When Relations Break: Analyzing Relation Hallucination in Vision-Language Model Under Rotation and Noise

Philip Wootaek Shin, Ajay Narayanan Sridhar, Sivani Devarapalli,
Rui Zhang, Jack Sampson, Vijaykrishnan Narayanan
The Pennsylvania State University
{pws5345,afs6372,lfd5379,rmz5227,jms1257,vxn9}@psu.edu

Abstract

Vision–language models (VLMs) achieve strong multimodal performance but remain prone to relation hallucination, which requires accurate reasoning over inter-object interactions. We study the impact of visual perturbations, specifically rotation and noise, and show that even mild distortions significantly degrade relational reasoning across models and datasets. We further evaluate prompt-based augmentation and preprocessing strategies (orientation correction and denoising), finding that while they offer partial improvements, they do not fully resolve hallucinations. Our results reveal a gap between perceptual robustness and relational understanding, highlighting the need for more robust, geometry-aware VLMs.

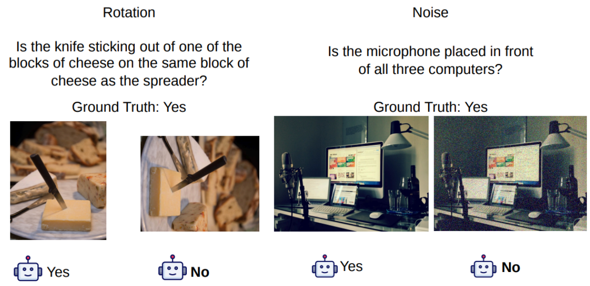

Figure 1: VLM response under visual perturbations.

1 Introduction

Vision–language models (VLMs) have been widely deployed across a range of applications, including object recognition (Feng et al., 2025; Jin et al., 2024), scene understanding (Liao et al., 2024; Selvam et al., 2025), and multimodal reasoning (Jiang et al., 2025; Xu et al., 2025). Despite their impressive capabilities, these models are known to exhibit hallucinations, where the generated outputs are inconsistent with the visual input. Such hallucinations typically manifest in three forms: object hallucination, attribute hallucination, and relation hallucination (Bai et al., 2025). Among these, relation hallucination remains particularly challenging, as it requires accurately capturing interactions and spatial relationships between objects rather than simply identifying their presence or attributes (Zheng et al., 2025).

In this work, we observe that relation hallucination is highly sensitive to visual perturbations, such as image noise and rotation, as illustrated in Fig. 1. While often studied independently, rotation and noise represent complementary failure modes—rotation disrupts geometric invariance, while noise degrades visual fidelity—and frequently co-occur in real-world settings, jointly affecting both structure and quality. While one might argue that rotating an image can alter perceived spatial relationships, humans naturally compensate for such transformations by mentally correcting orientation and still infer consistent object relations. In contrast, VLMs often fail to exhibit this invariance, leading to significantly degraded relational reasoning under such conditions.

To systematically investigate this phenomenon, we make the following contributions:

- Comprehensive Analysis: We conduct a systematic study of relation hallucination under varying noise and rotation—capturing both geometric and photometric perturbations—across multiple VLMs and datasets, providing a unified evaluation of robustness.

- Prompt-Based Intervention: We analyze how incorporating auxiliary prompts (e.g., rotation metadata or noise descriptions) influences model outcome and hallucination behavior.

- Mitigation Strategy: We propose a practical mitigation approach by introducing preprocessing modules—such as rotation correction and denoising—prior to VLM inference, demonstrating improved robustness in relational reasoning.

2 Related Work

Hallucination in vision–language models can be categorized into object, attribute, and relation hallucination, with the latter being the most challenging due to its reliance on inter-object dependencies and spatial reasoning (Bai et al., 2025). Existing benchmarks study relation hallucination using yes/no and multiple-choice formats (Nie et al., 2025; Wu et al., 2024; Zheng et al., 2025), but largely focus on clean visual inputs.

Recent work has explored VLM robustness under visual perturbations, showing that noise and rotation can significantly degrade performance (Zhang et al., 2025; Shin et al., 2025; Niu et al., 2026). However, these studies consider perturbations in isolation and do not directly address their impact on relation hallucination.

While preprocessing methods such as orientation detection and denoising (Barbosa, 2025; Chen et al., 2025; Yu et al., 2024) improve visual quality, their effectiveness in mitigating hallucination remains unclear. In contrast, our work systematically studies the combined effects of noise and rotation on relation hallucination and evaluates both prompt-based and preprocessing-based mitigation strategies, bridging robustness analysis with hallucination-specific evaluation.

3 Rotation Analysis

To evaluate our hypothesis on the effect of rotation on relation hallucination, we utilize three benchmark datasets: MMRel (Nie et al., 2025), R-Bench (Wu et al., 2024), and Reefknot (Zheng et al., 2025). Given that each dataset exhibits distinct characteristics and potential overlaps in image sources and annotations, we carefully curate subsets to ensure a fair and non-redundant evaluation.

Specifically, we remove duplicate image–question pairs and avoid overlapping samples across datasets, while maintaining diversity in relational queries. Through this selective curation process, we construct evaluation sets comprising 1,632 image–question pairs for MMRel, 3,466 pairs for R-Bench, and Reefknot subsets with 1,185 multiple-choice questions and 2,922 binary (yes/no) questions. This rigorous selection enables a consistent and unbiased analysis of relation hallucination under rotational perturbations.

Model	No Rotation (Orig)	90°	270°
Qwen2-VL 7B	78.02	72.07	72.65
InternVL2 8B	79.80	75.36	75.62
LLaVA-Next 8B	80.29	78.85	78.56
DeepSeek-Janus 7B	64.66	64.66	64.66
LLaMA-3.2 11B (Vision)	81.16	77.29	77.44

Table 1: Performance (%) of open-source vision–language models on R-Bench under different rotation settings.

3.1 Experimental Setup for Rotation-Induced Relation Hallucination

To investigate the effect of image rotation on relation hallucination, we apply controlled clockwise (90°) and counterclockwise (270°) rotations to the input images. We evaluate both rotation directions using five open-source vision–language models—Qwen2-VL 7B (Wang et al., 2024), InternVL2 8B (Chen et al., 2024), LLaVA-Next 8B (Liu et al., 2024), DeepSeek-Janus 7B (DeepSeek-AI, 2024), and LLaMA-3.2 11B (Vision) (Meta AI, 2024)—on the R-Bench dataset which we select as a representative benchmark due to its balanced coverage of relational reasoning tasks and manageable scale for controlled experimentation across multiple models.

As shown in Tab. 1, we observe no significant performance difference between 90 degrees clockwise and counterclockwise rotations. Therefore, for clarity and consistency, we report results using the clockwise rotation setting throughout the paper. We exclude 180° rotations, as such transformations are less representative of real-world viewing conditions and largely preserve global object configurations without introducing meaningful ambiguity in relational reasoning.

For evaluation, we employ three widely used closed-source vision–language models: GPT-5.1 (OpenAI, 2026), Gemini 2.5 Pro (DeepMind, 2026), and Claude Sonnet 4.5 (Anthropic, 2026). These models are selected due to their strong multimodal reasoning capabilities and broad adoption in recent studies.

During inference, each model is provided with the rotated image and the original question prompt, without any additional guidance or metadata. The generated responses are then compared against the ground-truth answers provided in each dataset. Quantitative results and comparative analyses are summarized in Tab. 2. Overall, we observe a con-

Category	GPT-5.1	Gemini 2.5 Pro	Claude Sonnet 4.5
Reefknot Y/N (orig)	76.61	75.41	72.25
Reefknot Y/N (cw)	73.03	71.73	65.20
Reefknot MCQ (orig)	88.51	86.08	80.08
Reefknot MCQ (cw)	82.61	80.92	73.33
R-Bench Y/N (orig)	80.15	75.97	79.69
R-Bench Y/N (cw)	78.51	48.93	72.74
MMRel Y/N (orig)	89.83	90.87	64.03
MMRel Y/N (cw)	58.82	53.25	30.45

Table 2: Accuracy (%) of vision-language models under original (orig) and clockwise rotated (cw) settings across Reefknot, R-Bench, and MMRel datasets.

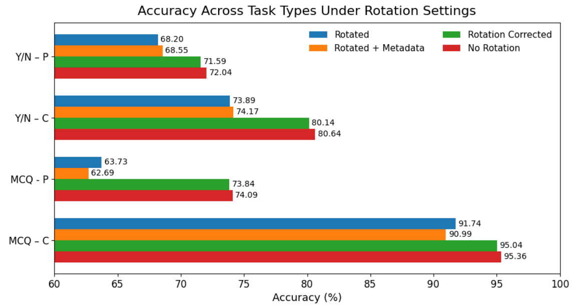

Figure 2: Effect of rotation metadata on VLM accuracy. (Reefknot, GPT-5.1)

sistent performance degradation under clockwise rotation across all models and datasets, confirming the sensitivity of VLMs to geometric perturbations in relational reasoning task. We focus on closed-source VLMs as they represent state-of-the-art, widely deployed systems and must be evaluated under realistic black-box conditions, ensuring our findings generalize to real-world applications.

3.2 Mitigating Rotation-Induced Relation Hallucination: Prompting vs. Preprocessing

From Tab. 1 and Tab. 2, we observe that image rotation consistently increases relation hallucination, motivating the need for effective mitigation strategies. To address this issue, we explore two approaches: (1) prompt-based augmentation using rotation metadata and (2) preprocessing via image orientation correction.

First, we investigate whether incorporating auxiliary prompts—specifically rotation information—can improve model robustness. Our hypothesis is that providing explicit orientation metadata may help VLMs compensate for geometric transformations and thus reduce hallucination. We conduct this analysis using GPT-5.1, which demonstrated relatively strong robustness on the Reefknot dataset under rotation. The Reefknot dataset contains both perceptual questions (orientation-related) and cognitive questions (action/interaction-related). However, as shown in Fig. 2, incorporating rotation metadata as an additional prompt yields minimal improvement in accuracy, suggesting that prompt-based guidance alone is insufficient to mitigate relation hallucination.

In contrast, we explore a preprocessing-based approach using a lightweight image orientation detector (Barbosa, 2025). We evaluate this detector on a modified Reefknot dataset with randomly applied 90° and 270° rotations. The detector achieves 99.66% accuracy on multiple-choice questions (MCQ) with a downstream accuracy of 95.04% (Cognitive) and 73.84% (Perceptual) for GPT-5.1 and 99.38% accuracy on yes/no (Y/N) questions with a downstream accuracy of 80.14% (Cognitive) and 71.59% (Perceptual) for GPT-5.1. These results indicate that correcting image orientation prior to VLM inference is a highly effective strategy, and is more reliable than relying on additional prompting to convey rotation information.

4 Noise Analysis

To evaluate noise effects on relation hallucination, we adopt 19 corruption types from (Zhang et al., 2025; Hendrycks and Dietterich, 2019) across Reefknot, R-Bench, and MMRel, grouped into six categories (A–F). We exclude weather (E) and geometric/structural (F) corruptions, as they introduce unrealistic artifacts and alter scene geometry, thereby invalidating ground-truth object relationships.

From the remaining geometry-preserving categories (A–D: noise-based, blur/distortion, compression/resolution, and photometric/color corruptions), we select one representative corruption each—Impulse Noise, Gaussian Blur, Pixelate, and Saturate—to enable a controlled and interpretable evaluation while preserving spatial relationships critical for assessing relation hallucination.

4.1 Impact of Noise Corruptions and Severity on Relation Hallucination

To examine the effect of noise on relation hallucination, we apply four representative corruptions—Impulse Noise, Gaussian Blur, Pixelate, and Saturate—at severity level 2. We hypothesize that introducing such perturbations would increase hallucination by degrading visual fidelity and disrupting relational cues. We evaluate this using the Reefknot dataset with GPT-5.1. As shown in Tab. 3, we

Category	Baseline	Gaussian Blur	Impulse Noise	Pixelate	Saturate
MC – Perceptual	74.09%	62.18%	64.77%	63.47%	63.73%
YN – Perceptual	72.04%	69.69%	69.36%	69.88%	70.22%
MC – Cognitive	95.36%	91.99%	91.86%	92.61%	92.49%
YN – Cognitive	80.64%	75.05%	76.14%	76.92%	77.19%

Table 3: Impact of noise corruptions (severity level 2) on relation hallucination performance using GPT-5.1 on the Reefknot dataset. Results are reported as correct predictions (accuracy %). All corruption types lead to performance degradation compared to the baseline.

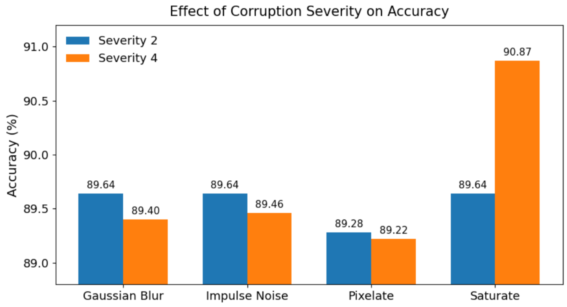

Figure 3: Effect of corruption severity on relation hallucination performance. (MMRel dataset, GPT-5.1)

observe a consistent decrease in accuracy across both yes/no and multiple-choice settings, as well as across perceptual and cognitive tasks, confirming the negative impact of noise on relational reasoning.

Furthermore, we analyze the effect of increasing corruption severity using GPT-5.1 on the MMRel dataset. As reported in Fig. 3, performance generally degrades as severity increases. However, *saturate* occasionally improves accuracy, likely because enhanced color contrast strengthens salient cues, making object relationships easier for VLMs to detect.

4.2 Mitigating Noise-Induced Relation Hallucination: Prompting vs. Denoising

From Tab. 3 and Fig. 3, we observe that relation hallucination consistently worsens as image noise increases, confirming the sensitivity of VLMs to visual corruptions. Motivated by this, we explore two mitigation strategies: (1) prompt-based augmentation using noise information and (2) preprocessing via image denoising prior to model inference.

From Tab. 4, we compare the effectiveness of prompt-based augmentation and preprocessing-based denoising for mitigating noise-induced relation hallucination. Overall, both approaches provide limited but non-negligible improvements, with effectiveness varying across datasets and corruption types.

Dataset	Corruption	Acc (%)	Drop (pp)	Acc-D (%)	Drop-D (pp)
MMRel	Gaussian Blur	86.70	3.12	89.34	0.49
	Impulse Noise	87.32	2.51	88.97	0.86
	Pixelate	86.58	3.25	89.34	0.49
	Saturate	85.54	4.29	87.68	2.15
R-Bench	Gaussian Blur	80.84	8.99	80.06	9.76
	Impulse Noise	80.35	9.48	80.18	9.65
	Pixelate	80.24	9.59	80.18	9.65
	Saturate	78.04	11.78	77.44	12.39

Table 4: Comparison of noise corruption (S2) and preprocessing-based denoising on relation hallucination across MMRel and R-Bench using GPT-5.1. Acc/Drop denote results under corruption, while Acc-D/Drop-D denote results after denoising.

On R-Bench, performance drops remain large (9–12 pp) under both settings, indicating that denoising alone cannot fully recover relational reasoning under severe noise. In contrast, on MMRel, denoising yields smaller drops (0.5–4 pp) and generally outperforms prompt-based augmentation, suggesting benefits when underlying visual structure is preserved.

For denoising, we evaluate state-of-the-art restoration models (Chen et al., 2025; Yu et al., 2024) and select outputs using LPIPS (Zhang et al., 2018), PSNR, and SSIM (Wang et al., 2004). Despite improved perceptual quality, gains in relational reasoning are inconsistent, indicating a gap between low-level restoration and high-level understanding.

Overall, denoising is a useful but condition-dependent strategy: it helps when noise affects low-level features, but is less effective when relational cues are disrupted or task complexity is high

5 Conclusion

In this work, we study relation hallucination in VLMs under rotation and noise, showing that even simple perturbations significantly degrade relational reasoning, with larger effects on complex datasets. While prompting and preprocessing (orientation correction and denoising) provide partial improvements, they fail to fully recover performance, and gains in perceptual quality do not consistently translate to better reasoning. These results reveal a gap between perceptual robustness and relational understanding, motivating more robust, geometry-aware VLM designs.

Acknowledgments

This work was supported in part by NSF Awards 2243979 and 2318101

References

Anthropic. 2026. Claude sonnet 4.5. https://claude.ai. Large language model.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2025. Hallucination of multimodal large language models: A survey.

Duarte Barbosa. 2025. Deep image orientation detection. https://github.com/duartebarbosadev/deep-image-orientation-detection. GitHub repository.

I Chen, Wei-Ting Chen, Yu-Wei Liu, Yuan-Chun Chiang, Sy-Yen Kuo, Ming-Hsuan Yang, and 1 others. 2025. Unirestore: Unified perceptual and task-oriented image restoration model using diffusion prior. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17969–17979.

Zhe Chen, Wenhai Wang, Yue Cao, and 1 others. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.

Google DeepMind. 2026. Gemini 2.5 pro. https://deepmind.google. Large language model.

DeepSeek-AI. 2024. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*.

Yongchao Feng, Yajie Liu, Shuai Yang, Wenrui Cai, Jinqing Zhang, Qiqi Zhan, Ziyue Huang, Hongxi Yan, Qiao Wan, Chenguang Liu, and 1 others. 2025. Vision-language model for object detection and segmentation: A review and evaluation. *arXiv preprint arXiv:2504.09480*.

Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.

Chaoya Jiang, Yongrui Heng, Wei Ye, Han Yang, Haiyang Xu, Ming Yan, Ji Zhang, Fei Huang, and Shikun Zhang. 2025. Vlm-r3: Region recognition, reasoning, and refinement for enhanced multimodal chain-of-thought. *arXiv preprint arXiv:2505.16192*.

Sheng Jin, Xueying Jiang, Jiaxing Huang, Lewei Lu, and Shijian Lu. 2024. Llms meet vlms: Boost open vocabulary object detection with fine-grained descriptors. *arXiv preprint arXiv:2402.04630*.

Guibiao Liao, Jiankun Li, and Xiaoqing Ye. 2024. Vlm2scene: Self-supervised image-text-lidar learning with foundation models for autonomous driving scene understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 3351–3359.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Meta AI. 2024. Llama 3.2: Open multimodal foundation models. *arXiv preprint arXiv:2409.17146*.

Jiahao Nie, Gongjie Zhang, Wenbin An, Yun Xing, Yap-Peng Tan, Alex C. Kot, and Shijian Lu. 2025. Mmrel: Benchmarking relation understanding in multi-modal large language models. *Preprint*, arXiv:2406.09121.

Tianyi Niu, Jaemin Cho, Elias Stengel-Eskin, and Mohit Bansal. 2026. Rotbench: Evaluating multimodal large language models on identifying image rotation. *Preprint*, arXiv:2508.13968.

OpenAI. 2026. Chatgpt-5.1. https://chat.openai.com. Large language model.

Surya Selvam, Ravi K Rajendran, Murugan Sankaradas, Anand Raghunathan, and Srimat T Chakradhar. 2025. Simcache: similarity caching for efficient vlm-based scene understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3327–3336.

Philip Wootaek Shin, Jack Sampson, Vijaykrishnan Narayanan, Andres Marquez, and Mahantesh Halappanavar. 2025. Losing the plot: How vlm responses degrade on imperfect charts. *Preprint*, arXiv:2509.18425.

Peng Wang, Shuai Bai, Shengbang Tan, Shuai Wang, Zhihao Fan, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.

Mingrui Wu, Jiayi Ji, Oucheng Huang, Jiale Li, Yuhang Wu, Xiaoshuai Sun, and Rongrong Ji. 2024. Evaluating and analyzing relationship hallucinations in large vision-language models. *Preprint*, arXiv:2406.16449.

Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, and 1 others. 2025. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. *arXiv preprint arXiv:2504.15279*.

Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. 2024. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. *Preprint*, arXiv:2401.13627.

Hao Zhang, Wenqi Shao, Hong Liu, Yongqiang Ma, Ping Luo, Yu Qiao, Nanning Zheng, and Kaipeng Zhang. 2025. B-avibench: Toward evaluating the robustness of large vision-language model on black-box adversarial visual-instructions. *Trans. Info. For. Sec.*, 20:1434–1446.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595.

Kening Zheng, Junkai Chen, Yibo Yan, Xin Zou, and Xuming Hu. 2025. Reefknot: A comprehensive benchmark for relation hallucination evaluation, analysis and mitigation in multimodal large language models. *Preprint*, arXiv:2408.09429.