SAMPLING VERSUS LONGER RESPONSES FOR VISUAL REASONING: A MATCHED-PROMPT MATHVISTA STUDY

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual reasoning can receive additional inference budget through longer responses or repeated sampling, but these interventions need not have the same effect. We compare a 512-token single response, a 2,048-token single response, and voting over four 512-token responses on the same 300 multiple-choice MathVista questions, using Qwen2.5-VL-3B and 7B with matched prompts and image preprocessing. Under the fixed explicit-answer scorer, four-response voting improves accuracy by 15.3 and 6.7 percentage points, respectively; paired question-bootstrap (Efron, 1979) 95% intervals are [11.3, 19.7] and [3.7, 10.0]. Increasing the single-response ceiling yields gains of 2.7 and 2.0 points with intervals spanning zero. Actual generated-token counts explain why equal ceilings are not equal compute: four-response voting uses roughly four times as many output tokens, whereas most single responses finish well below either ceiling. A broader, label-blind answer-parser sensitivity and an average-single-path baseline preserve the voting gains. Most newly correct voting outcomes arise on questions where the first response lacks a parsed final answer. A separate analysis of larger checkpoints uses saved predictions; incomplete response records prevent uniform regrading. The findings support repeated sampling for these two checkpoints and this subset; they establish neither a general model-size scaling law nor a compute-efficiency advantage at equal realized cost.

1 INTRODUCTION

Chain-of-thought prompting (Wei et al., 2022), repeated-sample aggregation (Wang et al., 2023), and other inference strategies can alter model performance. Their transfer to visual tasks depends on implementation and task. MathVista (Lu et al., 2024) provides a setting for studying these outcomes. Our primary experiment separates response ceilings from repeated sampling under a common prompt and image processor. A secondary analysis places the result alongside saved predictions from larger checkpoints, whose less complete records impose a narrower interpretation.

We compare accuracy across strategies and examine how answer extraction and measured token use affect the comparison. Complete responses and token counts support these analyses on the fixed multiple-choice subset. It does not identify a causal effect of model size. Earlier outputs are analyzed separately and are not pooled with the new experiment.

2 RELATION TO TEST-TIME INFERENCE METHODS

Reasoning, aggregation, and search. Chain-of-thought and zero-shot reasoning prompts alter the generation procedure (Wei et al., 2022; Kojima et al., 2022); self-consistency aggregates sampled answers (Wang et al., 2023). Our four-response condition applies that established principle. Tree of Thoughts searches over intermediate reasoning states (Yao et al., 2023), and Self-Refine revises a response using generated feedback (Madaan et al., 2023). Neither is equivalent to raising an output ceiling. Process verification (Lightman et al., 2024), compute allocation (Snell et al., 2025), and analyses of verification-based scaling (Setlur et al., 2025) introduce additional selection or learning mechanisms. Our comparison does not test those mechanisms.

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.

Table 1: Fixed 300-question comparison under the original explicit-answer scorer. Tokens are mean generated tokens per question, including all four paths for voting.

Model	Strategy	Accuracy (%)	Output tokens
3B	Single, 512 ceiling	54.7	202.1
3B	Single, 2,048 ceiling	57.3	215.7
3B	Vote, 4×512	70.0	806.0
7B	Single, 512 ceiling	69.7	257.5
7B	Single, 2,048 ceiling	71.7	269.2
7B	Vote, 4×512	76.3	1,027.6

Visual interventions and model training. DDCoT separates multimodal reasoning duties (Zheng et al., 2023); Visual Sketchpad creates intermediate visual representations (Hu et al., 2024); V* explicitly searches image details (Wu & Xie, 2024). A text-only increase in the response allowance does not reproduce these visual operations. Likewise, BLIP-2, InstructBLIP, LLaVA, and InternVL make different choices about visual alignment and instruction tuning (Li et al., 2023; Dai et al., 2023; Liu et al., 2023; Chen et al., 2024). Two Qwen2.5-VL checkpoints (Bai et al., 2025) reduce some implementation differences but do not isolate model size from training. The broader inference study of Ahmadpour et al. (2025) already reports model- and task-dependent gains; we add a narrower analysis of saved paths, parsing, and measured token use.

Benchmark scope. MathVista (Lu et al., 2024) combines visual mathematical tasks. Related benchmarks emphasize scientific explanations (ScienceQA), charts (ChartQA), and geometry (GeoQA) (Lu et al., 2022; Masry et al., 2022; Chen et al., 2021). MMMU and MMBench cover broader multimodal abilities (Yue et al., 2024; Liu et al., 2024); MathVerse varies the information supplied through diagrams and text (Zhang et al., 2024). These distinctions matter: our fixed multiple-choice MathVista subset does not establish gains on those benchmarks, nor on text-only competition mathematics such as MATH (Hendrycks et al., 2021). The contribution is the controlled within-subset comparison, not benchmark breadth.

3 MATCHED-PROMPT COMPARISON

Questions and models. We evaluate Qwen2.5-VL-3B-Instruct and Qwen2.5-VL-7B-Instruct on an identical fixed subset of MathVista testmini. We retain multiple-choice items with two to eight choices, order their IDs by a deterministic hash, reserve the first 32 for development, and evaluate the next 300. This selection is independent of model outputs. It is a multiple-choice subset, not a score on the full 1,000-item benchmark. Prior use of MathVista also prevents interpreting the dataset as untouched across the research project.

Both models receive the same question, letter-coded choices, image, and request to return an explicit final letter. Their processors use minimum and maximum image areas of $4 \cdot 28^2$ and $512 \cdot 28^2$ pixels. Inference uses BF16 and SDPA. Sampling uses temperature 0.7, top-p 0.95, top-k 50, and repetition penalty 1.05. Seeds are fixed by question ID and sample index before generation. Each question has four 512-token paths and one 2,048-token path. The first short path is also the single-short baseline, preserving paired comparisons and making its contribution to the voting condition explicit.

Aggregation and cost. Voting counts valid letters, treats missing answers as abstentions, and breaks ties by the earliest sample among the tied labels. All-missing votes are incorrect. The long-response and four-short-response conditions have the same 2,048-token total output ceiling. They need not consume the same tokens: EOS can end each response early, and voting repeats the input four times. Consequently, this experiment compares strategies under equal output ceilings; it is not an equal-FLOP or equal-latency comparison. We retain prompt and generated-token counts for each path.

Paired effects. Resampling the 300 question blocks 20,000 times gives a voting-minus-short difference of 15.3 percentage points for 3B, with a percentile 95% interval of [11.3, 19.7], and

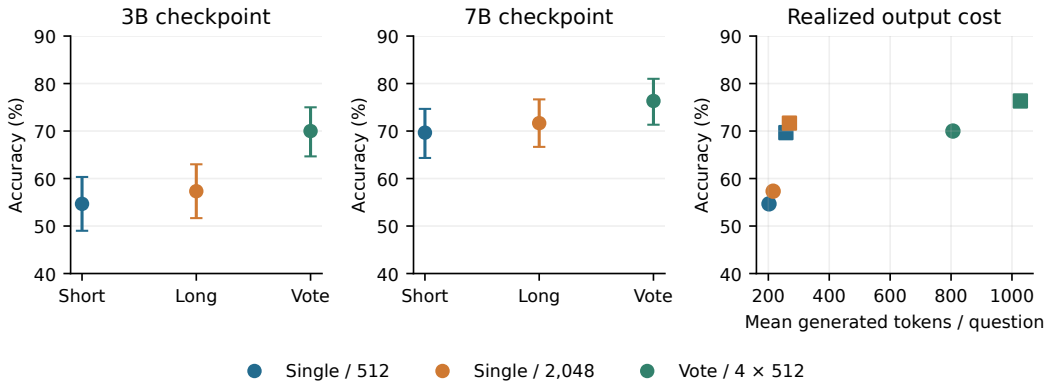


Figure 1: Measured accuracy and generated-token cost on the fixed 300-question subset. Left and middle: marginal 95% question-bootstrap intervals under the original scorer. Right: circles denote 3B and squares 7B; voting consumes substantially more tokens despite sharing the long-response output ceiling. Paired differences are evaluated separately in the text.

6.7 points for 7B, with interval [3.7, 10.0]. The difference between the two checkpoints' voting gains is 8.7 points [3.7, 13.7] under a shared-question bootstrap. This is a contrast between fixed checkpoints, not a randomized intervention on parameter count. Long-minus-short differences are 2.7 points [−3.7, 9.0] and 2.0 points [−2.7, 7.0]. Intervals describe variation over sampled questions, conditional on these checkpoints and saved stochastic paths. They are unadjusted descriptive intervals, not simultaneous guarantees or replication over model training seeds.

The single-long condition increases mean output length by only 13.6 tokens for 3B and 11.7 tokens for 7B. Increasing the ceiling therefore creates little additional realized computation on average. The voting condition instead uses about four times the single-short output tokens. Its advantage is evidence for this repeated-sampling strategy, not evidence that the same amount of actual compute is better allocated to voting than to a longer response.

Answer-parser sensitivity. The fixed scorer requires an explicit terminal letter declaration. Inspection revealed that variants such as "Final answer: A" could fail its grammar. We therefore apply a second, label-blind grammar uniformly to every path, accepting optional "answer", parentheses, and simple emphasis around the final letter. This analysis is post hoc; the original results remain in Table 1. Under the broader grammar, 3B accuracies are 56.7%, 59.3%, and 70.7%, and 7B accuracies are 70.0%, 72.3%, and 76.7%. Voting-minus-short differences remain positive: 14.0 points [10.0, 18.3] and 6.7 points [3.7, 10.0]. Neither parser is the official MathVista model-assisted extraction pipeline.

Missing answers remain part of the outcome. Under the broader grammar, single-short, single-long, and voting produce 58, 37, and 10 missing answers for 3B, and 41, 30, and 16 for 7B. Thus part of the voting improvement can reflect availability of an extractable final answer. Conditioning only on parsed outputs would remove this practical failure mode and change the estimand.

4 What repeated sampling changes

4.1 Sensitivity to the single-response baseline

The primary comparison uses the first short response as its baseline, a choice fixed before examining the confirmation outputs. Nevertheless, one stochastic path need not represent a checkpoint's typical single-response performance. Across the four short-response indices, 3B accuracy ranges from 54.7% to 61.3%; 7B ranges from 66.7% to 69.7%. The first draw is the lowest of the four for 3B and the highest for 7B. This does not invalidate the prespecified contrast, but it affects its magnitude and the comparison between checkpoints.

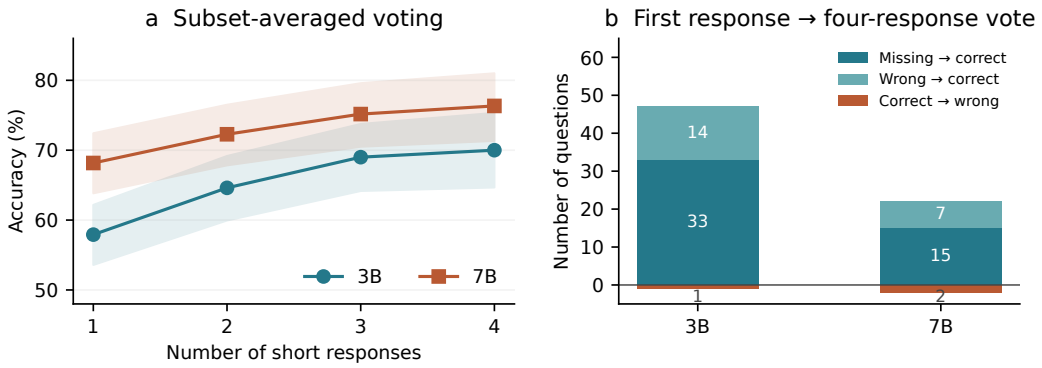


Figure 2: Secondary analyses of the same four short responses per question under the original scorer. (a) Mean accuracy over every size-k subset, with descriptive question-bootstrap intervals. (b) Counts of questions that become correct from a missing or wrong first answer, and correct first answers that become wrong. Recovery from a missing declaration is distinct from correcting an extracted wrong answer.

We therefore conduct a secondary analysis that averages the four single-path correctness indicators within each question. Its estimand is the performance of choosing a path uniformly from the four saved paths, conditional on this generated corpus. It is not the outcome of four independent question datasets, nor does it provide four independent model replications. Relative to this average-single baseline, voting improves 3B by 12.1 points (95% question-bootstrap interval [9.1, 15.1]) and 7B by 8.2 points ([5.6, 10.8]). Thus a voting advantage persists without selecting one privileged response index, while the magnitude of the cross-checkpoint gap is less pronounced. This sensitivity is another reason not to interpret two checkpoint contrasts as a scaling law.

4.2 Within-corpus response-count curve

For a question i, let y_{ij} be the parsed answer from short path j, and let g_i be its reference answer. Define the subset-averaged score

$$a_i(k) = \binom{4}{k}^{-1} \sum_{S \subseteq \{1,2,3,4\}:|S|=k} \mathbf{1}\{\text{vote}(y_{ij} : j \in S) = g_i\}.$$

Every subset retains the original index order for tie-breaking. Missing answers remain abstentions. We average $a_i(k)$ over the 300 questions and bootstrap questions, keeping their entire four-path vector together. The $k = 1$ point is the average-single sensitivity above; $k = 4$ is the primary voting condition. Intermediate points average all subsets and are secondary summaries of existing outputs, not independently executed confirmation experiments.

The resulting 3B curve is 57.9%, 64.6%, 69.0%, and 70.0% for $k = 1, 2, 3, 4$; the 7B curve is 68.2%, 72.3%, 75.2%, and 76.3%. Over these saved paths, the largest increments occur before the fourth response. The fourth response adds about one percentage point to the subset-averaged score while adding another full response's token cost. This observation describes the selected corpus; it does not establish an optimal stopping rule. Choosing k after looking at this curve would require a separate evaluation to estimate the selected policy's performance without selection bias.

The cost curve is linear in k for the subset average because each saved response appears in the same number of subsets. Mean generated tokens per short path are 201.5 for 3B and 256.9 for 7B. This accounting identity does not imply that a deployed system has linear latency: responses might be batched or run concurrently, and each additional response repeats image and prompt processing. We report generated tokens as one reproducible resource measure rather than as a complete serving-cost model.

4.3 Answer availability and correction transitions

A vote can improve an endpoint score by changing an extracted wrong answer, by obtaining a valid answer when the first response has none, or by both kinds of events on different questions. The recorded paths distinguish these possibilities without assigning a latent reasoning mechanism. Under the original scorer, the first short answer is correct, wrong, or missing on 164, 71, and 65 of the 3B questions. Four-response voting preserves 163 of the 164 correct first answers, corrects 14 of the 71 extracted wrong answers, and answers 33 of the 65 missing cases correctly. There is one correct-to-wrong transition. The net increase is therefore $(33 + 14 - 1)/300 = 15.3$ percentage points.

For 7B, the corresponding first-answer counts are 209 correct, 49 wrong, and 42 missing. Voting preserves 207 correct answers, corrects seven wrong answers, and recovers 15 missing cases; two correct answers become wrong. The net gain is $(15 + 7 - 2)/300 = 6.7$ points. Missing-to-correct transitions account for 33 of 47 newly correct 3B outcomes and 15 of 22 newly correct 7B outcomes. This is a substantial component of the score change under the specified response contract.

A missing parsed answer is not evidence that the model had no useful reasoning, and an extracted answer is not evidence that its reasoning was faithful. In particular, a final-letter grammar can miss a semantically clear response. The broader-parser sensitivity reduces this ambiguity but does not eliminate it. The transition table is therefore an endpoint accounting, not a decomposition into perceptual, reasoning, and formatting causes. A causal explanation would require interventions that separately change answer formatting and reasoning while preserving the relevant comparison.

4.4 Ties, abstentions, and the meaning of agreement

Four responses do not always produce a unique plurality. Under the original grammar, 28 of the 300 3B questions and 26 of the 7B questions have more than one answer tied for the largest nonzero vote count. The primary rule resolves such a tie by the earliest sampled path whose answer attains that count. It does not use the reference answer, a verifier score, response length, or a second model. This deterministic rule makes the saved result reproducible, but its dependence on response order merits a sensitivity check.

For this check, we assign each tied winning label equal probability and compute the expected correctness of the resulting vote analytically. If M_i is the set of most frequent nonmissing labels, the score for a question is $\mathbf{1}\{g_i \in M_i\}/|M_i|$; an empty M_i receives zero. This changes only the tie rule on existing paths. It is not a new sampling run, and a fractional expected correct-answer count should not be described as the result of a realized random tie draw. The expected scores are 212.5/300 (70.8%) for 3B and 229/300 (76.3%) for 7B, compared with the primary 210/300 and 229/300. Thus the observed voting gain is not dependent on an unusually favorable first-path tie rule in this corpus.

Abstentions also change what agreement means. There are 12 3B questions and 16 7B questions for which none of the four responses has a parsed answer; the vote remains missing and is scored as a failure. At the other extreme, 194 and 211 questions have exactly one distinct nonmissing label. This includes 18 and 12 questions, respectively, with only one parseable response: that response wins by default, even though the other three paths supply no extracted supporting answer. It would be misleading to call every such case four-path agreement. Our algorithm uses abstention-aware plurality, rather than requiring a strict majority of all four generated paths.

More broadly, agreement is not calibrated confidence. Repeated samples can agree on a wrong answer, share a perceptual error, or omit a necessary step in similar ways. Counting their labels does not make the four answers independent evidence about correctness. The bootstrap therefore keeps their vector together, and our conclusions concern observed endpoint accuracy rather than a confidence guarantee. This distinction matters when using voting inside a larger system: a system that requires a minimum number of valid votes would trade off answer availability and accuracy differently from the rule studied here.

The tie sensitivity changes neither the raw responses nor the fixed-parser result in the primary table. Together with the average-single baseline and broader-parser analysis, it checks three distinct choices: which path represents one call, how text becomes an answer, and how equal vote counts are resolved. The direction of the comparison is stable across these choices. Its transfer to other benchmarks and fresh generations remains untested.

Table 2: Distribution of actual single-response lengths. Ceiling counts indicate outputs whose recorded token count reaches the configured cap; they are not inferred finish-reason labels. Input counts include processor-expanded tokens and are identical for the two single-response prompts.

Model	Ceiling	Median	90th percentile	At ceiling	Mean input
3B	512	167.5	479.7	28/300	317.64
3B	2048	172.5	470.3	0/300	317.64
7B	512	223.0	503.0	30/300	317.64
7B	2048	217.5	489.1	0/300	317.64

Table 3: Correct answers by public MathVista task category under the original scorer. Counts preserve the denominator for each subgroup. S: single short; L: single long; V: four-response vote. All five categories are included.

		3B			7B		
Task	n	S	L	V	S	L	V
Textbook question answering	60	44	44	48	48	52	51
Visual question answering	60	39	40	50	51	50	51
Figure question answering	63	35	44	45	46	49	47
Geometry problem solving	106	41	36	60	55	55	70
Mathematical word problems	11	5	8	7	9	9	10
All selected questions	300	164	172	210	209	215	229

4.5 Why a larger ceiling often changes little

The long condition gives the model permission to generate more tokens but does not require it to do so. Its median length is close to the short condition's median in both checkpoints, and no saved long response reaches 2,048 tokens. Increasing the cap is consequently a weak intervention on average realized output length here. The result should not be generalized to deliberate long-reasoning prompts, models trained to use extended reasoning, or tasks whose solutions normally approach the output limit.

The four-response vote repeats the same input, yielding 1,270.56 mean input tokens, in addition to its larger output-token total. Whether its accuracy gain is worth the additional cost is a deployment question requiring a specified cost function or accuracy target. The present comparisons provide the measured ingredients for that decision but do not claim a universal preferred strategy. They also do not compare against all possible allocations of the same resource budget, such as changing image processing or using a stronger checkpoint for one call.

5 Variation across question types

MathVista groups questions into five task categories (Lu et al., 2024). We recover those public metadata labels by problem ID for the selected questions and report every category in Table 3. The partition was not used to choose questions or tune generation. This is a secondary descriptive analysis of the same 300 questions, not five new benchmark replications. The category counts are uneven: geometry contributes 106 questions, whereas mathematical word problems contribute only 11. Aggregate accuracy consequently weights these tasks according to this selected subset, not equally across task types.

For 3B, voting has a higher point score than the first short response in every category. Geometry accounts for 19 of its 46 net additional correct answers, followed by visual question answering with 11 and figure question answering with 10. For 7B, geometry accounts for 15 of 20 net additional correct answers; visual question answering has the same correct-answer count in the two conditions. These counts locate the observed gain within the evaluated sample. They do not establish a statistically reliable ordering of category-specific effects, particularly for the smallest category.

Longer responses have a different pattern. On 3B, the longer ceiling adds nine correct figure-question answers but loses five correct geometry answers in the aggregate. On 7B, its largest point gain is four textbook-question answers, while visual question answering loses one. Such variation is compatible with a small pooled effect and cautions against assigning the pooled result to every task. Since the single-long and single-short paths use different random seeds, individual transitions also include sampling variation; they are not deterministic continuations of the same short path.

The task labels are not an experimental separation of perception and reasoning. A geometry question can require both reading an image and solving a mathematical relation; a figure question can require substantial reasoning. Thus the concentration of gains on geometry does not show that voting specifically repairs a reasoning component. Likewise, a flat visual-question score at 7B does not prove that visual perception is saturated. Identifying those mechanisms would require interventions beyond category labels.

Statistical unit and interpretation. The question is the resampling unit throughout the controlled analysis. A shared draw is used for contrasts across models; within each question, the relevant paths remain together. Treating the 3,000 paths as independent test questions would inflate the apparent evidence: multiple paths share an image, prompt, and gold answer. The primary intervals quantify variation under empirical question resampling with the recorded generations fixed. They do not include a fresh draw of generation seeds, a different selection of checkpoints, or uncertainty in the benchmark labels. Subset-count curves and task breakdowns are exploratory summaries, and their individual intervals are not simultaneous family-wise guarantees.

Our primary accuracy counts missing declarations as failures. This makes the outcome operationally meaningful for a system that must return one valid option, but it also makes the prompt and parser part of the tested system. The broader grammar and transition analysis expose this dependence rather than eliminating it. An evaluation that uses a semantic judge, allows interactive answer repair, or restricts attention to parsed responses would answer a different question and should report that change explicitly.

6 Earlier larger-checkpoint comparisons

The saved configurations name InternVL2.5-26B, InternVL2.5-78B, and QVQ-72B. Their direct-answer scores under the archived scorer are 41.5%, 70.9%, and 50.9%, respectively. These checkpoint/model comparisons do not isolate scale from pretraining, instruction tuning, family, or baseline capability.

Direct prompting uses a 1,024-token output cap. The CoT condition changes the prompt and raises the cap to 4,096 tokens. Thus its effect cannot be attributed to reasoning text alone or to extra computation alone. Self-consistency uses four or eight sampled responses (temperature 0.7, top-p 0.95, each capped at 1,024 tokens) and votes on extracted answers. It changes randomness and the aggregator as well as call count. A fixed cap is not a measured token count. The archive lacks complete generation lengths and finish reasons, so k calls must not be described as exactly k times the compute or cost.

InternVL uses dynamic image tiling. Raising the tile cap from six to twelve does not necessarily double the actual number of tiles for every image. Its two-pass strategy includes a second image presentation with a larger tile cap and up to 500 characters of the first response. QVQ two-pass instead reuses its image-processing setup, also with a truncated previous-response prefix. These are different intervention bundles; the cross-family two-pass comparison is not a controlled test of the same resolution mechanism.

7 Data, scorer provenance, and pairing

7.1 What is available

For each retained condition there are exactly 1,000 prediction records. The records contain predicted answer, gold answer, answer type, and a correctness flag. Most InternVL non-sampling records retain only the first 300 response characters; self-consistency omits individual sampled answers, and QVQ

Table 4: Saved-scorer strategy-minus-direct differences in percentage points. Each row has 1,000 positions. Intervals are unadjusted paired bootstrap intervals; p_H is the exact paired test adjusted over 12 comparisons.

Model	Strategy	Difference	95% interval	p_H
26B	CoT + larger cap	+17.8	[+14.5, +21.1]	1.31e-23
26B	SC-4	+3.8	[+1.4, +6.2]	0.022
26B	SC-8	+5.8	[+3.7, +8.1]	5.15e-06
26B	Two-pass	+26.1	[+23.2, +29.0]	6.85e-61
78B	CoT + larger cap	+1.5	[-1.1, +4.0]	1
78B	Higher tile cap	+0.2	[-1.1, +1.5]	1
78B	SC-4	-1.7	[-3.9, +0.5]	0.825
78B	SC-8	-1.3	[-3.6, +0.9]	1
78B	Two-pass	-1.8	[-3.6, -0.1]	0.321
QVQ	CoT + larger cap	+13.0	[+9.8, +16.3]	3.06e-13
QVQ	SC-4	+2.9	[+0.6, +5.3]	0.126
QVQ	Two-pass	-0.1	[-3.3, +3.0]	1

records omit response text. Consequently, we can reproduce equality checks on saved predictions but cannot rerun full answer extraction or reconstruct alternative self-consistency votes. Missing response tails are not treated as complete answers in the audit.

The prediction arrays omit problem IDs. We recover the archived metadata file with 1,000 unique IDs and verify exact agreement of gold-answer strings, answer types, and array length for every condition. Positional pairing is still conditional on the generating scripts using that common metadata order; matching repeated gold answers does not independently prove identity. The code iterates the metadata order and can skip missing images. Full-length arrays and row-wise checks are consistent with no skipped item, but explicit question IDs in each run would establish pairing more directly. The analysis conditions on this positional correspondence; explicit per-run IDs are unavailable.

7.2 Why official grading is not established

The archived extractor uses hand-written regular expressions and choice-text rules. It can truncate decimal answers at a period in an answer phrase and mistake the initial capital letter of ordinary choice text for a choice label. It also uses substring fallback rules. The official MathVista extraction code has a separate extraction procedure with a model-based fallback[1]. Reproducing string equality after a custom extraction stage does not establish parity with that full pipeline. We therefore call the values below retained scorer outcomes, not independently validated official benchmark accuracies. Correcting the extraction implementation without complete original responses cannot retroactively correct these observations.

7.3 Paired statistical analysis

For each strategy and model we subtract the direct-answer correctness indicator on the same assumed problem position. We bootstrap whole positions 5,000 times (seed 20260907), preserving paired outcomes. We also compute exact two-sided McNemar tests and Holm-adjust them across all 12 strategy-versus-direct comparisons. Bootstrap intervals are descriptive unadjusted percentile intervals. They cover sampling variability under the empirical problem-resampling model, not scorer errors, pairing uncertainty, model-version uncertainty, or causal confounding. One stochastic generation run is not replication over sampling seeds or model training seeds.

8 Observed contrasts

The CoT-plus-cap contrast differs substantially between the two InternVL checkpoints: the 26B gain minus the 78B gain is 16.3 points, with a descriptive paired bootstrap interval $[12.2, 20.6]$. This

[1]Official source inspected for this audit: `https://github.com/lupantech/MathVista/blob/main/evaluation/extract_answer.py`.

compares effects of the recorded bundles between fixed checkpoints; it is not a randomized capability or size intervention. QVQ's observed +13.0-point contrast is intermediate, but three checkpoint observations do not establish a monotone capability-scaling law.

On 78B, the CoT contrast is +1.5 points with interval $[-1.1, 4.0]$. This leaves both modest benefit and modest harm compatible with the data under the retained scorer; it does not prove zero effect or that the change is "only noise". The higher-tile-cap contrast is +0.2 points with interval $[-1.1, 1.5]$. Its uncertainty and bundled preprocessing do not identify attention allocation as the bottleneck, nor show that image resolution cannot matter elsewhere.

For 78B two-pass, the percentile interval narrowly excludes zero while the exact paired unadjusted test gives $p = 0.0535$ and the Holm-adjusted value is 0.321. These procedures need not be exact inverses, especially for small discrete differences. We report both, do not select the favorable procedure, and make no adjusted significance claim. Likewise, QVQ SC-4's unadjusted interval is positive but its adjusted paired test does not reject at 0.05.

9 Outcome transitions are not mechanistic interference

The saved paired indicators permit a transparent transition count. For 78B SC-4, 58 baseline errors become correct and 75 baseline-correct answers become wrong, giving a net $-17/1000$ difference. The total is an accounting identity:

$$\Delta = \frac{\#(\text{wrong} \to \text{right}) - \#(\text{right} \to \text{wrong})}{1000}.$$

It does not show that textual reasoning overwrites a correct visual perception. The changes can also arise from sampling, different prompts, aggregation, extraction, or truncation. "Interference" can at most name the observed right-to-wrong event; it is not an identified internal mechanism.

Similarly, an error taxonomy based on subjective inspection cannot establish that a fixed percentage of errors is causally perceptual without a validated annotation protocol and appropriate interventions. The available observations do not identify a mechanism or support an architecture-level recommendation. No attention intervention, image-oracle intervention, or matched-compute perception-versus-reasoning decomposition was performed in this audit.

10 Limitations

For the earlier larger-checkpoint comparisons, complete raw responses and individual self-consistency samples are unavailable for official regrading. The new two-checkpoint comparison retains both, but it covers only multiple-choice items and uses a transparent custom scorer. Measured output tokens omit hardware-specific latency, prefill cost, and energy. The earlier cross-family two-pass implementations also differ in image processing, so their effects cannot be attributed solely to scale or resolution.

One task, two related scales, and one additional family do not provide a general scaling law. Prespecified paired comparisons, replication of sampling runs, and held-out tasks would strengthen external validity. If equivalence is the aim, an explicit practical margin and a powered equivalence design are needed; non-rejection in a difference test is insufficient. These requirements are not satisfied by a post-hoc power calculation based on the observed small effects.

11 Conclusion

On the matched MathVista subset, four-response voting improves accuracy for both Qwen2.5-VL checkpoints, whereas a larger single-response ceiling changes little realized computation. The voting gains persist across parser, baseline, and tie-breaking sensitivities, with additional token cost. Larger-checkpoint predictions show model-dependent outcomes under less controlled protocols. Together, these comparisons favor evaluating the implemented strategy and its measured cost for each checkpoint; they do not identify a causal effect of model size or an attention bottleneck.

Reproducibility Statement

The supplement provides per-question responses, predictions and token counts for the matched experiment, along with the generation and analysis code. The question split, model identifiers, image-area settings, sampling parameters and answer grammars are specified above. Public MathVista images and model weights are obtained from their providers. The larger-checkpoint records permit reanalysis of saved predictions, but do not support full response regrading.

AI Use Statement

OpenAI Codex was used to aid and polish manuscript writing and to assist with literature retrieval and discovery, including finding related work. The authors are responsible for the final manuscript and citations.

References

Mohammadjavad Ahmadpour, Amirmahdi Meighani, Payam Taebi, Omid Ghahroodi, Amirmohammad Izadi, and Mahdieh Soleymani Baghshah. Limits and gains of test-time scaling in vision-language reasoning. *arXiv preprint arXiv:2512.11109*, 2025.

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.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. GeoQA: A Geometric Question Answering Benchmark Towards Multimodal Numerical Reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021. URL https://aclanthology.org/2021.findings-acl.46/.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Chen_InternVL_Scaling_up_Vision_Foundation_Models_and_Aligning_for_Generic_CVPR_2024_paper.html.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In *Advances in Neural Information Processing Systems*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/9a6a435e75419a836fe47ab6793623e6-Abstract.html.

Bradley Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979. URL https://doi.org/10.1214/aos/1176344552.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring Mathematical Problem Solving With the MATH Dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah Smith, and Ranjay Krishna. Visual Sketchpad: Sketching as a Visual Chain of Thought for Multimodal Language Models. In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/fb82011040977c7712409fbdb5456647-Abstract-Conference.html.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large Language Models are Zero-Shot Reasoners. In *Advances in Neural Information Processing Systems*, 2022. URL https://proceedings.neurips.cc/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning*, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step by Step. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/aca97732e30bcf1303bc22ac3924fd16-Abstract-Conference.html.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is Your Multi-modal Model an All-around Player? In *European Conference on Computer Vision*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/00959.pdf.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. In *Advances in Neural Information Processing Systems*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/11332b6b6cf4485b84afadb1352d3a9a-Abstract-Conference.html.

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*, 2024.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative Refinement with Self-Feedback. In *Advances in Neural Information Processing Systems*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2022. URL https://aclanthology.org/2022.findings-acl.177/.

Amrith Setlur, Nived Rajaraman, Sergey Levine, and Aviral Kumar. Scaling test-time compute without verification or RL is suboptimal. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 54058–54094, 2025. URL https://proceedings.mlr.press/v267/setlur25a.html.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, 2025.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022.

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*, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wu_V_Guided_Visual_Search_as_a_Core_Mechanism_in_Multimodal_CVPR_2024_paper.html.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In *Advances in Neural Information Processing Systems*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/271db9922b8d1f4dd7aaef84ed5ac703-Abstract-Conference.html.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 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*, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yue_MMMU_A_Massive_Multi-discipline_Multimodal_Understanding_and_Reasoning_Benchmark_for_CVPR_2024_paper.html.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, Peng Gao, and Hongsheng Li. MathVerse: Does Your Multimodal LLM Truly See the Diagrams in Visual Math Problems? In *European Conference on Computer Vision*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/01270.pdf.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. DDCoT: Duty-Distinct Chain-of-Thought Prompting for Multimodal Reasoning in Language Models. In *Advances in Neural Information Processing Systems*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/108030643e640ac050e0ed5e6aace48f-Abstract-Conference.html.