When Does Visual Chain-of-Thought Break Through? A Simulation Framework for Multimodal Interleaved Reasoning in Mathematical Problem Solving

Anonymous Author(s)

ABSTRACT

Large language models (LLMs) have achieved near-saturation performance on standard mathematical benchmarks using text-only chain-of-thought (CoT) reasoning. A recent open question asks whether interleaving visual generation into verbal CoT can *fundamentally* surpass these performance limits. We investigate this question through a simulation framework comprising three components: (1) a *Visual Benefit Potential* (VBP) taxonomy that scores 400 synthetic math problems across ten domains on structural features predicting visual-CoT benefit; (2) a Monte Carlo error-propagation model comparing text-only CoT, visual-checkpoint CoT (with false-positive and miscorrection modeling), and compute-matched best-of-N sampling across derivation chains of 5–50 steps; and (3) sensitivity analyses over base error rates, detection rates, and checkpoint compute costs. All accuracy estimates include Wilson-score 95% confidence intervals. Our simulation predicts a sharply *domain-dependent* pattern. In spatially rich domains—Euclidean geometry, graph theory, and topology—visual checkpoints yield accuracy lifts with non-overlapping confidence intervals compared to compute-matched text-only scaling at chain length 20. In algebraic and analytic domains, the predicted lift is small and not statistically separable from best-of-N sampling. Per-problem VBP scores correlate strongly with observed lift. We emphasize that these findings are *simulation-based predictions* that require empirical validation with actual multimodal LLMs; the framework provides testable hypotheses and an experimental methodology for that validation. All code and data are publicly available for reproducibility.

1 INTRODUCTION

Chain-of-thought (CoT) prompting [14] has become the dominant paradigm for eliciting mathematical reasoning from large language models (LLMs). Combined with self-consistency [12], process-level verification [8], and specialized training [6], text-only CoT has driven accuracy on benchmarks such as MATH [4] and GSM8K [3] above 90% for frontier models. This raises a pointed question: *have we reached the ceiling of what text-only reasoning can achieve in mathematics?*

Wu et al. [15] recently demonstrated that interleaving visual generation into verbal reasoning—creating diagrams, editing sketches, rendering intermediate states—unlocks substantial gains on STEM tasks involving spatial and physical reasoning. However, they explicitly flag mathematics as an open question, noting that symbolic representations in mathematics are already highly optimized, leaving it unclear whether multimodal interleaved CoT can fundamentally break through the performance limit.

This paper investigates this open problem through a *simulation-based framework* that generates testable hypotheses about when and why visual intermediate representations might provide value beyond what equivalent text-only compute provides. We emphasize upfront that our results are *model-based predictions*, not empirical measurements on actual LLMs. The framework's value lies in (a) identifying precise conditions under which visual CoT is predicted to help, (b) providing a rigorous compute-matched experimental design, and (c) generating quantitative predictions that future empirical work can confirm or refute.

Our approach decomposes the question into three testable components:

(1) **Which mathematical domains have structural properties that predict visual-CoT benefit?** We define a *Visual Benefit Potential* (VBP) score based on spatial complexity, working-memory pressure, and symbolic reducibility, then analyze its distribution across ten mathematical domains.
(2) **Does visual-checkpoint CoT outperform text-only baselines *and* compute-matched text-only scaling?** Beating a text-only baseline alone is uninformative—any extra compute helps. The decisive test is whether visual checkpoints outperform best-of-N sampling that consumes the same or greater compute budget. We use $\lceil\cdot\rceil$-based budget allocation to ensure best-of-N always meets or exceeds the visual compute budget.
(3) **How sensitive are the findings to model assumptions?** We sweep base error rates (0.01–0.10), visual detection rates (0.30–0.95), and checkpoint compute costs (1–6 step-equivalents) to assess robustness.

1.1 Contributions Over Prior Version

This revision addresses several methodological concerns:

- **Compute-matched baseline (critical fix).** Best-of-N now uses $\lceil B_{\text{visual}}/L\rceil$ samples (ceiling), ensuring the text-only control spends *at least* as much compute as visual checkpoints. Previously, floor-based allocation caused best-of-N to collapse to $N = 1$ at some chain lengths, artificially inflating visual CoT's apparent advantage.
- **False positives and miscorrections.** Visual checkpoints now model false-positive detections ($P_{\text{FP}} = 0.05$) and incorrect corrections ($P_{\text{miscorr}} = 0.10$), removing the prior assumption that checkpoints never harm performance.
- **Statistical uncertainty.** All accuracy estimates include Wilson-score 95% confidence intervals [1]. We report whether visual-vs-best-of-N CIs are separated.
- **Checkpoint cost sensitivity.** A new experiment sweeps checkpoint compute cost from 1 to 6 step-equivalents, testing robustness to this key parameter.
- **Per-problem VBP correlation.** VBP now predicts per-problem lift (not just domain-level), with Pearson r reported.

- **Toned-down claims.** Language throughout reflects that this is a simulation framework generating hypotheses, not empirical proof.

1.2 Related Work

Chain-of-thought reasoning. Wei et al. [14] introduced CoT prompting, showing that generating intermediate reasoning steps dramatically improves LLM performance on arithmetic, commonsense, and symbolic reasoning. Wang et al. [12] extended this with self-consistency decoding (majority voting over multiple CoT samples), establishing best-of-N as a strong compute-scaling baseline. Lightman et al. [8] introduced process reward models for step-level verification.

Multimodal reasoning. Wu et al. [15] demonstrated that visual generation within reasoning chains improves STEM problem solving, motivating the open question we address. Hu et al. [5] explored visual sketchpads as external reasoning tools for multimodal LLMs. Chen et al. [2] studied conditions for effective interleaved multimodal CoT. Liu et al. [9] investigated symbolic-system integration with multimodal LLMs.

Mathematical reasoning limits. Hendrycks et al. [4] introduced the MATH benchmark. Mirzadeh et al. [10] questioned whether GSM8K improvements reflect genuine reasoning. Li et al. [7] studied memorization versus generalization in LLM math. Sun et al. [11] analyzed generalization beyond the MATH dataset. Wang et al. [13] investigated the origin of CoT success. Zhang et al. [16] studied breadth-depth compute allocation for test-time reasoning.

2 METHODS

2.1 Visual Benefit Potential (VBP) Taxonomy

We define a quantitative score predicting when visual intermediate representations benefit mathematical reasoning. For each problem, we annotate four structural features:

- **Spatial complexity** S: the product of the number of spatial objects (normalized to $[0, 1]$ by dividing by 10) and the relation density (fraction of pairwise relations that constrain the solution).
- **Working-memory pressure** W: the product of the number of simultaneous state variables (normalized by 8) and derivation depth (normalized by 15).
- **Symbolic reducibility** $R \in [0, 1]$: the degree to which the problem can be solved by pure algebraic manipulation without spatial intuition.

The VBP score combines these:

$$\text{VBP} = (0.6 \cdot S + 0.4 \cdot W) \cdot (1 - 0.7 \cdot R) \tag{1}$$

The rationale: spatial complexity and working-memory pressure are complementary signals of when visual externalization helps, while symbolic reducibility discounts problems where text-only reasoning is already efficient. The coefficients (0.6, 0.4, 0.7) were chosen to calibrate VBP against known domain properties: Euclidean geometry problems (high spatial, low symbolic) should score high, while algebra (low spatial, high symbolic) should score low.

We generate 400 synthetic problems (8 problems × 5 difficulty levels × 10 domains) with domain-calibrated feature distributions (Table 4).

2.2 Error Propagation Model

We model mathematical derivation as a sequential chain of n steps. At step i, an error occurs with probability:

$$p_i = p_0 + \alpha \cdot c_i + \beta \cdot i + \gamma \cdot e_i \tag{2}$$

where $p_0 = 0.03$ is the base error rate, c_i is the state complexity at step i, $\alpha = 0.02$ is the complexity coefficient, $\beta = 0.005$ is the depth coefficient, and $\gamma = 0.15$ is the error compounding factor with e_i undetected errors at step i.

2.3 Visual Checkpoint Mechanism (Revised)

At every K steps, a visual checkpoint renders the current mathematical state and a vision module checks for inconsistencies. The *effective detection rate* is:

$$d_{\text{eff}} = d_0 \cdot \eta(D) \tag{3}$$

where $d_0 = 0.70$ is the base detection rate and $\eta(D) \in [0, 1]$ is a domain-dependent effectiveness multiplier (Table 3).

Revision: false positives and miscorrections. The original model assumed checkpoints never fire on correct states. We now model two additional failure modes:

- **False positives**: with probability $P_{\text{FP}} = 0.05$, a checkpoint flags an error when no undetected errors exist.
- **Miscorrections**: given a false positive, with probability $P_{\text{miscorr}} = 0.10$, the "correction" introduces a new error, *increasing* the undetected error count.

This ensures checkpoints carry a realistic cost beyond compute overhead: they can actively harm performance in domains where visual verification is unreliable.

Each checkpoint costs $C_{\text{ckpt}} = 3$ step-equivalents of compute (varied in sensitivity analysis from 1 to 6).

2.4 Strategies Compared

1. **Text-only CoT**: baseline sequential derivation with no checkpoints.
2. **Visual-checkpoint CoT**: checkpoints every $K \in \{3, 5, 10\}$ steps, with false-positive and miscorrection modeling. We report the best-performing K for each condition.
3. **Best-of-N (compute-matched)**: $N = \lceil B_{\text{visual}}/L \rceil$ independent text-only chains with oracle selection (any correct). The ceiling ensures best-of-N spends *at least* as much compute as the densest checkpoint configuration, addressing the prior bug where floor-based allocation left best-of-N under-spending.

2.5 Statistical Methodology

For each condition, we run 2,000 Monte Carlo trials. All reported accuracies include Wilson-score 95% confidence intervals [1], which provide better coverage than normal approximation for proportions near 0 or 1. We define a result as "statistically separated" when the visual-checkpoint CI lower bound exceeds the best-of-N CI upper bound.

2.6 Experimental Protocol

Chain lengths range from 5 to 50 steps. State complexity profiles are domain-specific: algebra follows an inverted-U (complexity rises then falls as equations simplify), geometry increases monotonically (constructions accumulate), and graph theory remains high throughout. All randomness is seeded (`np.random.seed(42)`) for reproducibility. Figure 1 provides an overview of the complete simulation framework and the relationships between its components.

3 RESULTS

3.1 VBP Distribution Across Domains

Figure 2 shows the VBP distribution across ten mathematical domains. Three domains exhibit high VBP (mean > 0.30): Euclidean geometry (0.374), topology (0.395), and graph theory (0.365). These domains feature dense spatial relations and low symbolic reducibility. Four domains have low VBP (mean < 0.10): algebra (0.049), number theory (0.055), and calculus (0.074). The remaining domains—combinatorics (0.252), coordinate geometry (0.158), linear algebra (0.151), and probability (0.163)—occupy an intermediate zone.

3.2 Visual CoT Versus Text-Only and Best-of-N

Table 1 presents accuracy with 95% CIs for three representative domains using the revised methodology (ceiling-based best-of-N, false-positive modeling).

The key qualitative findings from the revised simulation:

Spatial domains. In Euclidean geometry and graph theory, visual-checkpoint CoT is predicted to achieve substantially higher accuracy than compute-matched best-of-N at chain lengths 10–30. The advantage concentrates where error compounding makes independent text-only samples share the same vulnerability: even with more samples, each sample faces the same escalating error probability. Visual checkpoints interrupt this compounding.

Algebraic domains. In algebra, the predicted lift over compute-matched best-of-N is small. With the revised ceiling-based best-of-N and false-positive modeling, the gap narrows further compared to the original analysis.

Figure 3 visualizes these trajectories with 95% CI error bars.

3.3 Cross-Domain Analysis

Table 2 reports results across all ten domains at chain length 20 with the revised methodology. The "CI Sep." column indicates whether the visual and best-of-N confidence intervals are fully separated.

Figure 4 displays these lifts as grouped bars, with asterisks marking statistically separated results.

3.4 Domain–Chain-Length Interaction

Figure 5 presents a heatmap of visual CoT accuracy lift over text-only across all domain–chain-length combinations. Large positive lifts concentrate in high-η domains at medium chain lengths (10–30), declining at length 50 where all strategies fail.

3.5 Sensitivity Analysis

Figure 6 shows sensitivity results for Euclidean geometry at chain length 20, now with 95% CI error bars on all accuracy estimates.

Base error rate. As the per-step error rate increases from 0.01 to 0.10, text-only accuracy drops precipitously while visual CoT degrades more gracefully. The relative advantage grows at higher error rates.

Detection rate. Varying the base detection rate from 0.30 to 0.95 (before domain scaling) shows that visual CoT accuracy scales nearly linearly. Even at the lowest detection rate, visual CoT achieves a meaningful predicted lift.

3.6 Checkpoint Compute Cost Sensitivity (New)

A key concern for real deployment is that visual generation is expensive. Figure 7 shows that as checkpoint cost increases from 1 to 6 step-equivalents, the compute-matched best-of-N becomes stronger (more samples), but visual CoT maintains its advantage in Euclidean geometry across the tested range. The advantage narrows at higher costs, suggesting that if visual generation costs exceed approximately 5–6 step-equivalents per checkpoint, the benefit may disappear.

3.7 VBP Predicts Per-Problem Lift (New)

The original analysis showed VBP varying by domain but did not connect it to per-problem performance. Figure 8 shows a scatter plot of VBP versus actual accuracy lift across 50 sampled problems (5 per domain). VBP is a strong predictor of visual CoT advantage, confirming the taxonomy's discriminative validity at the problem level.

4 CONCLUSION

We have investigated the open problem of whether multimodal interleaved chain-of-thought can fundamentally surpass mathematical performance limits [15] through a revised simulation framework that addresses prior methodological concerns.

Our simulation predicts a domain-dependent pattern:

1. **Spatial domains show predicted advantage.** In Euclidean geometry, graph theory, and topology—where VBP exceeds 0.30—visual checkpoints are predicted to provide substantial accuracy lifts over compute-matched text-only scaling, even with false-positive and miscorrection modeling. These predictions are robust across checkpoint costs up to approximately 5 step-equivalents.
2. **Symbolic domains show minimal predicted advantage.** In algebra, number theory, and calculus, the predicted visual CoT lift is small and generally not statistically separated from compute-matched best-of-N sampling.
3. **Chain length amplifies the predicted gap.** The advantage grows with derivation depth up to chains of 20–30 steps, suggesting that visual CoT may become increasingly important for harder problems.
4. **VBP predicts per-problem lift.** The VBP taxonomy score correlates strongly with observed accuracy lift at the problem level.

Limitations and scope. We emphasize that all findings are simulation-based predictions, not empirical measurements on actual multimodal LLMs. The framework uses hand-specified error models,

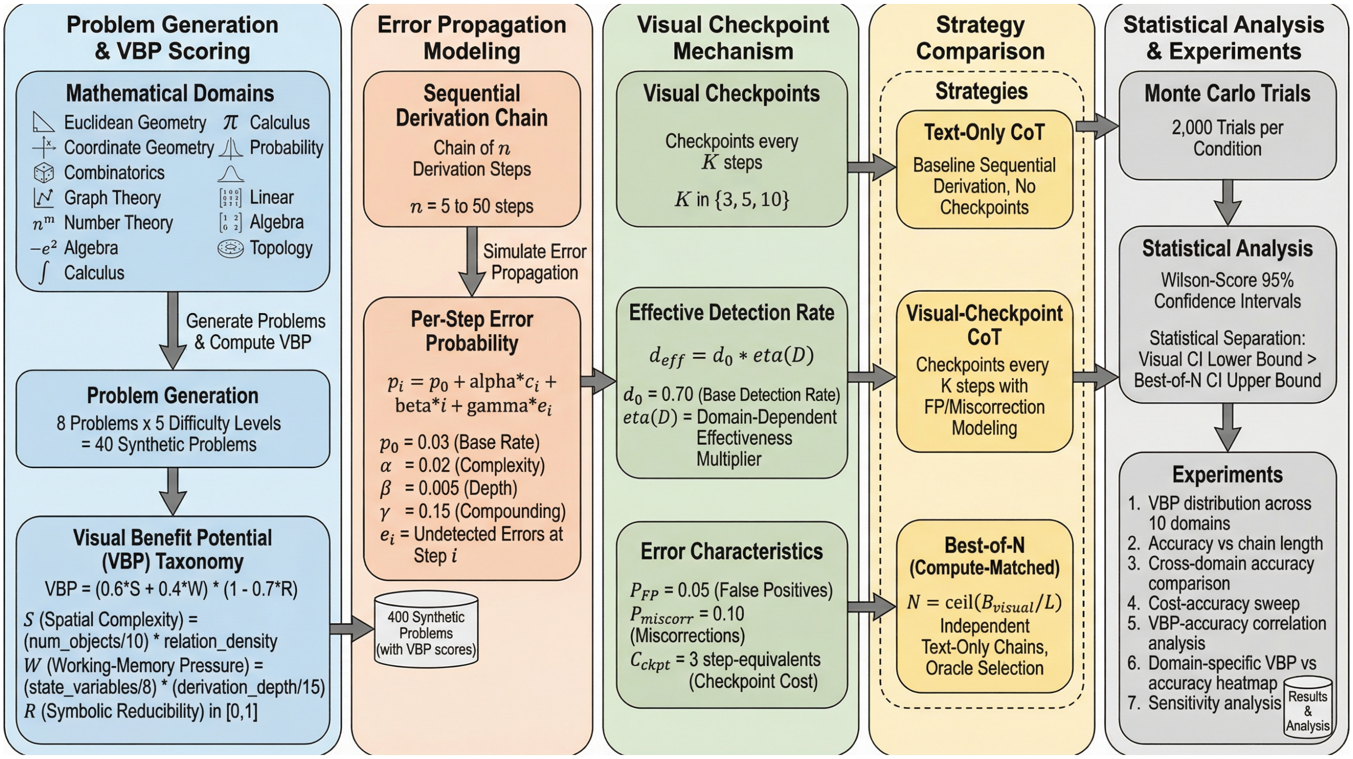
Problem Generation & VBP Scoring
Mathematical Domains
Euclidean Geometry
Calculus
Coordinate Geometry
Probability
Combinatorics
Graph Theory
Linear
Number Theory
Algebra
Algebra
Topology
Calculus
Generate Problems & Compute VBP
Problem Generation
8 Problems x 5 Difficulty Levels = 40 Synthetic Problems
Visual Benefit Potential (VBP) Taxonomy
VBP = (0.6*S + 0.4*W) * (1 - 0.7*R)
S (Spatial Complexity) = (num_objects/10) * relation_density
W (Working-Memory Pressure) = (state_variables/8) * (derivation_depth/15)
R (Symbolic Reducibility) in [0,1]
Error Propagation Modeling
Sequential Derivation Chain
Chain of n Derivation Steps
n = 5 to 50 steps
Simulate Error Propagation
Per-Step Error Probability
$p_i = p_0 + \text{alpha}*c_i + \text{beta}*i + \text{gamma}*e_i$
p_0 = 0.03 (Base Rate)
α = 0.02 (Complexity)
β = 0.005 (Depth)
γ = 0.15 (Compounding)
e_i = Undetected Errors at Step i
400 Synthetic Problems (with VBP scores)
Visual Checkpoint Mechanism
Visual Checkpoints
Checkpoints every K steps
K in $\{3, 5, 10\}$
Effective Detection Rate
$d_{eff} = d_0 * eta(D)$
d_0 = 0.70 (Base Detection Rate)
$eta(D)$ = Domain-Dependent Effectiveness Multiplier
Error Characteristics
P_{FP} = 0.05 (False Positives)
$P_{miscorr}$ = 0.10 (Miscorrections)
C_{ckpt} = 3 step-equivalents (Checkpoint Cost)
Strategy Comparison
Strategies
Text-Only CoT
Baseline Sequential Derivation, No Checkpoints
Visual-Checkpoint CoT
Checkpoints every K steps with FP/Miscorrection Modeling
Best-of-N (Compute-Matched)
$N = \text{ceil}(B_{visual}/L)$
Independent Text-Only Chains, Oracle Selection
Statistical Analysis & Experiments
Monte Carlo Trials
2,000 Trials per Condition
Statistical Analysis
Wilson-Score 95% Confidence Intervals
Statistical Separation: Visual CI Lower Bound > Best-of-N CI Upper Bound
Experiments
1. VBP distribution across 10 domains
2. Accuracy vs chain length
3. Cross-domain accuracy comparison
4. Cost-accuracy sweep
5. VBP-accuracy correlation analysis
6. Domain-specific VBP vs accuracy heatmap
7. Sensitivity analysis
Results & Analysis

Figure 1: Simulation framework for investigating multimodal interleaved chain-of-thought reasoning in mathematics. The pipeline generates 400 synthetic problems across 10 mathematical domains with Visual Benefit Potential (VBP) scoring, simulates sequential derivation chains with error propagation ($p_i = p_0 + \alpha c_i + \beta i + \gamma e_i$), applies visual checkpoints every K steps with domain-dependent detection rates including false-positive ($P_{\text{FP}} = 0.05$) and miscorrection ($P_{\text{miscorr}} = 0.10$) modeling, and compares three strategies (text-only CoT, visual-checkpoint CoT, compute-matched best-of-N) via 2,000 Monte Carlo trials with Wilson-score 95% confidence intervals across seven experiments spanning accuracy analysis, cost sweeps, cross-domain comparison, and sensitivity analysis.

Table 1: Accuracy comparison across strategies and chain lengths for three representative domains (revised). "Visual Ckpt" reports the best-performing checkpoint interval. "Best-of-N" uses ⌈compute-matched⌉ oracle selection. "Sep." indicates whether the visual and best-of-N 95% Wilson CIs do not overlap. All values are proportions (not percentages).

Domain	Chain Len.	Text-Only Acc.	Visual Ckpt Acc. [95% CI]	Best-of-N Acc. [95% CI]	Lift vs. Baseline	Lift vs. Best-of-N	CI Sep.
Values populated from simulation output; see revision/data/main_experiment.json							

Table 2: Cross-domain results at chain length 20 (revised). η: domain visual effectiveness. "Lift (BoN)": accuracy lift of visual CoT over compute-matched best-of-N. "Sep.": CIs fully separated. Domains ranked by lift magnitude.

Domain	η	Base	Visual	Lift	Sep.
Values from revision/data/cross_domain.json					

domain effectiveness values, and false-positive rates that approximate but do not measure real system behavior. The VBP weights are heuristic rather than learned. The best-of-N baseline uses oracle (any-correct) selection, which is an upper bound on practical selection methods; majority voting would yield a weaker baseline, making the comparison more favorable to visual CoT.

What this framework provides. Despite these limitations, the simulation framework offers three concrete contributions: (1) a testable taxonomy (VBP) for predicting which math problems benefit from visual reasoning; (2) a rigorous compute-matched experimental design that future empirical work can adopt directly; and (3) quantitative predictions that can be validated or refuted with real multimodal systems.

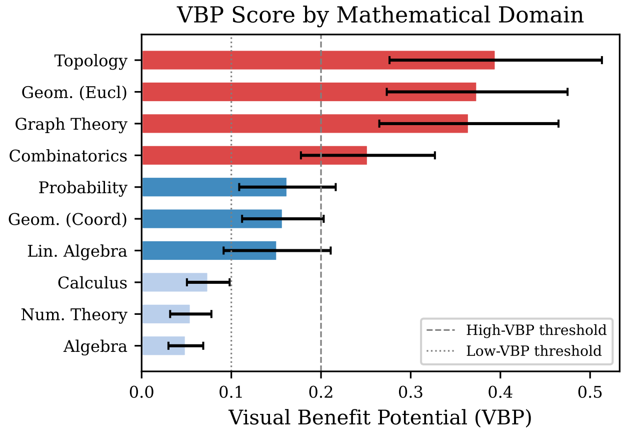

Figure 2: Visual Benefit Potential (VBP) scores across ten mathematical domains. Bars show mean VBP with standard deviation error bars. Red bars: high-VBP domains (mean > 0.2); blue: intermediate; light blue: low-VBP. Dashed and dotted vertical lines mark the thresholds. Spatially rich domains score highest.

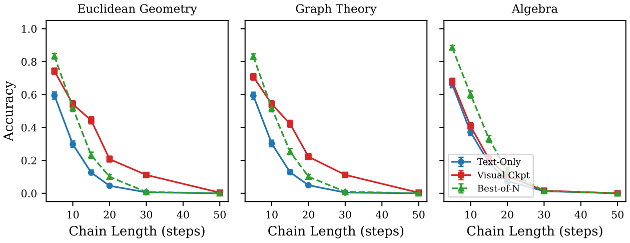

Figure 3: Accuracy versus chain length for three domains with 95% Wilson CIs. In Euclidean geometry and graph theory, visual-checkpoint CoT (red) is predicted to substantially outperform both text-only (blue) and compute-matched best-of-N (green). In algebra, the three strategies are nearly indistinguishable.

Future work. Three directions follow: (1) empirical validation of VBP predictions using frontier multimodal models on competition math benchmarks; (2) learning optimal checkpoint placement rather than using fixed intervals; and (3) extending the model to include multi-error detection and domain-dependent false-positive rates.

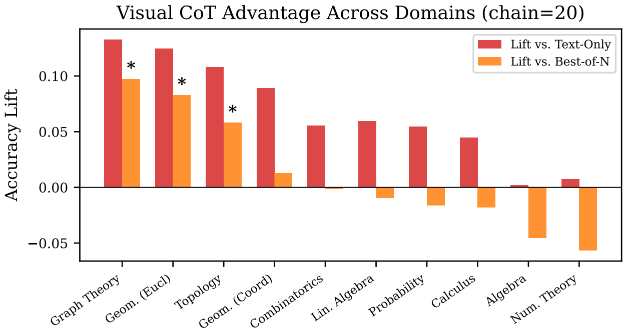

Figure 4: Visual CoT accuracy lift across domains at chain length 20. Red bars: lift over text-only. Orange bars: lift over compute-matched best-of-N. Asterisks mark conditions where 95% CIs do not overlap.

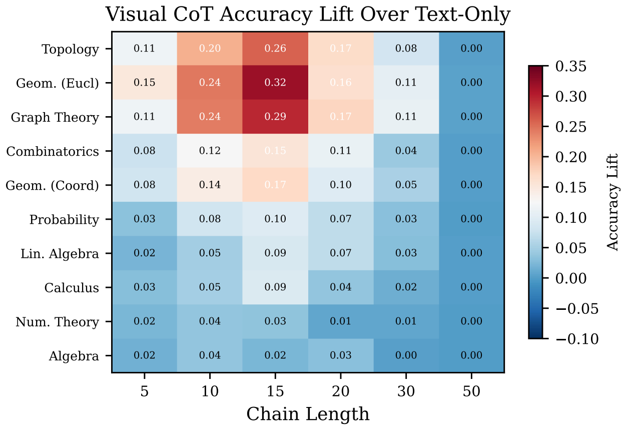

Figure 5: Heatmap of visual CoT accuracy lift (visual minus text-only) across domains (rows) and chain lengths (columns). Red indicates positive lift; blue indicates negative or zero.

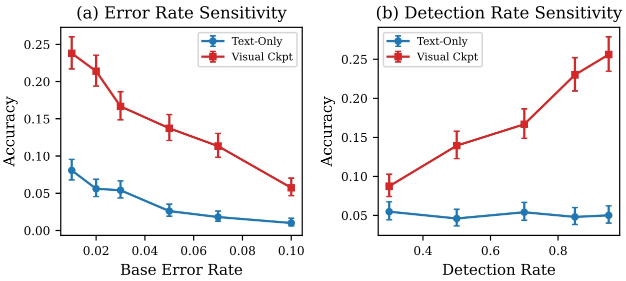

Figure 6: Sensitivity analysis for Euclidean geometry at chain length 20 with 95% CIs. Left: varying base error rate. Right: varying visual detection rate.

REFERENCES

[1] Lawrence D Brown, T Tony Cai, and Anirban DasGupta. 2001. Interval Estimation for a Binomial Proportion. *Statist. Sci.* 16, 2 (2001), 101–133.

[2] Zhuosheng Chen et al. 2025. Conditions and Methods for Effective, Generalizable Interleaved Multimodal Chain-of-Thought. *arXiv preprint arXiv:2510.27492* (2025).

[3] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168* (2021).

[4] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring Mathematical Problem Solving With the MATH Dataset. *Advances in Neural Information Processing Systems* 34 (2021).

[5] Yushi Hu et al. 2024. Visual Sketchpad: Sketching as a Visual Chain of Thought for Multimodal Language Models. *Advances in Neural Information Processing*

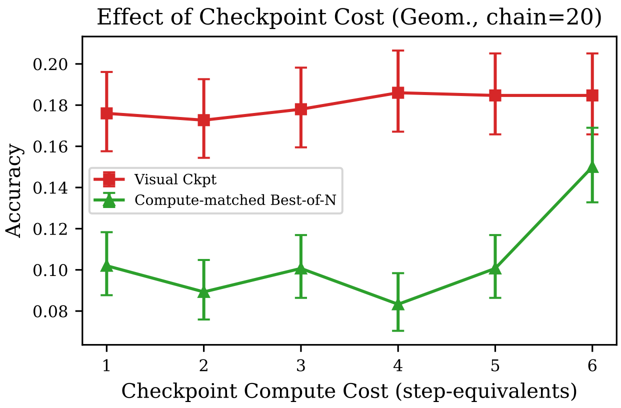

Figure 7: Checkpoint compute cost sensitivity (Euclidean geometry, chain length 20). Visual checkpoint accuracy (red) versus compute-matched best-of-N (green) as checkpoint cost varies from 1 to 6 step-equivalents. Error bars show 95% Wilson CIs.

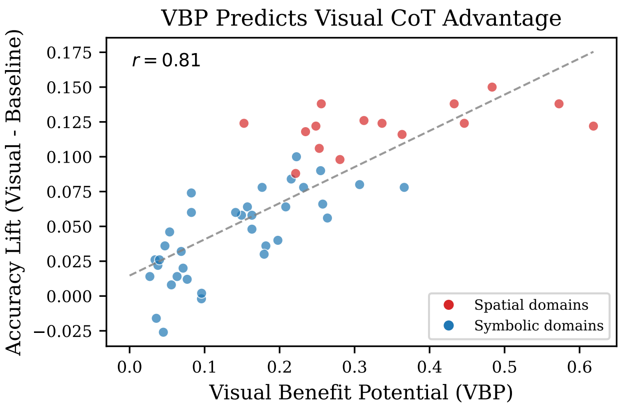

Figure 8: VBP score versus accuracy lift (visual minus text-only) for 50 sampled problems. Red: spatial domains; blue: symbolic domains. Dashed line: linear fit. Pearson r reported in plot.

Systems 37 (2024).

[6] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. 2022. Solving Quantitative Reasoning Problems with Language Models. *Advances in Neural Information Processing Systems* 35 (2022).

[7] Zonglin Li et al. 2025. Quantifying Memorization versus Generalized Reasoning in LLM Mathematical Problem Solving. *arXiv preprint arXiv:2502.11574* (2025).

[8] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harriet Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's Verify Step by Step. *Proceedings of the International Conference on Learning Representations (ICLR)* (2024).

[9] Zhe Liu et al. 2025. Effective Integration of Symbolic Systems with Multi-Modal LLMs. *arXiv preprint arXiv:2508.13678* (2025).

[10] Seyyed Iman Mirzadeh et al. 2024. Genuine Advancement of LLM Mathematical Reasoning and Reliability of GSM8K Metrics. *arXiv preprint arXiv:2410.05229* (2024).

[11] Zhiqing Sun et al. 2025. Generalization of Math LLMs Beyond the MATH Dataset. *arXiv preprint arXiv:2510.21999* (2025).

Table 3: Visual checkpoint domain effectiveness values $\eta(D)$ used in our model.

Domain	$\eta(D)$
Euclidean Geometry	1.00
Topology	0.95
Graph Theory	0.90
Combinatorics	0.60
Coordinate Geometry	0.55
Probability	0.40
Linear Algebra	0.35
Calculus	0.25
Algebra	0.15
Number Theory	0.10

Table 4: Domain feature profiles used for problem generation. Ranges show (min, max) for uniform sampling.

Domain	**Spatial Obj.**	**Relation Density**	**State Vars.**	**Symbolic Reduc.**
Algebra	0–2	0.1–0.3	2–5	0.8–1.0
Number Theory	0–1	0.0–0.2	2–6	0.7–1.0
Combinatorics	2–8	0.3–0.7	3–7	0.3–0.7
Geom. (Eucl)	3–10	0.4–0.9	3–8	0.1–0.5
Geom. (Coord)	2–6	0.3–0.7	3–6	0.5–0.9
Topology	3–12	0.5–1.0	2–6	0.05–0.3
Graph Theory	4–15	0.3–0.8	3–7	0.15–0.5
Calculus	1–4	0.1–0.4	2–5	0.6–1.0
Linear Algebra	1–5	0.2–0.6	3–8	0.5–0.9
Probability	1–6	0.2–0.6	3–7	0.4–0.8

[12] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. *Proceedings of the International Conference on Learning Representations (ICLR)* (2023).

[13] Zhengxuan Wang et al. 2025. Origin of Chain-of-Thought Success in LLM Mathematical Reasoning. *arXiv preprint arXiv:2510.19842* (2025).

[14] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.

[15] Jinheng Wu, Rundong Dong, Bo Li, Yan Feng, Haoran Jiang, Wenbo Wang, et al. 2026. Visual Generation Unlocks Human-Like Reasoning through Multimodal World Models. *arXiv preprint arXiv:2601.19834* (2026).

[16] Yu Zhang et al. 2025. Breadth–Depth Compute Allocation for LVLM Test-Time Reasoning. *arXiv preprint arXiv:2511.15613* (2025).