When More Thought Hurts: Calibrated Early Stopping for Multimodal Reasoning

Anonymous Authors

Anonymous Institution

Abstract. Longer test-time reasoning is not monotonically beneficial for multimodal large language models (MLLMs): under a shared-prefix checkpoint protocol, an answer that is correct early can become wrong after additional reasoning. We call these transitions *harmful flips* and use them to study when extra inference has positive marginal value. Across VAPO-Thinker-3B and OpenVLThinker-7B, checkpoint oracles reveal substantial per-example headroom, but this headroom does not imply that a learned router will transfer. We therefore evaluate a conservative, uncertainty-aware early-stopping family under locked target-benchmark protocols. For OpenVL, frozen routers remain non-inferior to 512-token inference while reducing mean reasoning tokens by 68.8% on MMBench and 58.45% on SEED-Bench. The same approach yields little adaptive benefit for VAPO, and on A-OKVQA a fixed 32-token budget is already Pareto-optimal, exposing an important calibration boundary. A separate visual-intervention audit shows that original images remain causally useful relative to blank or counterfactual images, yet prompt-level re-view does not reliably repair harmful trajectories. Our conclusion is deliberately narrow: checkpoint instability creates exploitable heterogeneity, but safe savings require model- and distribution-calibrated stopping; neither uniformly longer reasoning nor unconditional visual refresh is a reliable policy.

Keywords: Multimodal reasoning · Vision-language models · Adaptive inference · Test-time compute · Visual grounding

1 Introduction

Chain-of-thought (CoT) prompting and inference-time search have established a powerful scaling principle for language models: additional computation at test time can substantially improve difficult reasoning problems [17,16,13]. This principle is increasingly important for multimodal large language models (MLLMs), where structured reasoning and test-time search can improve visual question answering and mathematical reasoning [7,18]. Yet the usual scaling intuition hides a critical assumption: that an additional reasoning step has non-negative value for the current example. Recent multimodal evidence challenges that assumption. Longer traces can weaken perceptual grounding or reverse an already correct answer [15,14].

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We study this failure at the *instance level*. Across shared-prefix checkpointed trajectories, we repeatedly observe *harmful flips*: examples whose answer readout is correct at an early checkpoint but incorrect at the 512-token checkpoint. We use this operational definition throughout; because the checkpoints share a generated prefix, it measures trajectory instability rather than the causal effect of independently rerunning the model at each budget. On locked A-OKVQA evaluation [12], VAPO-Thinker-3B falls from 83.75% at 32 tokens to 82.75% at 512 tokens, while OpenVLThinker-7B falls from 87.25% to 85.88%. The corresponding harmful-flip rates are 3.12% and 4.13%. At the same time, checkpoint oracles reach 86.50% and 91.25%, respectively. Thus, the central problem is not simply that reasoning is too long: some examples benefit from continuation while others are damaged by it. The missing capability is to estimate the *marginal value of the next reasoning step*.

Multimodal inference introduces a second source of risk. As the textual trace grows, it can become an increasingly dominant autoregressive context even when the answer still depends on the image. Visual-forgetting studies show that reconditioning on visual information can improve grounding [14], while visually cued reasoning and visual test-time scaling demonstrate that targeted visual evidence can repair specific perception or grounding failures [11,10]. These findings motivate a natural alternative to simply thinking longer: look again. However, whether visual refresh is a generally safe repair for a trajectory that is already drifting remains unresolved.

We therefore ask two questions. **(Q1)** Can a frozen, label-free early-stopping controller reduce reasoning cost while remaining non-inferior to a long fixed budget? **(Q2)** Is visual refresh a safe repair when a trajectory appears to drift? Unlike work that trains a globally shorter reasoning style [20,21], our deployable component is a conservative checkpoint router that estimates whether further computation is useful for the current instance. Re-view is evaluated separately as a mechanism audit rather than presented as a deployed action. We use locked target-benchmark protocols and retain negative transfer results.

The resulting picture is deliberately bounded. Conservative routing works well for OpenVL on two locked tests: on MMBench [8] it retains 88.75% versus 89.00% for fixed-512 while reducing mean reasoning by 68.8%; on SEED-Bench [5] it obtains 76.375% versus 76.25% with 58.45% fewer tokens. However, the same strategy does not yield useful adaptive behavior for VAPO, and on A-OKVQA the 32-token baseline is already Pareto-optimal. Controlled interventions further show that real images matter causally—blank and counterfactual images are damaging—but prompt-level re-view is not a reliable cross-model repair.

Our contributions are threefold and intentionally separate deployable results from diagnostic evidence:

- We establish harmful flips as a reproducible form of shared-prefix checkpoint instability across two reasoning MLLMs and multiple benchmarks, and quantify the gap between fixed budgets and per-example checkpoint oracles.

- We evaluate conservative adaptive stopping under locked transfer. The results demonstrate substantial accuracy–compute gains for OpenVL while exposing a clear cross-model calibration boundary for VAPO.
- We conduct a separate visual-intervention audit using original, blank, and counterfactual images plus two refresh variants. It separates *causal visual dependence* from *successful repair*: visual evidence remains important, but re-inserting it is not universally beneficial.

The central conclusion is therefore not that "long reasoning is bad." Rather, the utility of additional multimodal reasoning is heterogeneous. Safe inference requires calibrated control over when to stop, when to continue, and when additional visual evidence is actually useful.

2 Related Work

2.1 Multimodal reasoning and test-time scaling

CoT prompting [17] and self-consistency [16] showed that explicit intermediate reasoning and inference-time sampling can improve language-model reasoning. Process verification provides a complementary route by scoring intermediate steps rather than only final outcomes [6], and compute-optimal test-time scaling further shows that the best allocation of inference compute depends strongly on problem difficulty [13]. In vision-language modeling, visual instruction tuning established a general-purpose multimodal assistant paradigm [7]; more recent LLaVA-CoT combines structured stages of visual interpretation and reasoning with stage-wise retracing search, directly demonstrating multimodal test-time scaling [18].

Our focus is orthogonal to making search deeper or traces more structured. We ask whether a *particular ongoing trajectory* should receive more computation at all. This distinction matters because the same additional budget can produce beneficial flips on some examples and harmful flips on others. The relevant quantity is therefore conditional marginal utility rather than reasoning length alone.

2.2 Visual forgetting, grounding, and re-conditioning

The closest failure mode is visual forgetting. Tian et al. show that longer multimodal reasoning can improve logical inference while degrading basic visual recognition, and introduce Vision-Anchored Policy Optimization to favor visually grounded trajectories [15]. Take-along Visual Conditioning diagnoses declining visual influence during long CoT and strategically reintroduces compressed visual tokens at critical stages [14]. Remember-R1 instead introduces process-level reinforcement signals intended to preserve visual dependence over long contexts [1]. Related evidence from Forgotten Polygons shows that strong MLLMs can remain surprisingly weak at elementary shape grounding and that explicit visual cues can dramatically improve selected reasoning tasks [11]. RegionFocus

similarly uses selective visual zooming as test-time scaling for GUI grounding [10]. Listener-rewarded reasoning provides another example of supervising the reliability of multimodal reasoning traces rather than only their final answers [4].

These studies establish that visual evidence can be lost, reweighted, or usefully reacquired. Our intervention question is stricter: after a harmful trajectory has begun, does an original-image re-view *causally repair* the answer relative to simply continuing? By comparing original, blank, and category-matched counterfactual images, we separate dependence on visual evidence from the stronger claim that visual refresh is a safe correction mechanism.

2.3 Adaptive reasoning and inference-time compute allocation

Adaptive reasoning aims to allocate computation according to input difficulty rather than applying one budget to every example. IBPO formulates reasoning as utility maximization under an inference-budget constraint [20]. MARS extends the adaptive-compute objective to multimodal reasoning, learning to skip or retain reasoning steps according to question difficulty and reporting large reductions in CoT length [21]. Compute-optimal scaling likewise argues for prompt-dependent allocation rather than a uniform inference budget [13].

Our setting differs in two ways. First, the controller operates *online*: it observes answer stability, confidence, and grounding signals from the evolving trace, rather than assigning a budget only from the initial question. Second, our evaluation treats cross-model failure as part of the result. A useful controller must not only save tokens on average; it must preserve paired accuracy under locked transfer. This motivates conservative uncertainty-aware stopping rather than aggressive early exit.

2.4 Multimodal evaluation

Reliable claims about adaptive reasoning require benchmarks with complementary visual and reasoning demands. MMBench provides fine-grained capability evaluation for large VLMs [8]; SEED-Bench offers human-annotated multiple-choice evaluation across diverse multimodal dimensions [5]; and A-OKVQA requires commonsense and world knowledge grounded in images [12]. Our development suite additionally includes MathVista for mathematical reasoning in visual contexts [9], MMMU for expert-level multi-discipline reasoning [22], and MM-Star, which was designed to reduce text-only solvability and benchmark leakage [2]. Broader evaluation efforts such as MME and MM-Vet emphasize the need to test both perception and integrated multimodal capabilities rather than a single aggregate skill [3,19]. We use this diversity not to claim universal superiority, but to test whether the same inference-control principle survives changes in model and target distribution.

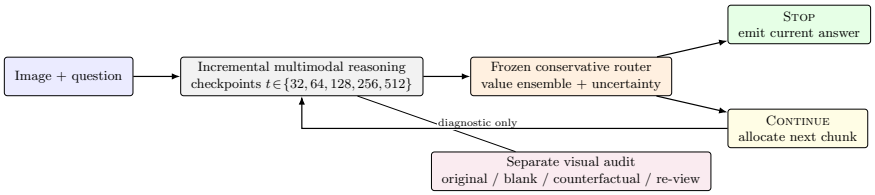

Fig. 1. Evaluation design. The deployable component incrementally generates multimodal reasoning and applies a binary early-stopping router at predefined checkpoints. The solid feedback arrow denotes online continuation to the next reasoning chunk. The dashed, arrow-free connection marks a separate diagnostic visual audit; re-view is not an action of the deployed router. Checkpoint answers are evaluated with a shared-prefix protocol, as detailed in Sec. 4.

3 Calibrated Early Stopping

3.1 Checkpoint instability and marginal utility

Let $x = (I, q)$ denote an image-question pair. A base MLLM produces a single reasoning trajectory from which we read candidate answers at checkpoints $t \in \mathcal{T} = \{32, 64, 128, 256, 512\}$. Let $C_t \in \{0, 1\}$ denote correctness of the checkpoint answer. We define a harmful flip relative to the final checkpoint as $C_t = 1, C_{512} = 0$ for any earlier t. Because all checkpoints share a prefix, this definition quantifies *within-trajectory answer instability*; it does not by itself establish that independently generated long-budget inference causally destroys an answer.

The oracle checkpoint accuracy,

$$A_{\text{oracle}} = \frac{1}{N} \sum_{i=1}^{N} \max_{t \in \mathcal{T}} C_{i,t}, \tag{1}$$

is used only to quantify heterogeneity. The deployable problem is harder: at checkpoint t, choose Stop or Continue without observing C_t or future labels. We define the latent marginal utility of continuation as

$$\Delta_t = \mathbb{E}[C_{t'} - C_t \mid s_t] - \lambda \, \mathbb{E}[N_{t'} - N_t \mid s_t], \tag{2}$$

where t' is the next checkpoint, s_t contains label-free trajectory features, and λ encodes compute cost. The ideal policy continues only when $\Delta_t > 0$.

3.2 Conservative value routing

Our router estimates separate correctness values for stopping and continuing from label-free state features such as answer consistency across nearby checkpoints, predictive uncertainty, trajectory drift, and image-ablation sensitivity. Five independently fitted predictors provide ensemble means $\mu_{\text{stop}}, \mu_{\text{cont}}$ and uncertainties $\sigma_{\text{stop}}, \sigma_{\text{cont}}$. To avoid aggressive early exits under distribution shift,

we stop only when

$$\mu_{\text{stop}} - \beta\sigma_{\text{stop}} \geq \mu_{\text{cont}} + \beta\sigma_{\text{cont}} - \epsilon, \tag{3}$$

otherwise we continue. Here β and the non-inferiority tolerance ϵ are frozen using development data before target labels are revealed. This asymmetry is deliberate: uncertainty spends more compute rather than silently claiming a safe exit.

The method is a *router family*, not a single universal calibration. Development proceeds in stages, producing frozen snapshots V2 and V3 before their respective untouched target evaluations. We therefore do not pool MMBench and SEED-Bench as if one identical parameter vector had seen both targets; instead, each target is a locked test of the snapshot that existed before that target's labels were inspected. Cross-model failures are reported rather than repaired post hoc.

3.3 Visual intervention audit

Visual refresh is evaluated separately from the final stopping policy. We compare continuation with original, blank, and category-matched counterfactual images, plus fresh visual re-reading and visual-evidence patches. These interventions test whether the trajectory still depends on image evidence and whether reintroducing that evidence repairs it; negative re-view results are not folded into the deployed method.

3.4 Metrics and statistical protocol

We report accuracy, mean and tail reasoning tokens, harmful-flip rates, and checkpoint-oracle headroom. Target-benchmark comparisons are paired at the example level. Non-inferiority is assessed against fixed-512 with a margin of -1.5 percentage points, fixed before target-label reveal (12 errors per 800-example test), using paired bootstrap confidence intervals; exact McNemar tests summarize changed predictions. The margin represents an explicit accuracy budget for compute reduction, not evidence of accuracy improvement. A controller is considered useful only when it clears this margin *and* yields material compute savings relative to the relevant fixed baseline.

4 Experiments

4.1 Setup and development diagnostics

We evaluate VAPO-Thinker-3B and OpenVLThinker-7B with reasoning checkpoints $\{32, 64, 128, 256, 512\}$. A *harmful flip* is an example whose shared-prefix checkpoint answer is correct earlier and wrong at 512 tokens; we avoid interpreting this diagnostic as an independent-budget causal effect. Development diagnostics span MMStar [2], which emphasizes genuinely vision-dependent evaluation, MathVista [9], which stresses mathematical reasoning in visual contexts, and a locked MMMU split [22], which covers expert-level multi-discipline problems.

Table 1. Development diagnostics. Oracle values use labels and quantify opportunity rather than deployable performance. HF is the early-correct/final-wrong rate.

Benchmark	Model	Best fixed	HF	Oracle	Oracle gain
MMStar (300)	VAPO	53.00	12.33	65.33	+12.33
MMStar (300)	OpenVL	22.00	11.33	33.00	+11.00
MathVista (140)	VAPO	70.00	10.71	85.00	+15.00
MathVista (140)	OpenVL	75.71	5.00	84.29	+8.58
MMMU (168)	VAPO	51.79	12.50	71.43	+19.64
MMMU (168)	OpenVL	54.76	7.74	70.24	+15.48

Final claims rely on untouched MMBench-EN [8] and SEED-Bench [5] tests, with A-OKVQA [12] used as an additional locked boundary diagnostic requiring image-grounded commonsense and world knowledge. This benchmark mix follows the broader lesson of MME and MM-Vet that perception and integrated reasoning should be evaluated across heterogeneous capability axes [3,19]. The controller is trained without target labels and frozen before final-label reveal. Policy comparisons use paired bootstrap confidence intervals and exact McNemar tests.

Table 1 shows the stable mechanism-level result: average accuracy may rise with budget, yet 5–12.5% of examples can lose an earlier correct answer, while checkpoint oracles remain 8.6–19.6 points above the best fixed budget. This motivates instance-wise control without supporting the stronger claim that longer reasoning is globally harmful. Visual interventions sharpen the interpretation. On MMStar harmful-flip cohorts, immediate stopping is strongest for both models; category-matched counterfactual images reduce accuracy relative to stopping by 26.5 points for OpenVL and 16.2 points for VAPO. Original-image re-view does not reliably beat a text-only continuation. On MathVista, direct transfer of the MMStar router is also worse than fixed-512, revealing a calibration problem rather than a universal controller. These negative development results motivate the conservative frozen policy evaluated below.

4.2 Final frozen evaluation on MMBench-EN

The MMMU outcomes motivate the conservative rule in Eq. (3). Snapshot V2 uses five-member value ensembles fitted on MMStar, MathVista, and the then-available MMMU development data with leave-one-benchmark-out diagnostics. At inference, uncertainty defaults to continuation; stopping requires conservative dominance of the stop value. Re-view was explored during development but disabled before final evaluation because it failed the safety criterion. All V2 weights and thresholds are frozen before examining MMBench-EN outcomes. Benchmark normalization uses only unlabeled state features.

We select 800 visual four-option examples from MMBench-EN by deterministic round-robin sampling over category and source. Both models use the same

Table 2. Final frozen MMBench-EN results ($n = 800$ per model). NI denotes whether the paired 95% CI lower bound exceeds the pre-specified -1.5 point non-inferiority margin.

Model	Policy	Acc.	Avg. tok.	Difference	95% CI	NI
VAPO	Fixed-512	86.75	512.00	–	–	–
VAPO	Frozen V2	86.75	511.68	0.00	[0.00, 0.00]	Yes
OpenVL	Fixed-512	89.00	512.00	–	–	–
OpenVL	Frozen V2	88.75	159.68	−0.25	[−0.875, 0.375]	Yes

shared 512-token trajectory, checkpoints, and constrained A/B/C/D action decoding as the development evaluations. We pre-specify a paired non-inferiority margin of -1.5 percentage points and use 5,000 paired bootstrap resamples for confidence intervals.

Table 2 provides the strongest deployable-policy result. For OpenVL, the frozen controller reduces average reasoning tokens by 68.8% while its paired accuracy difference is -0.25 points; the confidence interval lies entirely above the non-inferiority margin. It changes six outcomes relative to fixed-512 (four fixed-only versus two router-only; McNemar $p = 0.688$), consistent with matched rather than improved accuracy. Harmful flips decrease slightly from 2.50% to 2.25%. The 90th-percentile budget remains 512 tokens, showing that the policy preserves long reasoning for a difficult tail. For VAPO, the frozen controller exactly matches fixed-512 predictions and accuracy but averages 511.68 tokens, so it is statistically non-inferior without delivering a meaningful compute saving. The cross-model contrast is important: conservative uncertainty prevents the catastrophic early-stop collapse observed on MMMU, but useful compute reduction depends on calibrated separation between stopping and continuation values.

4.3 Second untouched evaluation on SEED-Bench

After the MMBench evaluation was closed, development continued only on previously available development resources, producing snapshot V3; no MMBench labels are used as V3 training targets or threshold-tuning labels. We then freeze V3 and apply it to a deterministic 800-example SEED-Bench manifest for each model. SEED-Bench was not used for training, leave-one-benchmark-out development, fallback selection, or threshold calibration. The only target-domain operation is the label-free batch feature normalization explicitly fixed before evaluation. All five stop checkpoints and original-image re-view outcomes use the same constrained A/B/C/D decoding protocol. We retain the pre-specified -1.5 point non-inferiority margin and compute 50,000 paired bootstrap resamples. Data hashes are `2bb9d817...e55f53` (VAPO) and `bc003e79...f9c5b` (OpenVL).

Table 3 independently confirms an OpenVL accuracy–compute improvement: the frozen controller reduces mean reasoning by 58.45%, selects checkpoints 32/64/128/256/512 for 179/102/170/166/183 examples, and changes only one

Table 3. Frozen V3 routing on untouched SEED-Bench ($n = 800$ per model). Differences are against fixed-512. NI uses the pre-specified -1.5 point margin.

Model	Policy	Acc.	Avg. tok.	Difference	95% CI	NI
VAPO	Fixed-512	72.000	512.00	–	–	–
VAPO	Frozen V3	73.875	32.00	+1.875	[0.000, 3.750]	Yes
OpenVL	Fixed-512	76.250	512.00	–	–	–
OpenVL	Frozen V3	76.375	212.76	+0.125	[0.000, 0.375]	Yes

Table 4. SEED-Bench trajectory and original-image re-view diagnostics. Oracle values use labels and are opportunity bounds, not method results. Re-view differences are relative to stopping at 512 tokens.

Model	32 acc.	512 acc.	HF	BF	Stop oracle	Re-view Δ
VAPO	73.875	72.000	38	23	77.250	+1.500
OpenVL	75.750	76.250	31	35	81.250	−0.250

fixed-512 error to correct (two-sided exact McNemar $p = 1.0$). Its interval clears the non-inferiority margin but should not be interpreted as evidence of an accuracy gain because the lower bound is zero. The 90th-percentile budget remains 512. For VAPO, the frozen policy chooses its pre-specified 32-token fallback for all 800 examples. It is non-inferior to fixed-512 and saves 93.75% of reasoning tokens, but this is global checkpoint transfer rather than successful instance-adaptive routing; relative to its own frozen fallback, the router changes neither predictions nor compute.

The diagnostic results sharpen the mechanism claim without making re-view universal (Table 4). VAPO has 38 early-correct/final-wrong harmful flips versus 23 beneficial flips; its stop-checkpoint oracle exceeds the best fixed budget by 3.375 points (95% CI $[2.125, 4.625]$). OpenVL has 31 harmful and 35 beneficial flips, and its oracle gain is 4.750 points (CI $[3.375, 6.250]$). At 512 tokens, original-image re-view recovers 29 VAPO errors and damages 17 correct answers, yielding +1.50 points, but the paired interval $[-0.125, 3.250]$ crosses zero ($p = 0.104$). For OpenVL it recovers six and damages eight, yielding -0.25 points (CI $[-1.125, 0.625]$; $p = 0.791$). Thus re-view has causal recovery capacity on individual examples, yet unconditional acceptance is not statistically reliable; a selective gate remains necessary.

4.4 Locked A-OKVQA diagnostic: the boundary becomes sharper

We finally evaluate 800 A-OKVQA examples per model using the same five checkpoints. This independent locked diagnostic is important because it distinguishes a genuine adaptive-compute gain from a case where a short global budget is already sufficient. Table 5 reports the fixed-budget endpoints and oracle headroom.

Table 5. Locked A-OKVQA diagnostic ($n = 800$ per model). HF denotes correct-at-32 but wrong-at-512 harmful flips. The oracle selects the best checkpoint using labels and is an upper bound, not a deployable method.

Model	Acc@32	Acc@512	HF	Oracle	Gain
VAPO-3B	83.75	82.75	3.12	86.50	+2.75
OpenVL-7B	87.25	85.88	4.13	91.25	+4.00

For VAPO, 25 examples flip from correct to wrong by 512 tokens while 17 move from wrong to correct; for OpenVL the counts are 33 and 22. Thus the aggregate degradation is explained by asymmetric instance-level transitions rather than uniformly poor late predictions. Importantly, the frozen router does not recover the oracle headroom on this benchmark: its deployable Pareto point collapses to the fixed 32-token solution for both models. This negative result prevents overclaiming. Adaptive routing can reduce compute on some target distributions, but a short fixed budget can dominate when the benchmark already favors early answers.

The visual intervention study further separates grounding from repair. Forced original-image re-view has only small, statistically non-significant net effects across checkpoints. In matched pre- and post-flip interventions, real images consistently outperform blank or category-matched counterfactual images, demonstrating that the visual input is causally consequential. However, real-image review does not reliably outperform simple continuation. Fresh Visual Re-read, which removes the prior reasoning trace, remains below continuation on error cohorts, and a Visual Evidence Patch that appends independently extracted visual facts yields only +1.67 points over continuation for VAPO while decreasing OpenVL by 5.00 points and harming stable-correct examples. We therefore treat these interventions as mechanism diagnostics, not as a deployable repair policy.

4.5 Cross-benchmark conclusion

Across the experiments, three findings remain after the controls available in the present data. First, shared-prefix checkpoint answers are non-monotonic and checkpoint oracles expose substantial per-example headroom; this is evidence of exploitable heterogeneity, not yet proof that independently sampled long-budget runs causally degrade. Second, conservative early stopping can convert that heterogeneity into a practical accuracy–compute trade-off for OpenVL: two separately frozen snapshots clear the same pre-specified non-inferiority criterion on two untouched targets while saving substantial mean tokens. This is replication of a routing principle, not evidence that one universal calibration transfers everywhere. VAPO and A-OKVQA provide explicit counterexamples. Third, visual evidence is causally relevant under image interventions, but visual refresh is not equivalent to repair. Together, the results support calibrated early stopping while ruling out the stronger prescriptions that one should always think longer, always stop early, or always look again.

5 Conclusion

We studied multimodal reasoning through the marginal value of additional inference. Shared-prefix checkpoint answers can reverse from correct to incorrect even when aggregate accuracy is flat or improving, and oracle checkpoint selection reveals substantial heterogeneity. Conservative early stopping converts part of this heterogeneity into large OpenVL compute savings under locked non-inferiority tests, but VAPO and A-OKVQA show that the calibration is not universal. Visual interventions further show that image evidence remains causally relevant without making re-view a safe repair. The supported claim is therefore narrower than "less reasoning is better": multimodal test-time compute can be reduced safely only when stopping is calibrated to the model, distribution, and uncertainty of the current trajectory.

6 Limitations and Reproducibility Notes

MMBench-EN and SEED-Bench are one-time locked tests for different frozen snapshots (V2 and V3): their revealed labels are not used to retune the snapshot reported on that target. Because the snapshots differ, the two results replicate a controller design principle rather than demonstrate that one identical router transfers across both benchmarks. The final results establish non-inferiority and compute reduction for OpenVL, not universal effectiveness or accuracy improvement. VAPO obtains negligible MMBench savings and falls back globally to 32 tokens on SEED-Bench; neither behavior demonstrates useful instance-adaptive routing. Neither frozen policy safely accepts re-view, and the positive VAPO SEED-Bench re-view difference is statistically inconclusive. Benchmark-aware normalization is transductive but label free; future evaluation should test a source-only or small streaming calibration alternative. The largest validity threat is the shared-prefix checkpoint protocol: early answers are read from one 512-token trajectory rather than from independently sampled fixed-budget runs. Independent generations at each budget are therefore required before making a causal claim that longer inference itself creates harmful flips. Length-matched text and duplicated-context controls are also necessary to separate visual refresh from extra-context effects. Finally, the non-inferiority claims use a pre-specified -1.5 point margin and paired bootstrap intervals; they should not be interpreted as evidence that the router improves accuracy.

References

1. Chen, J., Tang, J., Wei, W., Xu, X., Wu, J., Liu, Z., Zhou, Q., Yan, Y., Geng, B., Xia, Y., Zhang, L., Chen, Q.: Remember-r1: Mitigating long-context visual forgetting through reinforcement learning. arXiv preprint arXiv:2608.01314 (2026)
2. Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Wang, J., Qiao, Y., Lin, D., Zhao, F.: Are we on the right way for evaluating large vision-language models? In: Advances in Neural Information Processing Systems. vol. 37 (2024)

3. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., Wu, Y., Ji, R.: Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394 (2023)
4. Gambashidze, A., Li, P., Skripkin, M., Galichin, A., Gusarov, A., Sobolev, K., Kuznetsov, A., Oseledets, I.: Listener-rewarded thinking in vlms for image preferences. arXiv preprint arXiv:2506.22832 (2025)
5. Li, B., Ge, Y., Ge, Y., Wang, G., Wang, R., Zhang, R., Shan, Y.: Seed-bench: Benchmarking multimodal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13299–13308 (2024)
6. Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., Cobbe, K.: Let's verify step by step. In: International Conference on Learning Representations (2024)
7. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Advances in Neural Information Processing Systems. vol. 36 (2023)
8. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., Chen, K., Lin, D.: Mmbench: Is your multi-modal model an all-around player? In: European Conference on Computer Vision. pp. 216–233. Springer (2024)
9. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In: International Conference on Learning Representations (2024)
10. Luo, T., Logeswaran, L., Johnson, J., Lee, H.: Visual test-time scaling for gui agent grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19989–19998 (2025)
11. Rudman, W., Golovanevsky, M., Bar, A., Palit, V., LeCun, Y., Eickhoff, C., Singh, R.: Forgotten polygons: Multimodal large language models are shape-blind. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 11983–11998 (2025). https://doi.org/10.18653/v1/2025.findings-acl.620
12. Schwenk, D., Khandelwal, A., Clark, C., Marino, K., Mottaghi, R.: A-okvqa: A benchmark for visual question answering using world knowledge. In: European Conference on Computer Vision (2022). https://doi.org/10.1007/978-3-031-20074-8_9
13. Snell, C.V., Lee, J., Xu, K., Kumar, A.: Scaling llm test-time compute optimally can be more effective than scaling parameters for reasoning. In: International Conference on Learning Representations (2025)
14. Sun, H.L., Sun, Z., Peng, H., Ye, H.J.: Mitigating visual forgetting via take-along visual conditioning for multi-modal long cot reasoning. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics. pp. 5158–5171 (2025). https://doi.org/10.18653/v1/2025.acl-long.257
15. Tian, X., Zou, S., Yang, Z., He, M., Waschkowski, F., Wesemann, L., Tu, P.H., Zhang, J.: More thought, less accuracy? on the dual nature of reasoning in vision-language models. In: International Conference on Learning Representations (2026)
16. Wang, X., Wei, J., Schuurmans, D., Le, Q.V., Chi, E.H., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models. In: International Conference on Learning Representations (2023)
17. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q.V., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models. In: Advances in Neural Information Processing Systems. vol. 35 (2022)

18. Xu, G., Jin, P., Wu, Z., Li, H., Song, Y., Sun, L., Yuan, L.: Llava-cot: Let vision language models reason step-by-step. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2087–2098 (2025)
19. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: Evaluating large multimodal models for integrated capabilities. In: Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235, pp. 57730–57754 (2024)
20. Yu, Z., Xu, T., Jin, D., Sankararaman, K.A., He, Y., Zhou, W., Zeng, Z., Helenowski, E., Zhu, C., Wang, S., Ma, H., Fang, H.: Think smarter not harder: Adaptive reasoning with inference aware optimization. In: Proceedings of the 42nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 267, pp. 73304–73325 (2025)
21. Yue, T., Wu, Q., Zhao, D.: Mars: Multimodal adaptive reasoning model for avoiding overthinking. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 34539–34547 (2026). https://doi.org/10.1609/aaai.v40i41.40753
22. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B., Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y., Huang, W., Sun, H., Su, Y., Chen, W.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9556–9567 (2024)