Supplementary Material: Visual Refocusing while Thinking Improves Test-Time Scaling in Multi-Modal Large Reasoning Models

Soumya Suvra Ghosal[1*] Youngeun Kim[2*] Zhuowei Li[2] Ritwick Chaudhry[2] Linghan Xu[2]
Hongjing Zhang[2] Jakub Zablocki[2] Yifan Xing[2] Qin Zhang[3†]
[1]University of Maryland, College Park [2]Amazon [3]Physion Labs
sghosal@umd.edu {youngeuk,zhuoweli,ritwic,linghanx,zhhongji,jzablock,yifax}@amazon.com qin@physionlabs.ai

A. Limitations

Although VisRef provides consistent improvements across a variety of visual-reasoning benchmarks and model architectures, it introduces additional computational overhead. In particular, applying DPP-based token selection at every reasoning step increases inference latency compared to standard decoding. This accuracy–latency trade-off is inherent to test-time scaling methods; however, our approach offers higher accuracy for a given computational budget than existing alternatives.

B. Software and Hardware

We run all experiments with Python 3.10.18, PyTorch 2.7.0, and Transformers 4.55.0. For all experimentation, we use four Nvidia A10G GPUs.

C. Baselines and Implementation Details.

For evaluation, we compare VisRef with two baselines: (1) standard thinking (Equation 1), where the model generates a single reasoning trace without additional intervention, and (2) textual self-reflection [6], which extends reasoning through text-only reflection without visual refocusing. Based on ablations (Section 5), we set the adaptive stopping entropy threshold $\delta_{\text{entropy}} = 0.25$, the visual token budget $m = \lfloor 0.3|\mathcal{V}| \rfloor$ (i.e., 30% of the total visual tokens $\mathcal{V}$), and the maximum reasoning steps $K_{\max} = 10$.

D. Evaluation Criteria.

To evaluate reasoning performance, we report the accuracy of each model on the test set of each dataset. Specifically, for each image-text input $x_{\text{input}} = [I, T] \in \mathcal{D}^{\text{test}}$, the model first generates a thinking trace τ, followed by the final answer y. The accuracy metric is defined as: $\mathbb{E}_{x_{\text{input}} \sim \mathcal{D}^{\text{test}}, \tau \sim \pi_\theta(\cdot|x_{\text{input}}), y \sim \pi_\theta(\cdot|x_{\text{input}}, \tau)} [\mathbb{I}\{y = y^*\}]$, where y^* is the correct answer.

Datasets & Models. To validate the effectiveness of VisRef, we conduct experiments on three visual reasoning benchmarks.

- **MathVista** [5] unifies 31 visual-math datasets covering puzzles, functional plots, and scientific figures to assess diverse mathematical reasoning skills in visual contexts; we use the *testmini* split containing $1,000$ problems.
- **MathVision** [7] includes 304 visually grounded math competition problems across 16 disciplines and five difficulty levels, enabling fine-grained evaluation of visual mathematical reasoning.
- **MM-Star** [3] is a human-curated benchmark of $1,500$ vision-dependent questions designed to assess six core multimodal capabilities across 18 detailed axes, including perception, spatial reasoning, and commonsense understanding. We evaluate VisRef on three state-of-the-art MLRMs—InternVL3.5-8B [8], SAIL-VL2-Thinking [11], and Qwen-3-VL-8B-Thinking [2]—all evaluated in their reasoning ("thinking") mode, which generates explicit reasoning traces before producing the final answer.

E. Additional Results

In the main paper (Section 5), we evaluated VisRef on three visual reasoning benchmarks. Here, we extend our evaluation to two additional benchmarks: TallyQA [1] and RealWorldQA [10]. TallyQA [1] focuses on complex counting tasks in visual scenes, requiring models to identify and enumerate multiple objects while maintaining spatial awareness. This benchmark is particularly challenging as it tests whether models can preserve precise visual grounding throughout iterative counting processes. RealWorldQA [10] is designed to evaluate real-world visual understanding using over 700 images, including anonymized vehicular footage and diverse real-world scenes. This benchmark tests models' abilities to reason about authentic, uncurated visual scenarios encountered in everyday settings. Table 1 presents results on these

*Equal contribution. This work was done during Soumya Suvra Ghosal's Amazon internship at AWS AI Labs.
†Work done while at AWS AI Labs.

Table 1. **Evaluation on additional visual reasoning benchmarks.** We evaluate VisRef across three visual reasoning benchmarks: TallyQA, and RealQA. To ensure a fair comparison, all methods adopt the adaptive stopping criterion described in Section 4.1.2. For brevity, we denote *Standard Thinking* as **ST**, and *Textual Self-Reflection* [6] as **TSR**. All results are reported in accuracy (%), and the numbers in parentheses indicate the performance gain over the ST baseline.

Model	Method	TallyQA	RealWorldQA
InternVL3.5-8B	ST (Baseline)	79.4	44.6
	TSR [6]	79.6	44.9
	VisRef (Ours)	**84.5** (+5.1)	**47.2** (+2.6)
Qwen3-VL-8B	ST (Baseline)	74.3	55.4
	TSR [6]	75.1	56.9
	VisRef (Ours)	**78.9** (+4.6)	**59.1** (+3.7)
SAIL-VL2-8B	ST (Baseline)	69.3	57.3
	TSR [6]	71.7	58.0
	VisRef (Ours)	**73.9** (+5.4)	**61.2** (+3.9)

benchmarks. On TallyQA, VisRef achieves substantial improvements across all three models, with gains of 5.1%, 4.6%, and 5.4% for InternVL3.5-8B, Qwen3-VL-8B, and SAIL-VL2-8B respectively, compared to standard thinking. Similarly, RealWorldQA shows consistent improvements ranging from 2.6% to 3.9% across the model suite, demonstrating VisRef's effectiveness on real-world visual understanding tasks. Figure 1 illustrates the test-time scaling behavior of VisRef across both additional benchmarks using three MLRMs: InternVL-3.5-8B (top row), Qwen3-VL-8B (middle row), and SAIL-VL2-8B (bottom row). The star marker (☆) indicates the baseline with no additional test-time compute (standard thinking), while successive circles represent increasing test-time token budgets. We compare VisRef against parallel thinking [4, 9], which samples multiple text-only reasoning trajectories without visual refocusing. Across all benchmarks and models, VisRef consistently achieves superior accuracy for any given computational budget.

F. Additional Qualitative Evaluations

To provide deeper insights into how VisRef maintains visual grounding during reasoning, we visualize attention patterns before and after visual refocusing in Figure 2. The visualizations use images from the RealWorldQA [10] dataset with the InternVL-3.5-8B model. We observe that after applying VisRef's visual token reinjection, the attention patterns become substantially more focused on task-relevant regions, confirming that our method effectively counteracts visual token dilution during extended reasoning chains.

G. Derivation of the Relevance-Diversity Decomposition

In this section, we provide the complete derivation of Equation 10 from the main paper, which decomposes the log-determinant of the kernel matrix into relevance and diversity terms. Specifically, given the kernel matrix $L_k^{V_k} \in \mathbb{R}^{|V_k|\times|V_k|}$ restricted to visual token subset V_k, the log-determinant can be decomposed as:

$$\log\det(L_k^{V_k}) = \sum_{v_i\in V_k} \log(r_i^2) + \log\det(\bar{L}_k^{V_k}) \tag{1}$$

where r_i denotes the relevance score of token v_i, and $\bar{L}_k^{\mathcal{V}_k}$ is the normalized diversity kernel.

Derivation. For any selected subset $V_k = \{v_1, \dots, v_m\} \subseteq \mathcal{V}$, the kernel matrix entries are given by:

$$[L_k^{V_k}]_{ij} = L_k(v_i, v_j) = \phi_k(v_i)^\top \phi_k(v_j) = v_i^\top M_k v_j \tag{2}$$

where $\phi_k(v) = M_k^{1/2} v$ projects visual token v into the textual reasoning subspace defined by $M_k = \sum_{j=1}^{T_k} z_k^{(j)} (z_k^{(j)})^\top$.

The relevance of token v_i to the current reasoning state z_k is measured by:

$$r_i^2 = \|\phi_k(v_i)\|_2^2 = v_i^\top M_k v_i = \sum_{j=1}^{T_k} (v_i^\top z_k^{(j)})^2 \tag{3}$$

This quantity captures the alignment between visual token v_i and the textual context, with $r_i^2 = [L_k^{V_k}]_{ii}$. Next, we introduce the normalized kernel $\bar{L}_k^{V_k}$ with entries:

$$[\bar{L}_k^{V_k}]_{ij} = \frac{[L_k^{V_k}]_{ij}}{r_i r_j} = \frac{\phi_k(v_i)^\top \phi_k(v_j)}{\|\phi_k(v_i)\|_2 \|\phi_k(v_j)\|_2} \tag{4}$$

Note that $[\bar{L}_k^{V_k}]_{ii} = 1$ for all i, representing normalized correlations between tokens. Let $D_{V_k} = \mathrm{diag}(r_1, \dots, r_m)$. The kernel matrix can then be factorized as:

$$L_k^{V_k} = D_{V_k} \bar{L}_k^{V_k} D_{V_k} \tag{5}$$

Thus, we can write each element as: $[D_{V_k} \bar{L}_k^{V_k} D_{V_k}]_{ij} = r_i \cdot \frac{[L_k^{V_k}]_{ij}}{r_i r_j} \cdot r_j = [L_k^{V_k}]_{ij}$. Applying the multiplicative property of determinants:

$$\det(L_k^{V_k}) = \det(D_{V_k})^2 \det(\bar{L}_k^{V_k}) = \left(\prod_{v_i\in V_k} r_i^2\right) \det(\bar{L}_k^{V_k}) \tag{6}$$

Finally, taking the natural logarithm yields:

$$\log\det(L_k^{V_k}) = \underbrace{\sum_{v_i\in V_k} \log(r_i^2)}_{\text{relevance term}} + \underbrace{\log\det(\bar{L}_k^{V_k})}_{\text{diversity term}} \tag{7}$$

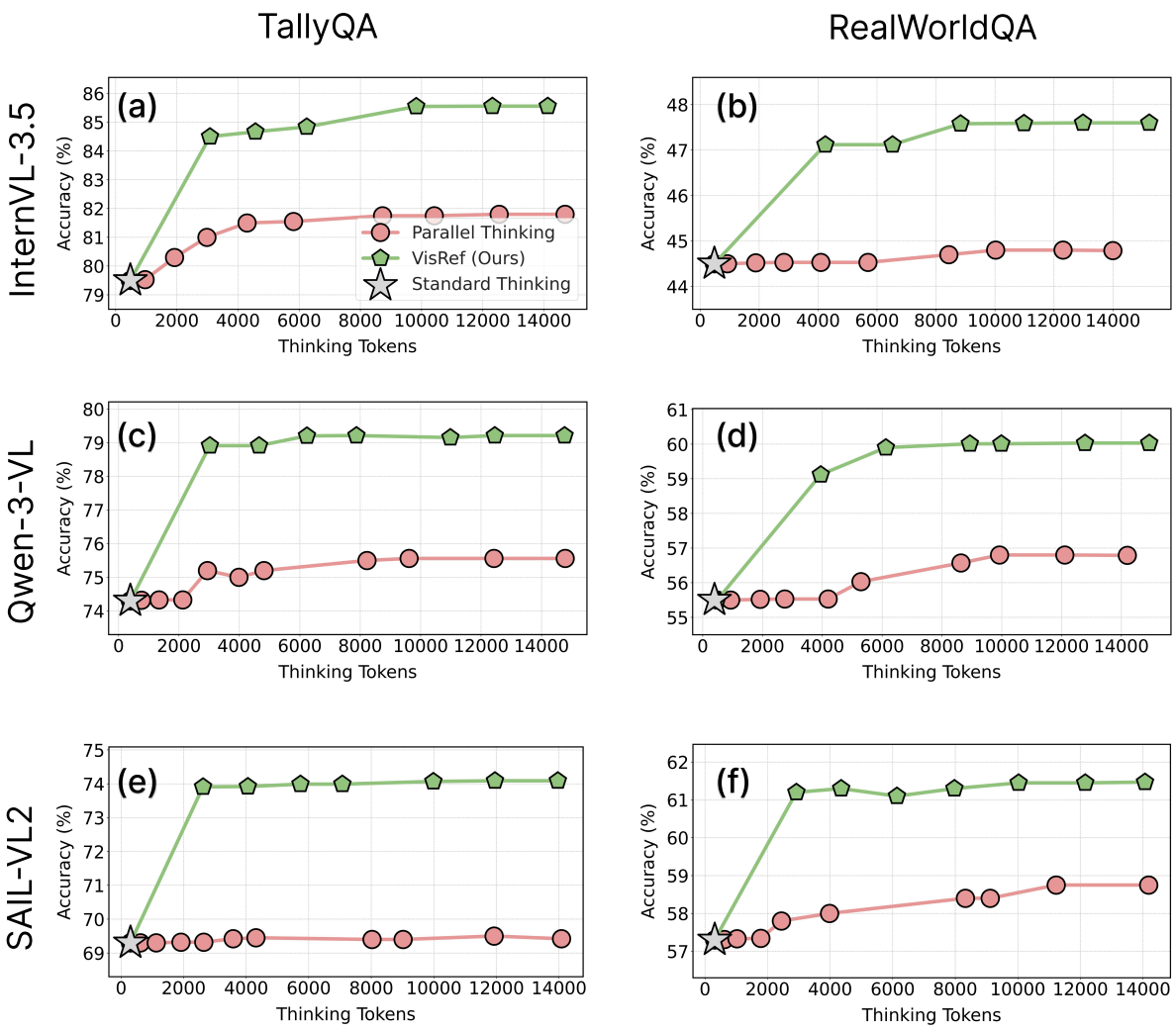

Figure 1. **Test-time scaling of VisRef.** We evaluate the test-time scaling behavior of VisRef by generating multiple parallel visual-integrated reasoning chains under a fixed token budget. Results are shown across two benchmarks (TallyQA, and RealWorldQA) and three MLRMs: InternVL-3.5-8B (first row), Qwen-3-VL-8B (second row), and SAIL-VL2 (third row). The star marker (☆) denotes standard thinking—the baseline with no additional test-time compute. Parallel thinking [4, 9] generates multiple parallel chains-of-thought without visual refocusing. Across all models and benchmarks, VisRef consistently achieves superior accuracy for any given computational budget.

The first term aggregates individual token relevances to the reasoning context, while the second term, through the normalized kernel determinant, penalizes redundancy and encourages diverse visual coverage. This decomposition provides theoretical justification for why maximizing $\log\det(L_k^{V_k})$ naturally balances both objectives.

H. Computational cost analysis.

Table 2 reports detailed latency measurements (on 1 H100 GPU) on the Mathvista dataset using InternVL-3.5-8B. On average, our DPP-based token selection adds only 0.5 secs of overhead compared to Textual self-reflection (TSR), and 1.1 secs compared to standard thinking (ST). Note that ST does not include self-reflection, so it is faster than others. The efficiency of VisRef stems from greedy approximation of Eq. 11.

Method	Time
ST	7.1s
TSR	7.7s
Look-Back [1]	7.6s
VisRef	8.2s

Table 2. Latency per prompt on MathVista.

I. Generalization Across Model Scales

We study whether VisRef consistently improves multi-step visual reasoning as the backbone model scales. Specifically,

Model	ST	TSR	VisRef
InternVL-1B	46.1	48.5	**52.0**
InternVL-2B	52.9	53.7	**58.1**
InternVL-8B	68.1	73.9	**79.3**

Table 3. Accuracy (%) across model scales on MathVista.

Selection	MVista	MVision	MM-Star
Random	67.3	40.8	57.3
Relevance-only	75.6	43.3	61.0
DPP (Ours)	**79.3**	**44.6**	**63.1**

Table 4. Token selection strategies (InternVL-8B).

we evaluate InternVL models spanning 1B, 2B, and 8B parameters under the same decoding setup and token budget. Table 3 shows that VisRef yields gains over both Standard Thinking (ST) and Textual Self-Reflection (TSR) at every scale: for the 1B model, VisRef improves accuracy from 46.1% (ST) and 48.5% (TSR) to 52.0%; for 2B, it increases accuracy from 52.9%/53.7% to 58.1%; and for 8B, it improves performance from 68.1%/73.9% to 79.3%. These results suggest that the benefit of visual refocusing is not confined to a particular parameter regime, but instead persists from small to larger models, indicating that VisRef effectively counteracts visual token dilution during extended reasoning across model capacities.

J. Random Sampling Baseline

We further verify that the improvement of VisRef is not merely due to selecting *any* subset of visual tokens under a fixed budget. To this end, Table 4 compares three selection strategies on InternVL-8B: (i) *Random* selection, (ii) *Relevance-only* selection that greedily keeps the most text-aligned tokens, and (iii) our DPP-based selection that jointly optimizes relevance and diversity. Random selection performs close to the ST baseline and is substantially worse than VisRef across all benchmarks, indicating that naive token subsampling fails to preserve the visual evidence needed for multi-step reasoning. Relevance-only selection improves over random sampling, but it remains consistently below DPP (Ours), suggesting that selecting only the most aligned tokens can still be redundant (e.g., repeatedly focusing on similar regions) and may miss complementary evidence elsewhere in the image. By explicitly encouraging diversity in addition to relevance, DPP (Ours) achieves the best results, supporting our claim that balancing relevance and diversity is essential for effective visual refocusing under tight token budgets.

K. Weighted Version of Eq. 10

We additionally experimented with a weighted objective of the form $\lambda \cdot \text{relevance} + (1-\lambda) \cdot \text{diversity}$. As shown in Table 5, performance peaks at λ=0.5 across both MathVista and MathVision, indicating that balancing relevance and diversity is important in practice. This result supports our default (unweighted) formulation in the main paper, where the two terms contribute equally.

λ	0.0 (Div)	0.25	0.5 (Ours)	0.75	1.0 (Rel)
MVista	71.2	76.8	**79.3**	77.4	75.6
MVision	41.5	43.1	**44.6**	44.2	43.3

Table 5. Effect of λ weighting (InternVL-8B).

Figure 2. **Attention Visualization.** Attention maps show how VisRef progressively refocuses on relevant visual regions during multi-step reasoning. Initially, the attention maps are noisy. With visual reinjection, VisRef reinforces grounding on task-critical objects, leading to more accurate visual reasoning.

References

[1] Manoj Acharya, Kushal Kafle, and Christopher Kanan. Tallyqa: Answering complex counting questions. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8076–8084, 2019. 1

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1

[3] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024. 1

[4] Soumya Suvra Ghosal, Souradip Chakraborty, Avinash Reddy, Yifu Lu, Mengdi Wang, Dinesh Manocha, Furong Huang, Mohammad Ghavamzadeh, and Amrit Singh Bedi. Does thinking more always help? understanding test-time scaling in reasoning models. *arXiv preprint arXiv:2506.04210*, 2025. 2, 3

[5] 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. *arXiv preprint arXiv:2310.02255*, 2023. 1

[6] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025. 1, 2

[7] 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. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. 1

[8] 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. 1

[9] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*, 2022. 2, 3

[10] xAI. Realworldqa, 2024. 1, 2

[11] Weijie Yin, Yongjie Ye, Fangxun Shu, Yue Liao, Zijian Kang, Hongyuan Dong, Haiyang Yu, Dingkang Yang, Jiacong Wang, Han Wang, et al. Sail-vl2 technical report. *arXiv preprint arXiv:2509.14033*, 2025. 1