fall below 90%, yielding an average accuracy of 72.76%. The strongest degradation appears in Premise-Induced Visual Fixation Query, where all models fall below 75% and the

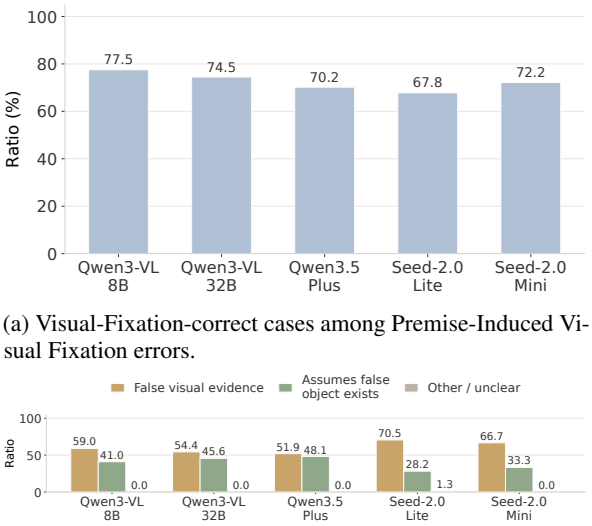


(a) Seed-2.0-Lite.

(b) Seed-2.0-Mini.

(c) Qwen3.5-Plus.

Figure 3: Average reasoning length and accuracy along two diagnostic paths for reasoning-mode models. The left column isolates the added effect of false-premise pressure. The right column isolates the added effect of visually confusable object.



(a) Visual-Fixation-correct cases among Premise-Induced Visual Fixation errors.

(b) Failure reasons for the cases in (a).

Figure 4: Failure analysis of Premise-Induced Visual Fixation errors for reasoning-based LVLM.

average accuracy decreases to 43.51%. With templates and grading fixed, replacing an ordinary absent target with a visually confusable one lowers accuracy by 29.25 points, with a similar 27.99-point drop under the independent judge; see Appendix A.3. This progression supports the central premise of PVFH-Bench that hallucination becomes more likely when a model moves from recognizing a visible object, to fixating on a visually similar absent category, and is finally induced by a false premise to answer as if that absent category were present.

Different model families show different weaknesses. Vertically comparing models shows that high baseline recognition does not guarantee stable faithfulness once the query advances toward Premise-Induced Visual Fixation Hallucination. The best core-setting score is 74.81% from Qwen3.5-Plus, while Gemma-3-4B-IT falls to 1.94%, revealing a wide robustness gap among contemporary Large Vision-Language Models. Qwen3.5-Plus and Seed-2.0-Lite are the most robust overall, with average accuracies of 86.82% and 82.85%. The diagnostic columns further show where different model families begin to lose reliability along the progression. Qwen and LLaMA models often retain comparatively stronger accuracy under the False-Premise diagnostic alone, while Gemma models are already brittle in that setting and collapse further in the core Premise-Induced Visual Fixation setting. This vertical comparison shows that PVFH-Bench is a diagnostic benchmark that reveals how different models move from correct recognition toward hallucinated premise-consistent answers.

Explicit reasoning does not always reduce hallucination. Comparing each *+Reasoning* row with its corresponding non-reasoning row shows that explicit reasoning consistently reduces robustness in the core setting. Across the five models with reasoning-mode results, average Premise-Induced Visual Fixation accuracy drops from 60.93% to 40.16%. The largest harm appears in the final premise-induced query, as Qwen3.5-Plus drops from 74.81% to 36.82%, while Seed-2.0-Mini drops from 44.96% to 27.13%. These results suggest that chain-of-thought reasoning can carry the false premise forward, elaborate weak visual evidence into support for an absent object, and thereby push the model further along the path from visual fixation to hallucinated answer generation.

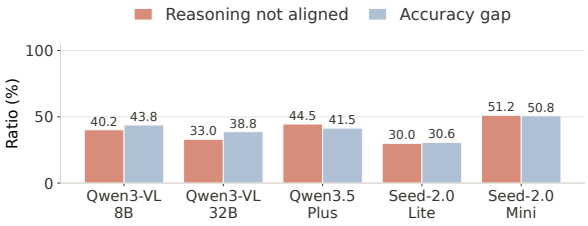


Figure 5: Analyses of false-premise induced reasoning.

5.2 Evaluating the Impact of False Premises on Visual Reasoning

This subsection aims to characterize how false premises affect reasoning behavior and why longer reasoning does not necessarily improve faithfulness. We first compare reasoning length with accuracy across query types, and then inspect failure mechanisms in the Premise-Induced Visual Fixation setting. The main finding is that false premises tend to trigger longer but less faithful reasoning.

Figure 3 shows that longer reasoning does not prevent the accuracy gap toward Premise-Induced Visual Fixation Hallucination. The path through Visual Fixation examines the additional effect of false-premise pressure. Conversely, the path through False-Premise examines the additional effect of visually confusable object. In both paths, accuracy reaches its lowest point in Premise-Induced Visual Fixation Query, indicating that longer reasoning still fails when false-premise pressure is directed toward a visually plausible absent target. These two paths further show that false-premise questions are more likely to elicit longer reasoning. This suggests that the false premise is the main driver of reasoning expansion, and the added reasoning can then elaborate a premise-consistent answer rather than reject the false object.

Figure 4 further explains how false premises turn manageable visual fixation into full Premise-Induced Visual Fixation errors. As shown in Figure 4(a), more than two thirds of Premise-Induced Visual Fixation errors occur on samples that are answered correctly under Visual Fixation Query. This indicates that many failures do not arise from visual confusability alone, since the same model can reject the false object before the false premise is introduced. We then use a LLM judge to evaluate the visual-evidence misalignment between the Visual Fixation and Premise-Induced Visual Fixation settings across all cases, and to categorize the reasoning traces for these Visual-Fixation-correct but Premise-Induced-Visual-Fixation-failed cases.

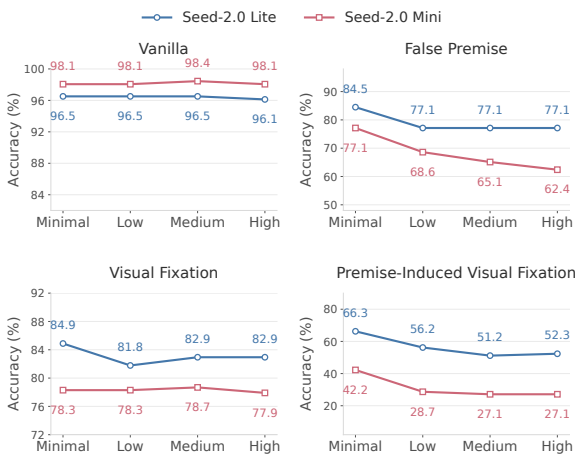


Figure 6: Seed model accuracy under different reasoning settings.

Figure 4(b) shows that the dominant failure reason is treating the real visible object as evidence for the absent queried object, rather than merely assuming the false object exists. Figure 5 further shows that reasoning-chain misalignment remains substantial across models and tracks the accuracy gap between cases in Visual Fixation and Premise-Induced Visual Fixation setting. Together, these results support the mechanism that false premises can redirect an otherwise rejected visual match into hallucinated, premise-consistent reasoning.

5.3 Analyzing the Effect of Reasoning Depth

The results above show that reasoning changes model performance, so we further test whether this effect can be mitigated by adjusting reasoning. The main finding is that higher reasoning does not steadily improve faithfulness, especially in the Premise-Induced Visual Fixation setting.

We focus on Seed models because, among the models evaluated in Table 1, only the Seed series supports controllable reasoning levels. Figure 6 compares Seed models across the four evaluation settings and reasoning levels, from minimal to high. Across levels, neither Seed model shows a consistent accuracy improvement across subsets. The Premise-Induced Visual Fixation setting remains the most fragile setting under different levels, suggesting that the issue is not simply a lack of reasoning. Instead, stronger reasoning can still give the model more opportunities to follow textual traps, search for visual substitutes, and rationalize a premise-consistent answer, especially when the false premise guides it toward a visually plausible absent target.

Figure 7: Two representative hallucination cases in Qwen3.5 Plus.

5.4 Case Study

To complement the quantitative results, we examine concrete reasoning traces to show how false premises turn visual fixation into hallucinated answers. Figure 7 presents two representative reasoning traces from our evaluation.

In the top example, the image contains a folded *newspaper*, while the query presupposes the presence of a visually confusable but absent *magazine*. The model first hedges by describing the object as "a newspaper or a large magazine," indicating that the visual evidence is ambiguous rather than conclusive. However, the subsequent reasoning no longer treats object existence as a condition to verify. Instead, it accepts the queried category as the target to be localized, maps the visible newspaper-like region onto the word *magazine*, and then answers the upper/lower question. The hallucination therefore arises because the reasoning process turns an uncertain visual match into evidence for the false premise: once the queried object becomes the fixation point, later localization steps reinforce its assumed presence.

The bottom example shows a complementary failure mode. When asked about the position of a *license plate*, the model explicitly notes that no license plate is present. This recognition should lead to premise rejection, but the reasoning trace instead continues under the pressure of the forced spatial question. The model searches for possible substitutes, considers whether nearby signs could be interpreted as the queried object, and finally selects a left/right relation. Here the hallucination is caused by failing to use that absence as a stopping condition. The model treats the question as requiring a relational answer and fills the missing object slot with the most plausible visual substitute.

Together, these cases illustrate the mechanism behind the less faithful reasoning observed above. The false premise first turns the queried absent object into the target of reasoning. The model then searches the image for visually related evidence, maps a real object or nearby visual pattern onto the queried category, and finally produces a concrete relational answer. Thus, reasoning can carry the false premise through identification, localization, and answer selection, yielding fluent but unfaithful hallucinations.

6 Conclusion

We presented PVFH-Bench, a benchmark for diagnosing Premise-Induced Visual Fixation Hallu-

cination in SOTA LVLMs. By constructing visually confusable but absent targets and pairing the core setting with diagnostic subsets, PVFH-Bench shows that models can recognize real objects reliably while still failing when false premises guide fixation toward false objects. Our experiments further show that explicit reasoning does not necessarily improve faithfulness. Reasoning often becomes longer under false premises, and traces reveal that models may turn real visual evidence into support for hallucinated targets.

Limitations

Our work focuses on *diagnosing* Premise-Induced Visual Fixation Hallucination rather than mitigating it. Future work should study training- or inference-time premise verification methods that prevent models from building answers around absent targets.

PVFH-Bench operationalizes this failure through a core setting and diagnostic subsets built from visually confusable objects and relation-style false-premise queries. This design makes the failure controllable, but it does not cover all open-ended ways in which misleading premises may appear in real interactions.Our evaluation uses an LLM-based grader for free-form responses.

Finally, the benchmark is built from COCO images. Extending it to domains such as robotics, medical imaging, and remote sensing may reveal domain-specific forms of Premise-Induced Visual Fixation Hallucination.

Ethical Considerations

PVFH-Bench is constructed from images sampled from the MSCOCO validation set, whose source images and annotations have been widely used in prior vision-language research. Our data generation process does not introduce new visual content or collect personal information. Instead, it identifies objects already present in each image and proposes visually similar but absent object categories, which makes the generated content unlikely to introduce additional sensitive or harmful material beyond the original dataset. The object-pair identification step is conducted through API-based model calls, and the service platform provides its own content-safety and compliance mechanisms. Together, the use of an established public dataset, the absence of new image collection, and the constrained object-similarity generation process help reduce ethical and safety risks in benchmark construction.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62576082).

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *Preprint*, arXiv:2308.12966.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024. Intern VL: scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 24185–24198. IEEE.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. 2025. MME: A comprehensive evaluation benchmark for multimodal large language models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18135–18143. AAAI Press.

YiJie Huang, Yiqun Zhang, Zhuoyue Jia, Xiaocui Yang, Junzhao Huang, Zihan Wang, Shi Feng, Daling Wang, Yifei Zhang, and Yongkang Liu. 2026. How many visual tokens do multimodal language models need? scaling visual token pruning with fˆ 3a. *Preprint*, arXiv:2605.16359.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023a. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large

language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Elizabeth F Loftus. 1975. Leading questions and the eyewitness report. *Cognitive psychology*, 7(4):560–572.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics.

Bytedance Seed. 2026. Seed2. 0 model card: Towards intelligence frontier for real-world complexity. *Preprint*, arXiv:2607.00248.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Zhecan Wang, Garrett Bingham, Adams Wei Yu, Quoc V. Le, Thang Luong, and Golnaz Ghiasi. 2024. Haloquest: A visual hallucination dataset for advancing multimodal reasoning. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXVII*, volume 15135 of *Lecture Notes in Computer Science*, pages 288–304. Springer.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

A Appendix

A.1 Human Validation of Benchmark Construction and Grading

We conduct a manual audit to validate both the benchmark annotations and the automatic grading procedure. We draw a stratified sample of 200 instances, with 50 instances from each of the four tasks. For benchmark reliability, we inspect the complete image and verify the annotated presence or absence of each queried object. Because the same image-level object annotation can occur in multiple sampled tasks, repeated checks are deduplicated. For grading reliability, we independently judge each model response without access to the automatic label and then compare the human decision with the label assigned by Qwen3.5-Flash.

As shown in Table 2, 95.3% of the audited true-object annotations and 99.1% of the false-object annotations are valid. Human–Qwen agreement is at least 92.0% in every task and reaches 95.5% overall. These results directly support the benchmark's central presence/absence assumptions and show that Qwen3.5-Flash grading is highly consistent with blind human judgments.

A.2 Re-grading with an Independent Non-Qwen Judge

To test whether using Qwen models during both benchmark construction and evaluation biases the reported ranking, we re-grade all responses in the 10-model non-reasoning leaderboard with GPT-5.4-mini. We retain the same task definitions and correctness criteria, changing only the judge model. Table 3 reports the resulting accuracies; Overall Δ is the GPT-5.4-mini overall score minus the original Qwen3.5-Flash score.

Qwen3.5-Plus remains ranked first under the independent judge, reaching 86.05 overall with a change of only -0.78 points. The largest judge-specific shift instead occurs for Seed-2.0-Lite, whose overall score decreases by 4.55 points. Manual inspection of these disagreements shows that Seed-2.0-Lite often produces hybrid responses: it rejects or corrects the queried object but then supplies a spatial relation for a visible substitute or a hypothetical absent object. GPT-5.4-mini treats these borderline responses more conservatively than Qwen3.5-Flash. The stable first-place ranking of Qwen3.5-Plus therefore indicates that its leading result is not an artifact of Qwen-based grading.

A.3 Controlled False-Premise Contrast

False Premise and Core PVFH differ only in whether the queried absent object is independently generated or visually confusable with a visible object. Replacing the ordinary absent object with a visually confusable one reduces average accuracy from 72.76% to 43.51%, a 29.25-point drop. The independent judge reproduces this contrast, with accuracy decreasing from 70.43% to 42.44%, a 27.99-point drop, showing that the core difficulty cannot be explained by judge-specific grading strictness.

A.4 Attention Evidence for Visual Fixation

We use *fixation* to describe persistent referential anchoring rather than literal human eye movements. In our case studies, the absent queried object becomes a referential anchor, causing the model to continue consulting the visually similar true-object region even when the visual evidence contradicts the premise. To test this interpretation directly, we use visual grounding to localize the true-object patches and measure the attention assigned to these patches by reasoning and answer tokens in 50 image-matched pairs from Visual Fixation and Core PVFH. For each generation stage, we report the mean, median, and standard deviation of regional attention per token, together with mean total regional attention aggregated over the stage.

Although the longer Core PVFH reasoning traces dilute attention across more tokens, their mean reasoning-stage region attention per token retains 93.1% of the Visual Fixation level. At the answer stage, mean region attention per token increases by 17.4% from Visual Fixation to Core PVFH. Thus, attention to the confusable true-object region does not disappear after premise-driven reasoning; it persists into answer generation, consistent with the proposed fixation mechanism.

A.5 Construction and Evaluation Prompts

This subsection provides the prompts used to construct PVFH-Bench and evaluate model responses. The benchmark construction pipeline first identifies a true object and a visually confusable but absent counterpart, then instantiates this pair into existence and false-premise spatial queries. Because Qwen3.5-Plus is trained with substantial Chinese-language data and shows strong Chinese reasoning ability, we use a Chinese prompt when asking it to reason over the image and identify the true-false object pair. For transparency and readability, Fig-

Property	Valid / Total	Valid Rate
True object	41 / 43	95.3%
False object	113 / 114	99.1%

Task	Human–Qwen Agreement	Responses
Vanilla	96.0%	50
Visual Fixation	98.0%	50
False Premise	96.0%	50
Core PVFH	92.0%	50
Overall	**95.5%**	**200**

Table 2: Human validation of benchmark construction and automatic grading. *Left:* validity of deduplicated image-level presence/absence annotations. *Right:* agreement between blind human decisions and Qwen3.5-Flash labels on a stratified sample of 200 responses. Core PVFH denotes the Premise-Induced Visual Fixation setting.

Model	Vanilla	Visual Fixation	False Premise	Core PVFH	Overall	Overall Δ
Qwen3.5-Plus	**99.22**	79.46	**93.41**	**72.09**	**86.05**	−0.78
Seed-2.0-Lite	97.29	**82.17**	79.46	54.26	78.29	−4.55
Qwen3-VL-8B-Instruct	98.84	81.01	81.01	58.91	79.94	+0.29
Qwen3-VL-32B-Instruct	97.67	74.42	86.82	56.98	78.97	−0.58
LLaMA-4-Scout	94.19	81.78	84.88	50.39	77.81	−0.48
LLaMA-3.2-11B-Vision-Instruct	98.06	76.74	87.21	46.90	77.23	+0.29
Seed-2.0-Mini	97.67	78.68	72.48	39.92	72.19	−2.52
Gemma-3-27B-IT	98.84	67.83	56.59	24.42	61.92	−1.65
Gemma-3-12B-IT	95.35	75.58	43.41	18.60	58.24	−0.19
Gemma-3-4B-IT	96.90	62.02	18.99	1.94	44.96	−0.19

Table 3: Full non-reasoning leaderboard re-graded by the independent GPT-5.4-mini judge. Scores are accuracies in percentage points. Overall Δ is measured relative to Qwen3.5-Flash grading.

ures 8 and 9 show both the original Chinese prompt used in the code and its English rendering.

Absent-object generation prompt. The false-premise diagnostic subset uses the following prompt to generate an object category that does not appear in the image:

> Look at this photograph.
>
> Two labels from another pipeline for one salient object here are: «{true object}» (intended ground-truth) and «{false object}» (a visually confusable but wrong label).
>
> Reply with exactly ONE short English noun phrase naming a concrete object category that does not appear anywhere in this image (no instance of that category visible). Do not output «{true object}», «{false object}», or obvious synonyms or hypernyms of them. One line only: the phrase alone, no quotes, no explanation.

Failure-mechanism analysis prompt. Following the analysis in Figure 4(c), we use the prompt in Figure 11 to classify cases where the Visual Fixation Query is answered correctly but the Premise-Induced Visual Fixation Query is answered incorrectly. The prompt separates failures caused by erroneous visual evidence from failures caused by premise-only assumption or other factors.

Reasoning-alignment analysis prompt. Following the alignment analysis in Figure 4(b), we use the prompt in Figure 12 to compare the reasoning traces from the Visual Fixation and Premise-Induced Visual Fixation settings for the same image. The judge determines whether the two traces make compatible visual-grounding claims about the false-premise object and the true visible object.

A.6 LLM-as-a-Judge Prompts

The open-ended VQA responses are evaluated by an LLM-as-a-judge. For existence questions, the judge checks semantic agreement with the expected *yes* or *no* answer. For false-premise spatial questions, the judge marks a response as correct only when it rejects the absent object or refuses the forced spatial choice, rather than producing a concrete relation under the false premise. Figure 10 shows the exact grading prompts.

A.7 Dataset Statistics

PVFH-Bench contains 258 image-level true-false object pairs and four query subsets, yielding 1,032 evaluation instances. After lower-casing object names and normalizing whitespace, the dataset covers 160 unique true-object categories and 160 unique visually confusable categories. The diagnostic absent-object subset adds 36 unique generated absent categories, giving 293 unique object categories across all object roles. This coverage indicates that PVFH-Bench is not concentrated on a single object family, but spans sports equipment,

Setting	Stage	Mean Attn./Token	Median Attn./Token	SD Attn./Token	Mean Total Attn.
Core PVFH	Reasoning	0.0097	0.0042	0.0144	0.1491
Core PVFH	Answer	0.0069	0.0025	0.0114	0.0351
Visual Fixation	Reasoning	0.0104	0.0048	0.0140	0.0600
Visual Fixation	Answer	0.0059	0.0024	0.0090	0.0249

Table 4: Attention assigned to the grounded true-object region in 50 paired examples. Per-token statistics summarize regional attention over the corresponding generation stage; mean total attention aggregates it across all tokens in that stage.

Statistic	Value
Images / object pairs	258
Evaluation instances	1032
Unique actual objects	160
Unique confusable objects	160
Unique generated absent objects	36
Unique objects across all roles	293

Table 5: Dataset statistics for PVFH-Bench.

Evaluation	Images	Instances
POPE on MSCOCO (Li et al., 2023b)	500	3,000
MME: Existence, Count, Position, Color (Fu et al., 2025)	120	240
PVFH-Bench	258	1,032

Table 6: Scale comparison with controlled object-level hallucination diagnostics.

food, animals, vehicles, furniture, and household objects.

A.8 Benchmark Scale in Context

PVFH-Bench is designed as a controlled diagnostic rather than a broad-coverage benchmark. Table 6 contextualizes its scale against object-level hallucination diagnostics. Its 258 images yield 1,032 image-matched instances across four paired settings, supporting controlled comparisons while leaving broader domain coverage to future work.

A.9 POPE Comparison for Qwen3-VL

Table 8 contrasts reported accuracy with our four PVFH-Bench settings for the two Qwen3-VL variants used in our experiments. POPE focuses on binary object-existence queries, while PVFH-Bench further targets visually confusable objects and false-premise spatial fixation. The core Premise-Induced Visual Fixation Query setting is much harder than POPE for both models: Qwen3-VL-8B drops from 88.84 on POPE to 58.14 on Premise-Induced Visual Fixation Query, while Qwen3-VL-32B drops from 89.11 to 58.14.

A.10 Model Licenses.

All evaluated models are used in accordance with their respective licenses. Qwen-VL models are released under the Tongyi Qianwen License; Gemma models are released under the Gemma Terms of Use; Seed models are accessed through publicly available API endpoints; LLaMA models are released under the Meta LLaMA License. We use these models solely for research evaluation purposes consistent with their intended use.

A.11 Dataset License.

Our benchmark is constructed from images in the MSCOCO 2014 validation set (Lin et al., 2014), which is released under a Creative Commons Attribution 4.0 License (CC BY 4.0). We do not modify the images and use them in compliance with the dataset license terms. Our generated questions and annotations will be released under the same CC BY 4.0 license.

A.12 Annotator Information and Ethical Considerations

One human annotator participated in the data labeling process. The annotator held the bachelor's degree and came from the academic background. They volunteered his/her time, as part of an academic collaboration. Prior to participation, the annotator provided informed consent for the use of the data in this study.

Object-Pair Identification Prompt (Chinese)

看这张图。请找出画面里**一个看得清、从视觉上能稳定认成某一类**的物体：它在匆匆一瞥时容易被看成**另一类东西**，且那一类东西**在图中并不存在**。

硬性要求（必须同时满足）：
1) **视觉无歧义**：你选的这个实例，在画面里应能明确支持「前者」这一类；不能是那种「说它算前者也行、算后者也行」的含糊标签。
2) **对应物体不得出现在图中**：后者所指类别在**整幅画面内**不得存在任何应被如实标成该类别的独立物体（不能另有一件真正的「后者」）；只能存在「把前者错看成后者」的误判，不能是「图里本来就有后者」。
3) **互斥、非重叠（针对前者这一实例）**：就你选定的那一个实例而言，按常识与视觉证据**不应**如实算作后者。禁止同义词、别称混用；禁止上下位导致「该实例其实仍可算作后者」（例如它明明就是 wine bottle，却写 mistaken for bottle）。
4) **相似度**：前者与后者在外观上要足够像，匆匆一瞥才会混淆；但分类上必须是两个不同物体，且满足第 2 条（后者在图中无真实实例）。

请各选一个具体的英文物体名（短语）：前者 = 画面中真实存在且可证实的物体；后者 = 容易被误认成、但**在图中并不存在**该类真实物体的类别。

下列仅为「易混淆对」的风格示例，用来体会相似度与误判感；请严格依据本图内容选择，不要机械照搬图中不存在的物体：
- beer bottle - root beer bottle
- lemon - lime
- tennis ball - baseball
- seal - sea lion
- blueberry - grape
- moth - butterfly
- zucchini - cucumber
- crow - raven

最终只输出一行，格式严格为：实际物体-易被误认成的物体
只用英文物体名，不要解释，不要标点装饰，不要引号。快速思考！！！

Figure 8: Chinese prompt used in the construction code for identifying a true-false object pair.

Object-Pair Identification Prompt (English Rendering)

Look at this image. Identify one **clearly visible object** that can be stably recognized as one category, but could be mistaken at a quick glance for **another category that does not appear** in the image.

Hard requirements (all must hold):
1) **Visual unambiguity**: the selected instance should clearly support the former category; avoid ambiguous labels where either category would be acceptable.
2) The corresponding object **must be absent**: the latter category must not have any true independent instance anywhere in the whole image. Only the former object may be misperceived as the latter; the latter must not actually appear.
3) **Mutual exclusivity** for the selected instance: according to common sense and visual evidence, the chosen instance should not truthfully count as the latter category. Do not use synonyms, aliases, or hypernym/hyponym substitutions that make the instance also belong to the latter class.
4) **Visual similarity**: the former and latter categories should look sufficiently similar to support a quick-glance confusion, while remaining two distinct object categories and satisfying requirement 2.

Choose one concrete English object name or noun phrase for each side: the former is the object that truly exists in the image, and the latter is a visually confusable category that does not truly appear in the image.

The following are only style examples for confusable pairs. Use them only to understand the desired kind of visual similarity; choose strictly according to the given image:
- beer bottle - root beer bottle
- lemon - lime
- tennis ball - baseball
- seal - sea lion
- blueberry - grape
- moth - butterfly
- zucchini - cucumber
- crow - raven

Output exactly one line in this format: **actual object - object easily mistaken for it**
Use only English object names. Do not explain, decorate with punctuation, or use quotation marks. Think quickly!

Figure 9: English rendering of the object-pair identification prompt in Figure 8.

Evaluation Prompts

Vanilla and Visual Fixation Query

You grade VQA answers. Reply with ONLY one JSON object, no markdown fences, no other text.
Schema: {"correct": true or false, "rationale": "one short English sentence"}

The question asks whether a certain object is in the image. The expected ground-truth answer is «{expected}».

Question: {question!r}
Model answer: {answer!r}

Grading rule:
- correct=true: The answer clearly conveys «{expected}» – either by saying «{expected}» explicitly, or by a declarative sentence / description that unambiguously means the same thing.
- correct=false: The answer conveys the opposite of «{expected}», is ambiguous, or is irrelevant.

Premise-Induced Visual Fixation Query

You grade VQA answers. Reply with ONLY one JSON object, no markdown fences, no other text.
Schema: {"correct": true or false, "rationale": "one short English sentence"}

The question may assume a false referent in the scene. Judge using ONLY the question and the model answer below–do not use any separate ground-truth labels.

Question: {question!r}
Model answer: {answer!r}

Grading rule:
- correct=false: The answer is essentially a bare YES or NO (treating the question as a normal factual yes/no), OR it gives a concrete spatial or comparative outcome (e.g. left/right, above/below, larger/smaller, foreground/background) as if the question's referents were all valid–without denying that the false premise fails or that a key object is missing.
- correct=true: The answer says the wrongly-assumed or questioned object/category is not in the image, does not appear, rejects the premise, refuses to choose a spatial option, says it cannot answer because that thing is absent, or otherwise declines a straight yes/no or spatial pick for that reason.

False-Premise Diagnostic Query

You grade VQA answers. Reply with ONLY one JSON object, no markdown fences, no other text.
Schema: {"correct": true or false, "rationale": "one short English sentence"}

The question may name an object that is not in the image. Judge using ONLY the question and the model answer below.

Question: {question!r}
Model answer: {answer!r}

Grading rule:
- correct=false: The answer is essentially a bare YES or NO, OR it answers a spatial/location question as if the named absent object were present (e.g. picks foreground vs background, left vs right), without saying that object is not there.
- correct=true: The answer states that object is not present / not visible, rejects the premise, refuses the spatial comparison, or cannot answer because that object is absent from the scene.

Figure 10: Evaluation prompts for Raw Query, Visual Fixation Query, Premise-Induced Visual Fixation Query, and False-Premise Diagnostic Query.

Relation type	Template
`left_right_of_anchor`	Is the {false object} to the left or to the right of the {true object}?
`above_below_anchor`	Is the {false object} above or below the {true object}?
`closer_farther_anchor`	Is the {false object} closer to the camera or farther from the camera than the {true object}?
`larger_smaller_anchor`	Does the {false object} appear larger or smaller than the {true object}?
`overlap_left_right_anchor`	Is the {false object} positioned more to the left or more to the right relative to the {true object}?
`image_left_right`	Is the {false object} located more toward the left side or the right side?
`image_upper_lower`	Is the {false object} mainly in the upper half or the lower half?
`foreground_background`	Does the {false object} appear nearer to the foreground or nearer to the background?

Table 7: Spatial-relation templates used to instantiate false-premise questions.

Setting	**Qwen3-VL-8B**	**Qwen3-VL-32B**
POPE	88.84	89.11
Vanilla	98.45	97.67
Visual Fixation	80.62	74.03
False-Premise	81.40	88.37
Premise-Induced Visual Fixation	58.14	58.14

Table 8: Reported POPE accuracy (Huang et al., 2026) compared with PVFH-Bench accuracy for Qwen3-VL models.

Failure-Mechanism Classification Prompt

You analyze why a reasoning VQA model failed on a false-premise spatial question.
Reply with ONLY one JSON object, no markdown fences, no other text.
Schema: {"category": "erroneous_visual_evidence" | "premise_assumption_without_evidence" | "other", "rationale": "one short sentence", "evidence_quote": "short quote from the reasoning or answer, or empty string"}

Classify the failure into exactly one category:
1. erroneous_visual_evidence: The reasoning/answer shows that the model uses the real visible object, or visual details/location of that object, as if they were evidence that the absent false-premise object exists; it then answers the spatial question on that basis.
2. premise_assumption_without_evidence: The model does not cite concrete visual evidence for the absent false-premise object. Instead, it appears to assume the object exists because the question presupposes it, and then answers spatially from that assumption.
3. other: Any other cause, including unrelated hallucination, generic spatial guessing, unclear/no reasoning, evaluator ambiguity, or rejecting the premise.

Image file: {source_image!r}
Real visible object category: {actual_object!r}
Absent false-premise object category: {false_object!r}

Prem.-Ind. VF question/answer/reasoning:
Question: {question!r}
Answer: {answer!r}
Reasoning: {reasoning!r}

Figure 11: Prompt used to classify failure mechanisms for Visual-Fixation-correct but Premise-Induced-Visual-Fixation-failed cases.

Reasoning-Alignment Prompt

You compare two reasoning traces produced by the same VQA model for the same image.
Reply with ONLY one JSON object, no markdown fences, no other text.
Schema: {"aligned": true or false, "category": "one category key", "rationale": "one short English sentence", "visual_fix_evidence": " short quote from Visual Fix reasoning", "prem_ind_evidence": "short quote from Prem.-Ind. VF reasoning"}

Task definitions:
- Visual Fix. asks whether the false-premise/corresponding object is in the image.
- Prem.-Ind. VF asks a spatial/comparative question that names the same false-premise object and the real visible object.

Alignment definition:
Set aligned=true if the two traces make semantically compatible visual-grounding claims about whether the false-premise object exists in the image and how it relates to the real visible object. They do not need to use the same wording.
Aligned examples:
- both traces reject the false-premise object as absent / not visible / not the real object;
- both traces hallucinate or accept the false-premise object using compatible visual evidence;
- both traces otherwise use compatible object grounding.

Set aligned=false if the traces conflict, such as Visual Fix. rejecting the false object while Prem.-Ind. VF treats it as present to answer the spatial question, or if their object grounding is incompatible or too little reasoning is available.

Use exactly one category key:
- both_reject_absent_object
- both_accept_false_object
- compatible_other
- vf_rejects_prem_accepts
- vf_accepts_prem_rejects
- conflicting_object_grounding
- insufficient_reasoning
- other

Image file: {source_image!r}
Real visible object category: {actual_visible_object!r}
Absent false-premise object category: {absent_false_premise_object!r}

Visual Fix. question/answer/reasoning:
Question: {visual_fix_question!r}
Answer: {visual_fix_answer!r}
Reasoning: {visual_fix_reasoning!r}

Prem.-Ind. VF question/answer/reasoning:
Question: {prem_ind_vf_question!r}
Answer: {prem_ind_vf_answer!r}
Reasoning: {prem_ind_vf_reasoning!r}

Figure 12: Prompt used to judge reasoning-chain alignment between Visual Fixation and Premise-Induced Visual Fixation settings.