Appendix: Thinking in Uncertainty: Mitigating Hallucinations in MLRMs with Latent Entropy-Aware Decoding

A Additional results with different model scales

A1. Details of Evaluation Benchmarks

In this section, we provide detailed descriptions of the benchmarks used in our evaluation, categorized by their primary focus: general reasoning, hallucination, mathematical reasoning, and scientific reasoning.

A1.1. General Reasoning & Understanding

- **V*Bench (vstar_bench) [16]** is a general-purpose multimodal benchmark constructed from diverse real-world images. It is designed to evaluate the visual understanding, spatial reasoning, and fine-grained attribute recognition capabilities of VLMs.
- **RealWorldQA [17]** is a benchmark consisting of real-world photographs, many captured from vehicle perspectives, paired with short-answer questions. It is designed to probe the basic spatial understanding and everyday scene reasoning of multimodal models.
- **MMVP (MultiModal Visual Patterns) [13]** constructs "CLIP-blind" image pairs alongside 300 carefully designed questions. It aims to test whether models can distinguish subtle visual pattern differences in scenarios where standard visual encoders typically fail.
- **MMEval-Pro [3]** is a calibrated multimodal benchmark that augments existing multiple-choice datasets (e.g., MMMU, ScienceQA, MathVista) with perception and knowledge-anchor questions. It forms question triplets to mitigate text-only shortcuts, offering a more trustworthy evaluation of LMMs.
- **VMCBench [20]** consolidates 20 VQA datasets into a unified multiple-choice format ($\sim$9k questions) via Auto-Converter. This enables the scalable and consistent evaluation of VLMs across diverse visual-linguistic scenarios.

A1.2. Hallucination Benchmark

- **MMHalu (MMHal-Bench) [12]** is a general-purpose multimodal hallucination benchmark built from diverse image-question pairs and categories. It assesses the frequency with which models generate visually inconsistent or fabricated content under realistic settings.
- **Bingo (Bias and Interference Challenges) [1]** is a benchmark that diagnoses both bias-driven errors (e.g., region, OCR, and factual bias) and interference-induced hallucinations through image-text challenge sets.
- **POPE (Polling-based Object Probing Evaluation) [7]** evaluates object hallucination by querying models with balanced yes/no questions regarding object existence. It provides a stable, instruction-agnostic metric to determine whether generated descriptions remain faithful to the image content.

A1.3. Mathematical Reasoning

- **MathVision (MATH-Vision) [14]** comprises 3,040 visual mathematics problems collected from real competitions. Spanning 16 disciplines and 5 difficulty levels, it assesses multimodal mathematical reasoning in realistic exam-style settings.
- **MathVista [9]** is a unified benchmark consolidating 6141 examples from 28 existing datasets and three newly introduced subsets (IQTest, FunctionQA, and PaperQA). It evaluates mathematical reasoning within visual contexts such as diagrams, plots, and scientific figures.
- **MathVerse [19]** is an all-round visual math benchmark consisting of 2612 diagram-based problems transformed into approximately 15k multimodal variants. It enables fine-grained evaluation to determine whether MLLMs genuinely leverage diagrams for reasoning rather than relying solely on textual cues.
- **VisuLogic [18]** comprises 1,000 single-choice visual reasoning problems across six categories (e.g., quantitative shifts, spatial relations, attribute comparisons). It is explicitly designed to test vision-centric logical reasoning by precluding language-only shortcuts.
- **Geometry3K [8]** is a dataset containing 3,002 multiple-choice Euclidean geometry problems, each paired with a diagram and formalized structural annotations. It targets high-school to competition-level geometric capabilities.

A1.4. Scientific Reasoning

- **MMK12 [10]** is a K-12 level multimodal reasoning dataset introduced by the MM-Eureka framework. Its evaluation split provides 500 multimodal multiple-choice questions for each of four disciplines—mathematics, physics, chemistry, and biology—to comprehensively as-

sess foundational scientific reasoning.

A2. General Multimodal Reasoning Benchmarks

In addition to the Qwen2.5-VL-7B-Instruct backbone, we implement LEAD on a diverse set of MLRMs to verify its effectiveness on heterogeneous architectures, as shown in Table 1. The evaluated models include VLM-R1-3B [11], GLM-4.1V-9B-Thinking [2], InternVL3.5-14B [15], Vision-R1-32B [4] and Vision-R1-72B. We report performance on RealWorldQA, MMVP, VStar, and VMCBench, adhering to the evaluation settings in Section 4.

A3. Domain-specific Multimodal Reasoning Benchmarks

We further evaluate these backbones on the scientific subsets of MMK12 to analyze the impact of LEAD on domain-specific reasoning, with results presented in Table 2. Specifically, we report performance across the Math, Physics, Chemistry, and Biology categories. This analysis focuses on structured scientific tasks and adheres to the consistent decoding protocol outlined in Section 4.

B Details about Baselines

To benchmark against existing approaches, we reproduce three representative decoding strategies: Visual Contrastive Decoding (VCD), Memory-space Visual Retracing (MemVR), and Self-Introspective Decoding (SID). All baselines are implemented on the Qwen2.5-VL backbone. We strictly follow the original methodologies and hyperparameter configurations to ensure a fair and consistent comparison.

- **Visual Contrastive Decoding (VCD) [6]:** For VCD, we execute two parallel forward passes at each decoding step: one using the clean image and another with a distorted input. Following the method's core principle, we derive a contrastive distribution to amplify visually consistent signals and suppress hallucinations. Additionally, we apply the feasibility filtering constraint (V_{head}) based on confidence scores from the clean image path. Final tokens are selected using top-p sampling. All hyperparameters, including noise injection methods, contrastive strength, and thresholds, strictly align with the original settings.
- **Memory-space Visual Retracing (MemVR) [21]:** We replicate the MemVR inference process by first pooling visual hidden states to obtain a global visual feature prior to generation. During autoregressive decoding, we monitor top-k entropy to evaluate model uncertainty. When the entropy exceeds a predefined threshold, the visual retrospect mechanism is triggered, adjusting the current token's hidden state towards the global visual feature by a fixed ratio. This intervention reinforces visual grounding and suppresses hallucinations. A greedy decoding strategy is employed.

Table 1. Comparisons of performance on general reasoning benchmarks across different MLRM parameter scales and architecture families. Results are reported as accuracy (%) on all benchmarks.

Model	RealWorldQA	MMVP	Vstar	VMCBench
VLM-R1-3B	61.7	29.1	69.6	71.4
+ LEAD (Ours)	62.8	30.2	71.2	72.9
GLM-4.1V-9B-Think	70.3	54.1	76.4	82.8
+ LEAD (Ours)	71.9	55.6	78.5	84.0
InternVL3.5-8B-Think	67.5	47.9	81.7	82.1
+ LEAD (Ours)	69.1	48.7	82.7	83.2
Vision-R1-32B	71.5	47.8	83.8	81.9
+ LEAD (Ours)	72.1	49.1	84.8	83.2
Vision-R1-72B	72.0	50.8	83.8	82.4
+ LEAD (Ours)	73.6	52.0	85.1	84.3

Table 2. Comparisons of performance on MMK12 subsets across different MLRM parameter scales and architecture families. Results are reported as accuracy (%) on all subsets.

Model	Math	Phys	Chem	Bio
VLM-R1-3B	37.2	35.4	42.2	45.6
+ LEAD (Ours)	38.4	37.8	44.0	47.0
GLM-4.1V-9B-Think	71.2	62.0	69.2	70.0
+ LEAD (Ours)	72.4	64.2	70.4	70.8
InternVL3.5-8B-Think	55.2	45.0	67.0	62.8
+ LEAD (Ours)	57.0	47.6	68.4	64.2
Vision-R1-32B	72.8	63.2	69.0	68.8
+ LEAD (Ours)	73.2	63.8	71.2	69.6
Vision-R1-72B	73.4	63.4	70.6	73.8
+ LEAD (Ours)	74.2	64.0	71.8	75.2

- **Self-Introspective Decoding (SID) [5]:** Our implementation strictly follows the SID pipeline. We extract cross-modal attention at specified decoder layers and compute visual token importance using the CT²S strategy. Visual tokens with the lowest scores are identified as low-relevance regions, and their hidden states are attenuated to construct the introspective logits. These are then combined with the original logits using the contrastive formula to mitigate hallucinations. Finally, we employ a greedy decoding strategy for generation.

C Pass@K Experiments

We report Pass@k results for $k \in [1, 64]$ on the R1-Onevision backbone, extending the analysis provided in the main text. As shown in Fig. 1, LEAD consistently outperforms baseline methods across the evaluated range. The steeper improvement rate at lower k values suggests enhanced sample efficiency, indicating that correct solutions are retrieved with fewer sampled paths. Furthermore, LEAD achieves a higher performance saturation point as k increases, indicating that our latent entropy-aware strategy

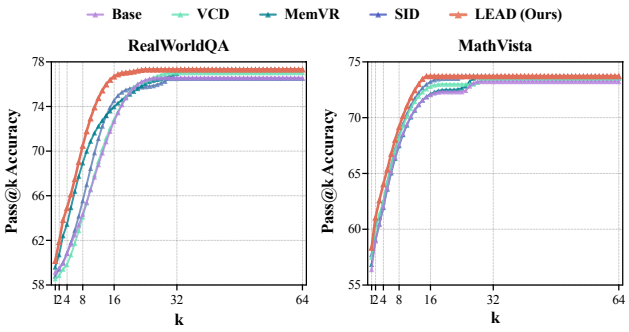

Figure 1. Pass@k accuracy evaluation of R1-Onevision-7B on sampled data of RealworldQA and MathVista, illustrating full results for $k \in [1, 64]$.

effectively promotes diversity in valid reasoning paths.

References

[1] Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. *arXiv preprint arXiv:2311.03287*, 2023. 1

[2] Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv e-prints*, pages arXiv–2507, 2025. 2

[3] Jinsheng Huang, Liang Chen, Taian Guo, Fu Zeng, Yusheng Zhao, Bohan Wu, Ye Yuan, Haozhe Zhao, Zhihui Guo, Yichi Zhang, et al. Mmevalpro: Calibrating multimodal benchmarks towards trustworthy and efficient evaluation. *arXiv preprint arXiv:2407.00468*, 2024. 1

[4] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025. 2

[5] Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-introspective decoding: Alleviating hallucinations for large vision-language models. *arXiv preprint arXiv:2408.02032*, 2024. 2

[6] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882, 2024. 2

[7] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, 2023. 1

[8] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*, 2021. 1

[9] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024. 1

[10] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, et al. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025. 1

[11] Haozhan Shen, Zilun Zhang, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. Vlm-r1: A stable and generalizable r1-style large vision-language model. 2025. Accessed: 2025-02-15. 2

[12] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Annual Meeting of the Association for Computational Linguistics*, 2024. 1

[13] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578, 2024. 1

[14] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024. 1

[15] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 2

[16] Penghao Wu and Saining Xie. V?: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094, 2024. 1

[17] X.AI. Grok-2 beta release, 2024. Accessed: 2024. 1

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

[19] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*. Springer, 2024. 1

[20] Yuhui Zhang, Yuchang Su, Yiming Liu, Xiaohan Wang, James Burgess, Elaine Sui, Chenyu Wang, Josiah Aklilu, Alejandro Lozano, Anjiang Wei, et al. Automated generation of challenging multiple-choice questions for vision language model evaluation. *arXiv preprint arXiv:2501.03225*, 2025. 1

[21] Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Kening Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, et al. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. *arXiv preprint arXiv:2410.03577*, 2024. 2