

---

# Supplementary Material of MIRAGE: Assessing Hallucination in Multimodal Reasoning Chains of MLLM

---

Anonymous Author(s)

Affiliation

Address

email

## 1 More Analysis

Table 1: Hallucination type rates in MIRAGE benchmark questions of Qwen-7/72B with different pretraining data. Pretraining with higher quality data leads to less logical, fabrication, and factual hallucinations.

Model	Logical	Factuality	Spatial	Context	Fabrication
Qwen2.5-VL-72B	47.7%	33.7%	29.2%	21.6%	16.5%
Qwen2-VL-72B	59.3%	45.4%	32.7%	32.6%	26.5%
Qwen2.5-VL-7B	64.7%	45.7%	33.4%	29.3%	25.5%
Qwen2-VL-7B	74.0%	60.6%	35.6%	42.7%	35.4%

Table 2: Hallucination type rates in MIRAGE benchmark questions of Qwen-2.5-VL. Larger Models lead to less logical, fabrication, and factual hallucinations.

Model	Logical	Factuality	Spatial	Context	Fabrication
Qwen2.5-VL-72B	47.7%	33.7%	29.2%	21.6%	16.5%
Qwen2.5-VL-7B	64.7%	45.7%	33.4%	29.3%	25.5%
Qwen2.5-VL-3B	78.9%	60.1%	36.7%	37.9%	38.1%

### 1.1 Relation Between Pretraining Data and Hallucination Types

We also explore relations between pretraining data and hallucination types. Specifically, we keep use Qwen-VL [2, 1] with different pretraining data (*i.e.*, Qwen2-VL and Qwen2.5-VL) and compare the hallucination rates of each hallucination type. As shown in Table 1, Qwen2.5-VL models have less logical, factual, and fabrication hallucination rates than those of Qwen2-VL models. A possible explanation is that pretraining data with higher quality provides more accurate factual knowledge and reasoning chains to models, such that models can avoid logical and factuality hallucinations during inference. Nevertheless, the spatial hallucination does not significantly reduced, which indicates that current MLLMs still show weak visual reasoning capabilities.

### 1.2 Relation Between Model Size and Hallucination Types

We also explore relations between pretraining data and hallucination types. Specifically, we keep use Qwen2.5-VL [1] with different model sizes (*i.e.*, 3B/7B/72B) and compare the hallucination rates of each hallucination type. As shown in Table 2, Larger Qwen2.5-VL models have less logical, factual,

Table 3: Hallucination type rates in MIRAGE benchmark questions of Qwen2.5-VL-3B/7B and corresponding Logos-3B/7B. Our proposed method leads to less logical and fabrication hallucinations.

Model	Logical	Factual	Spatial	Context	Fabrication
Qwen2.5-VL-7B	64.7%	45.7%	33.4%	29.3%	25.5%
Logos-7B	49.3%	39.7%	29.9%	23.8%	15.6%
Qwen2.5-VL-3B	78.9%	60.1%	36.7%	37.9%	38.1%
Logos-3B	57.1%	47.4%	36.7%	31.8%	24.0%

and fabrication hallucination rates than those of smaller models. A possible explanation is that models owning more model parameters have more capabilities for accurate factual knowledge and reasoning chains to models, such that models can avoid logical and factuality hallucinations during inference. Nevertheless, the spatial hallucination does not significantly reduced, which indicates that current MLLMs still show weak visual reasoning capabilities.

### 1.3 Whether Logos Reduces Reasoning Hallucination or Not

Finally we investigate the hallucination mitigate effect on each hallucination type. As shown in Table 3, Logos-7B reduces logical hallucination by 15.4% and fabrication hallucination by 10%. Similar results can also be found in Logos-3B. Nevertheless, we do not find significant hallucination mitigation on spatial and factuality hallucination on both Logos models. A possible reason is that reinforcement learning does not introduce new knowledge and only refines the logic of reasoning chains.

### 1.4 Conclusion

Our findings reveal that the model scale, data scale, and training stages of MLLMs: (1) significantly influence the degree of logical, fabrication, and factual hallucinations; (2) show no effective improvement on spatial hallucinations caused by misinterpretations of spatial relationships, suggesting that current MLLMs exhibit weak visual reasoning capabilities and struggle to benefit from simple scaling of training resources. Our findings will provide insights for future MLLM development.

## References

- [1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibong Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [2] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.