Investigating Inference-time Scaling for Chain of Multi-modal Thought: A Preliminary Study

Yujie Lin[1,3*], **Ante Wang**[1,3*], **Moye Chen**[2], **Jingyao Liu**[1], **Hao Liu**[2]
Jinsong Su[1,3,4†] and **Xinyan Xiao**[2]
[1]School of Informatics, Xiamen University, China [2]Baidu Inc., Beijing, China
[3]Key Laboratory of Digital Protection and Intelligent Processing of Intangible Cultural Heritage of Fujian and Taiwan (Xiamen University), Ministry of Culture and Tourism, China
[4]Shanghai Artificial Intelligence Laboratory
{yjlin,wangante}@stu.xmu.edu.cn jssu@xmu.edu.cn

Abstract

Recently, inference-time scaling of chain-of-thought (CoT) has been demonstrated as a promising approach for addressing multi-modal reasoning tasks. While existing studies have predominantly centered on text-based thinking, the integration of both visual and textual modalities within the reasoning process remains unexplored. In this study, we pioneer the exploration of inference-time scaling with multi-modal thought, aiming to bridge this gap. To provide a comprehensive analysis, we systematically investigate popular sampling-based and tree search-based inference-time scaling methods on 10 challenging tasks spanning various domains. Besides, we uniformly adopt a consistency-enhanced verifier to ensure effective guidance for both methods across different thought paradigms. Results show that multi-modal thought promotes better performance against conventional text-only thought, and blending the two types of thought fosters more diverse thinking. Despite these advantages, multi-modal thoughts necessitate higher token consumption for processing richer visual inputs, which raises concerns in practical applications. We hope that our findings on the merits and drawbacks of this research line will inspire future works in the field.[1]

1 Introduction

The remarkable performance of OpenAI's o1 in tackling reasoning tasks has sparked widespread research interest in studying inference-time scaling across various communities (Chen et al., 2025). This technique is built upon the success of Chain-of-Thought (CoT, Wei et al. 2022), which enables large language models (LLMs) to solve a problem through step-by-step reasoning. It makes use of additional computation at inference time to conduct "*slow thinking*" (Kahneman, 2011) to further improve the accuracy of responses. Taking the rich experience of pioneering research in language-only approaches (Xu, 2023; Yao et al., 2024b; Snell et al., 2024; Wu et al., 2024; Luo et al., 2024; Xie et al., 2024; Wang et al., 2024b,a; Li et al., 2025; Wang et al., 2025; Bi et al., 2025), rapid progress has been made in addressing multi-modal reasoning tasks leveraging typical methods such as Best-of-N, Beam Search, and Monte Carlo Tree Search (MCTS) (Xu et al., 2024; Yao et al., 2024a).

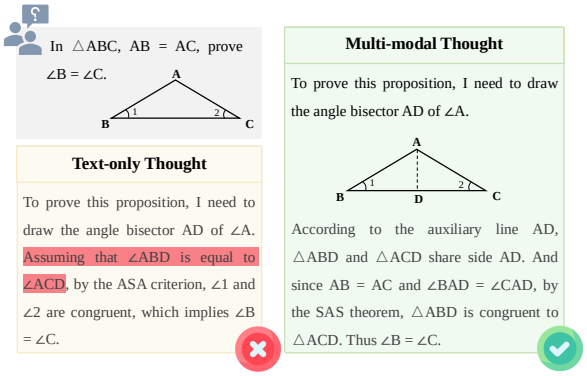

Figure 1: Solving the problem using text-only thought vs. multi-modal thought. Additional visual information from the latter offers richer and more intuitive features, making it easier to yield better subsequent steps.

Although effective, these methods mostly follow the practice of text-based thinking, overlooking the crucial role of thinking in the visual modality in multi-modal scenarios. This oversight leads to sub-optimal performance, thus highlighting the need for new approaches that integrate both modalities. Consider the process of solving geometric problems as an example: humans often utilize auxiliary lines to facilitate problem-solving. Analogously, rich and intuitive visual information, e.g., drawn auxiliary lines, can enhance the model's ability to generate improved solutions, as shown in Figure 1. Recent studies (Lu et al., 2022, 2023; Zhou et al., 2024; Hu

*These authors contributed equally.
†Corresponding author.
[1]Our code can be found at https://github.com/DeepLearnXMU/TTS_COMT

Findings of the Association for Computational Linguistics: ACL 2025, pages 15654–15667
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

et al., 2024; Zhang et al., 2024b; Chen et al., 2024b; Cheng et al., 2025) have also demonstrated the benefits of incorporating supplementary visual information at each reasoning step compared with conventional text-based methods (Zhang et al., 2023; Zheng et al., 2023; Mitra et al., 2024). Therefore, an important yet under-explored question arises: *What is the potential of inference-time scaling of multi-modal thinking?*

To this end, we conduct the first study on this problem, shedding light on the merits and constraints of this new research line. We follow Wei et al. (2022) and Hu et al. (2024) to elicit text-only and multi-modal thought, respectively. For inference-time scaling, we investigate popular sampling-based and tree search-based methods. Sampling-based method generates multiple independent samples in parallel, applying techniques like Self-Consistency (Wang et al., 2023) or Best-of-N (Chen et al., 2024a; Wan et al., 2024) to select the most reliable reasoning trails. However, this approach is inefficient because it requires exploring full solution paths, even if a mistake has occurred early. Tree search-based method tackles this issue by utilizing sophisticated search algorithms, such as Beam Search (Yao et al., 2024b) and MCTS (Tian et al., 2024a). Both these methods ask for a critic to discriminate the promising samples or search steps. We consistently employ an advanced Large Vision Language Model (LVLM) as a verifier to infer critic scores and, alternatively, enhance this method by aggregating results from multiple trials. In this way, it effectively provides reliable feedback without the need for specific training.

We conduct a comparison of inference-time scaling using multi-modal and text-only thought across 10 datasets, encompassing tasks related to geometric reasoning, mathematical reasoning, and visual question answering. The main findings are as follows:

- The use of multi-modal thought achieves significantly better performance and higher upper bounds compared to text-only thought on average, demonstrating the potential of this line of research.

- Although effective, it requires higher token consumption to process richer visual inputs. This result calls for more efficient reasoning mechanisms, especially for practical concerns.

- Further analyses show that tree search-based methods effectively mitigate reasoning errors, such as invalid processed images. However, their success heavily relies on verifier performance, highlighting the need to develop more effective multi-modal verifiers.

We hope that these findings can inspire future research to develop more advanced and robust methods for multi-modal reasoning.

2 Preliminaries

In this section, we firstly establish the formulation for both text-only and multi-modal thought (§2.1), and then provide a detailed explanation of two inference-time scaling paradigms (§2.2): sampling-based and tree search-based methods. In addition, we present a prompting-based verifier equipped with a consistency-enhanced mechanism to effectively guide these inference-time scaling methods (§2.3).

2.1 Thought Formulation

Chain-of-Thought (CoT, Wei et al. 2022) is a technique that encourages the model to generate intermediate reasoning steps $\mathbf{S} = (\mathbf{s}_1, ..., \mathbf{s}_n)$ before reaching the final answer $\mathbf{a}$. In the context of multi-modal reasoning tasks, given a problem $\mathbf{q}$ and images $\mathbf{I}$, each reasoning step $\mathbf{s}_i$ is sampled from an LVLM $\mathcal{M}$:

$$\mathbf{s}_i \sim \mathcal{M}(\mathbf{q}, \mathbf{I}, \mathbf{s}_{1:i-1}). \quad (1)$$

This expansion allows for more computations to be deployed and can unlock the ability to solve more complex problems (Li et al., 2024b).

Following the success on language-only reasoning tasks, most previous works (Zhang et al., 2023; Zheng et al., 2023; Mitra et al., 2024) study *text-only thought*, where $\mathbf{S}$ is defined as a text string. In this work, we aim to explore the potential of *multi-modal thought*, where any reasoning step $\mathbf{s}_i$ can blend multi-modal information, e.g., both text and image.

However, as most prevalent LVLMs still struggle to generate reliable images, we follow Hu et al. (2024) to allow the LVLM to generate executable code for visual manipulation instead. Implementation details can be found in Appendix A.

2.2 Inference-Time Scaling

2.2.1 Sampling-based Methods

These methods scale inference-time computation by generating multiple reasoning chains in parallel, and then determining the most promising answer from them. Here we investigate two widely used techniques, Self-Consistency (Wang et al., 2023) and Best-of-N (Chen et al., 2024a; Wan et al., 2024).

Self-Consistency This method leverages the intuition that the correct answer typically can be reached using multiple different ways of thinking (Wang et al., 2023). Given N sampled reasoning chains $\mathbf{S}_1, ..., \mathbf{S}_N$ generated from the LVLM, it selects the most consistent answer via majority voting over their corresponding answers $\mathbf{a}_1, ..., \mathbf{a}_N$:

$$\mathbf{a}^* = \arg \max_{\mathbf{a} \in \{\mathbf{a}_1, ..., \mathbf{a}_N\}} \sum_{i=1}^{N} \mathbb{I}(\mathbf{a}_i = \mathbf{a}), \quad (2)$$

where $\mathbb{I}(\cdot)$ denotes the indicator function and $\mathbf{a}^*$ represents the selected answer.

Best-of-N This approach utilizes trained verifiers to evaluate the correctness of model generated solutions. Compared to Self-Consistency based on voting, it can be more efficient when a strong verifier is available. Formally, given N sampled reasoning chains and a verifier $\mathcal{F}(\cdot)$, we have

$$\mathbf{S}^* = \max_{\mathbf{S} \in \{\mathbf{S}_1, ..., \mathbf{S}_N\}} \mathcal{F}(\mathbf{S}), \quad (3)$$

where $\mathbf{S}^*$ is the reasoning chain with the highest verifier score, from which the final answer $\mathbf{a}^*$ is derived.

2.2.2 Tree Search-based Methods

Sampling-based methods are easy to implement due to their simplicity. However, they are inefficient because they require exploring full solution paths, even if a mistake occurs early (Xiang et al., 2025). To address this challenge, tree search-based methods are later explored. They model the problem-solving as a tree search process, where each node represents a reasoning step. Here, we investigate the most prevalent Beam Search (Welleck et al., 2022; Xie et al., 2023; Yao et al., 2024b; Wu et al., 2024) and MCTS (Xu, 2023; Ding et al., 2024; Luo et al., 2024; Xie et al., 2024; Liu et al., 2024a; Tian et al., 2024a; Yao et al., 2024a) algorithms in both language-only and multi-modal tasks.

Beam Search Adapted from classical beam search algorithms used in sequence generation tasks (e.g., machine translation (Sutskever, 2014; Wu, 2016)), this method dynamically explores and prunes reasoning paths at each step for deductive reasoning. The algorithm operates with a beam width B (number of retained reasoning chains per step) and an expansion size N (new thoughts generated per candidate). Starting with N initial candidate thoughts sampled from a question q via Equation 1, a verifier $\mathcal{F}(\cdot)$ scores all candidates, retaining only the top-B chains. Each retained chain is then expanded by sampling N new thoughts, producing $B \times N$ candidates, after which the verifier reevaluates and prunes the pool to the top-B chains. This cycle of scoring and pruning repeats iteratively until all active chains reach finished states (final answers), at which point the highest-scoring finished chain across all iterations is selected as the final answer $\mathbf{a}^*$.

MCTS It is a decision-making algorithm that is powerful for solving complex problems, as demonstrated by AlphaGo (Silver et al., 2016, 2017). MCTS operates by iteratively exploring the most promising reasoning paths through four key phases: selection, expansion, evaluation, and backpropagation.

In the selection phase, the algorithm traverses the tree from the root node, recursively choosing child nodes using the Upper Confidence Bound (UCB) metric (Kocsis and Szepesvári, 2006). This metric strategically balances the exploitation of high-value paths and the exploration of under-visited paths. The UCB formula is defined as:

$$\text{UCB}(\mathbf{s}_i) = \frac{V(\mathbf{s}_i)}{C(\mathbf{s}_i)} + w \times \sqrt{\frac{\ln C(\mathbf{s}_{i-1})}{C(\mathbf{s}_i)}}, \quad (4)$$

where $V(\mathbf{s}_i)$ and $C(\mathbf{s}_i)$ denotes the accumulated value and count of being visited of $\mathbf{s}_i$. w is a weighting parameter that adjusts the exploration-exploitation balance. During expansion, the algorithm extends the selected node by generating a new thought through Equation 1. The evaluation phase then estimates the value of this newly expanded node using a verifier $\mathcal{F}(\cdot)$. Finally, backpropagation updates the values of all ancestors, propagating the evaluation results back to the root.

After multiple iterations, the finished reasoning chain with the highest accumulated value is selected, yielding the optimal final answer $\mathbf{a}^*$.

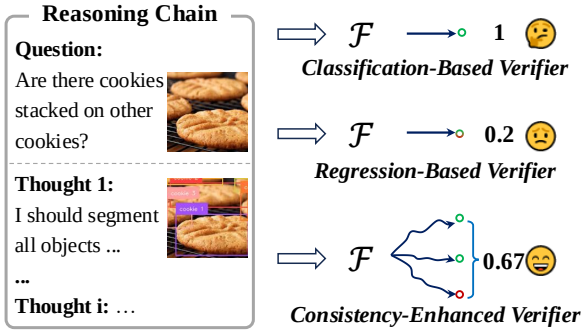

Figure 2: Three types of verifiers investigated in this work, where the classification-based verifier outputs sparse binary scores (0 or 1) and the regression-based verifier provides dense but inaccurate scores. We introduce the consistency-enhanced verifier to compute dense and accurate scores by aggregating multiple evaluations sampled from the classification-based verifier.

2.3 Verifiers for Inference-Time Scaling

The verifier $\mathcal{F}$ is pivotal in inference-time scaling, tasked with evaluating the validity of reasoning chains. Existing approaches fall into two categories: training-based (Xiong et al., 2024; Wang et al., 2024c) and prompting-based (Tian et al., 2024b; Shinn et al., 2024; Jones et al., 2024). Training-based approaches hold theoretical promise by optimizing verifiers on corresponding labeled data. However, acquiring such annotations is challenging, resulting in datasets limited in scale and diversity, thereby restricting the generalizability of trained verifiers. Thus, this work focuses on prompting-based approaches, capitalizing on the strong instruction-following capabilities of LVLMs to assess reasoning chains.

Formally, given a reasoning chain $\mathbf{s}_{1:i}$, the verifier $\mathcal{F}$ operates as: $\mathbf{r} = \mathcal{M}(\mathbf{P}, \mathbf{s}_{1:i})$, where $\mathbf{P}$ represents the verification instruction, and $\mathbf{r}$ is the evaluation output, which includes a scalar score. Particularly, we investigate two instruction designs for $\mathbf{P}$:

- **Classification-Based** (Tian et al., 2024b; Zhang et al., 2024a; Shinn et al., 2024): It prompts the verifier to classify the reasoning chain as "correct" or "incorrect." While intuitive, this binary output provides sparse feedback, offering limited granularity to guide searching.

- **Regression-Based** (Jones et al., 2024; Xiong et al., 2024): Here, the verifier is prompted to output a continuous score $\in [0, 1]$ reflecting the quality of chains. However, directly predicting precise scores is error-prone, often yielding high-variance results that degrade performance.[2]

To address these limitations, we introduce the **Consistency-Enhanced Verifier**, compared with the above two in Figure 2. This method aggregates N_v independent evaluations using classification-based instructions, computing the proportion of "correct" classifications across trials. In this way, it not only avoids variance in regression-based outputs but also overcomes the sparsity of single-trail classification, enabling more reliable guidance for inference-time scaling.

3 Experiments

3.1 Setup

Datasets We conduct experiments on 10 datasets spanning three distinct domains:

- **Geometric Reasoning:** To assess geometry problem-solving capabilities, we utilize the Geometry3K dataset (Lu et al., 2021) and three tasks from IsoBench (Fu et al., 2024), including Connectivity, Maxflow, and Isomorphism.

- **Mathematical Reasoning:** We evaluate mathematical reasoning performance using the Parity, Convexity, and Winner ID tasks from IsoBench (Fu et al., 2024).

- **Visual Question Answering (VQA):** To test generalizability, we also employ three challenging VQA benchmarks: V^* Bench (Wu and Xie, 2024), MMVP (Tong et al., 2024), and BLINK (Fu et al., 2025).

More detailed introduction to these datasets can be found in Appendix C.

Baselines CoT (Wei et al., 2022) and Visual Sketchpad (Hu et al., 2024) are employed in our experiments to implement text-only thought and multi-modal thought, respectively. Both of them are further compared under four inference-time scaling methods: Self-Consistency, Best-of-N, Beam Search, and MCTS. Additionally, the following latest multi-modal reasoning frameworks are also compared in our main experiment, including Multimodal-CoT (Zhang et al., 2023), DD-CoT (Zheng et al., 2023), and CCoT (Mitra et al., 2024). Detailed baseline descriptions are provided in Appendix D.

[2]The prompts used for these two designs are in Appendix B.

Methods	Geometry				Math			VQA		
	Geometry3K	Maxflow	Isomorphism	Connectivity	Convexity	Parity	Winner ID	V^* Bench	MMVP	BLINK
Direct	35.4	28.9	50.8	52.3	90.2	72.1	51.8	51.7	64.0	60.7
Multimodal-CoT	43.8	25.8	50.0	71.1	91.8	75.8	57.6	49.2	65.3	63.3
DDCoT	29.2	25.8	51.6	60.2	77.3	52.1	43.2	47.9	49.3	64.0
CCoT	25.0	26.6	35.9	12.5	87.5	56.5	39.3	53.4	54.0	64.4
Text-only Thought + Inference-time Scaling										
CoT	39.6	30.9	43.0	53.9	91.0	77.9	56.4	46.2	65.7	62.2
+Self-Consistency	37.5	35.9	51.6	62.5	93.4	76.0	57.2	52.1	62.3	64.6
+Best-of-N	43.8	39.8	53.1	60.9	93.4	81.0	56.0	53.4	67.3	65.5
+Beam Search	41.7	54.7	59.4	65.6	94.5	85.2	**61.1**	37.8	70.0	59.8
+MCTS	45.8	51.6	62.5	64.8	94.9	81.0	58.4	53.4	66.7	66.2
Multi-modal Thought + Inference-time Scaling										
Visual Sketchpad	33.3	81.3	86.1	85.9	93.0	81.0	53.7	49.1	66.3	58.2
+Self-Consistency	39.6	85.2	88.3	87.5	96.5	83.6	54.1	51.3	68.3	61.5
+Best-of-N	43.8	91.4	89.8	91.4	94.1	82.8	54.9	55.0	67.7	66.1
+Beam Search	**54.2**	**95.3**	92.2	**99.2**	96.5	**87.2**	58.4	**64.3**	71.3	64.8
+MCTS	52.1	94.5	**93.0**	**99.2**	**96.9**	86.7	60.3	62.2	**74.0**	**68.7**

Table 1: Accuracy across 10 datasets from three domains. For each dataset, the best result is highlighted in **bold**, and the second-best is emphasized with underlining.

Metrics We evaluate the performance using two metrics: **Accuracy** and **Pass@K**. Accuracy measures whether the highest-ranked answer is correct, while Pass@K represents whether the correct answer can be recalled within K answers.

Implementation Details We employ GPT-4o-mini[3] as the policy and verifier by default. For all experiments, we set the temperature to 1.0 during generation. For Self-Consistency and Best-of-N, the number of samples N is set to 5. For Beam Search, we set both the beam width B and expansion size N to 4. For MCTS, we set the maximum expansion per node to 4, the maximum number of iterations to 30, and the exploration constant w in UCB to 1.4. We adopt the consistency-enhanced verifier by default and set the number of samples N_v to 5.

3.2 Multi-modal vs. Text-only Thought

Multi-modal thought shows greater advantages compared with text-only thought. Table 1 presents compelling experimental results that highlight the superiority of multi-modal thought over both text-only thought and the latest multi-modal reasoning frameworks, particularly when enhanced by inference-time scaling.

A closer analysis reveals notable performance variations across different domains. In geometric reasoning, multi-modal thought consistently outperforms text-only thought under single-chain reasoning, with the exception of Geometry3K, where it initially lags but surpasses text-only thought when inference-time scaling is applied. In mathematical reasoning, the advantages of multi-modal thought become even more evident. For example, while text-only thought initially excels on Winner ID due to the problem's rich textual context, multi-modal thought progressively closes the gap under inference-time scaling. The similarly superior performance observed across VQA tasks further underscores the adaptability and effectiveness of multi-modal thought, strengthening its potential for broader applications.

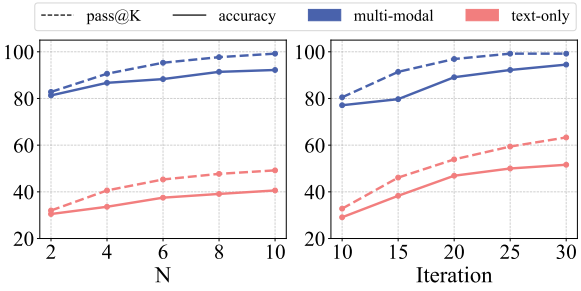

Figure 3: Performance comparison between text-only thought and multi-modal thought on the Maxflow dataset under Best-of-N (left) and MCTS (right).

Multi-modal thought exhibits a higher performance upper limit. To further investigate the potential of text-only and multi-modal thought, we compare their performance upper limits on the Maxflow dataset using Best-of-N and MCTS methods. As shown in Figure 3, within Best-of-N, accuracy consistently improves as N increases, with multi-modal thought achieving a significantly

[3] gpt-4o-mini-2024-07-18

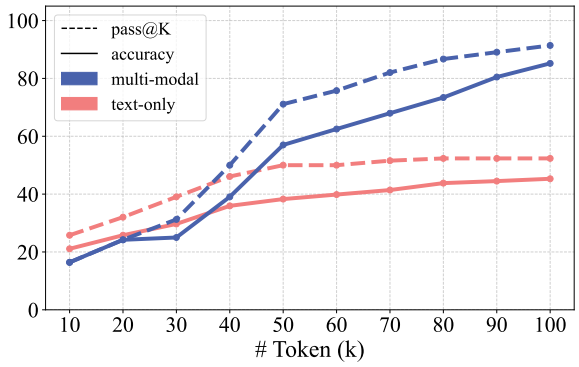

Figure 4: Performance comparison between text-only thought and multi-modal thought varies with the maximum token consumption on the Maxflow dataset under Self-Consistency.

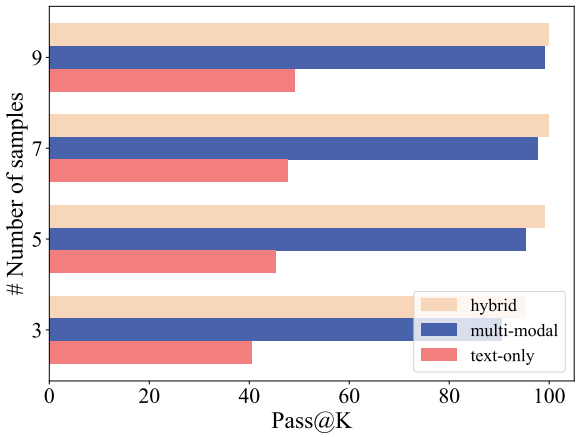

Figure 5: Performance of Self-Consistency on the Maxflow dataset as the number of samples increases, comparing text-only thought, multi-modal thought, and the hybrid form.

higher upper limit. This trend is further validated by the MCTS results, which exhibit a similar pattern, reinforcing the superior potential of multi-modal thought.

The effectiveness of multi-modal thought relies on higher token consumption. Given the substantial difference in token cost between processing text and image information in LVLMs, it is crucial to understand the trade-off between performance and token usage for text-only and multi-modal thought. To this end, we conduct experiments on the Maxflow dataset using Self-Consistency to analyze their performance under inference-time scaling with varying token limits per sample.

Figure 4 illustrates the relationship between token limits and performance for both text-only and multi-modal thought. Under strict token constraints, multi-modal thought underperforms compared to text-only thought. This is likely due to the limited number of reasoning steps multi-modal thought can sample when the token budget is severely restricted, preventing the model from effectively processing visual information.

However, as the token limits are relaxed, the performance of multi-modal thought improves rapidly. With a more generous token budget, the model can fully leverage visual information, leading to significant performance gains. This highlights a key trade-off: while multi-modal thought enables more complex reasoning, it requires significantly more tokens to process visual inputs effectively. This suggests that future research should prioritize developing methods for compressing or optimizing the token consumption associated with image processing in LVLMs, enabling more efficient utilization of multi-modal thought.

Multi-modal and text-only thought are complementary. While our results demonstrate the overall effectiveness of multi-modal thought, we further examine whether text-only thought retains value in certain scenarios. We hypothesize that the two approaches may be complementary, each leveraging different aspects of the input. To test this, we conduct experiments on the Maxflow dataset using the Self-Consistency method, comparing text-only thought, multi-modal thought, and a hybrid form, i.e., voting from both text-only thought and multi-modal thought.

As presented in Figure 5, the hybrid form outperforms both text-only and multi-modal thought individually, suggesting that text-only thought provides valuable insights that enhance reasoning, even in cases where multi-modal reasoning alone may fall short. This performance boost underscores the complementary nature of the two approaches and highlights the potential of a balanced integration of them for more robust multi-modal reasoning.

3.3 Ablation Study on the Verifier

Consistency-enhanced verifier outperforms naive prompting-based alternatives. To evaluate the superiority of our proposed consistency-enhanced verifier, we compare its performance against classification-based and regression-based verifiers, as described in §2.3. We conduct this comparison on the Geometry3K, Maxflow, and Isomorphism datasets using MCTS with multi-modal thought. As shown in Figure 6, our verifier consistently outperforms both naive prompting-based

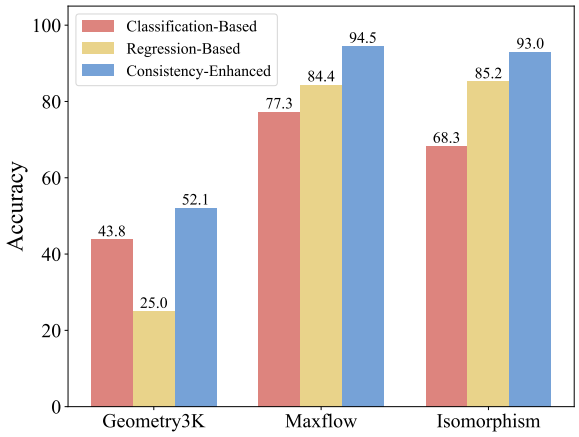

Figure 6: Performance comparison of Classification-Based, Regression-Based, and Consistency-Enhanced Verifiers using MCTS with multi-modal thought across three datasets.

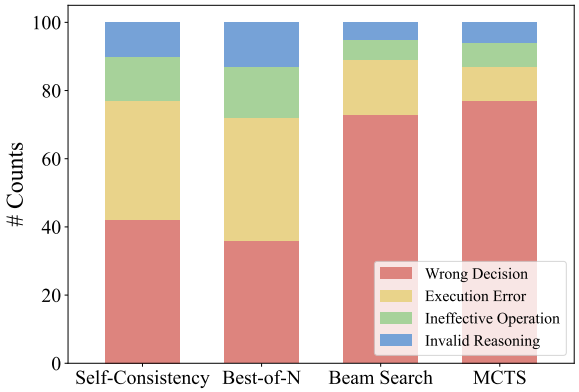

Figure 7: Error statistics on four inference-time scaling methods with multi-modal thought.

approaches across all three datasets, demonstrating its effectiveness in addressing the limitations of existing approaches and providing more reliable guidance for inference-time scaling.

Consistency-enhanced verifier performance scales with N_v. Our consistency-enhanced verifier is designed based on the hypothesis that aggregating multiple verifications enhances robustness and performance. Specifically, we expect accuracy to improve as the number of sampled verifications increases. To validate this, we conduct experiments using MCTS on the Geometry3K, Maxflow, and Isomorphism datasets, systematically varying N_v. From the results in Table 2, we observe a clear trend: increasing N_v consistently improves performance. This finding strongly supports our hypothesis, confirming the effectiveness of the design rationale for the verifier.

	Geometry3K	Maxflow	Isomorphism
N_v=3	41.7	89.8	91.4
N_v=5	52.1	94.5	93.0
N_v=7	58.3	97.7	96.1

Table 2: Performance variants with N_v under MCTS method with multi-modal thought.

3.4 Error Analysis

To gain deeper insights into the behavior of multi-modal thought under different inference-time scaling methods, we conduct a detailed error analysis. We identify four primary sources of error: Wrong Decision, Execution Error, Ineffective Operation, and Invalid Reasoning (detailed explanations are provided in Appendix E) and then randomly sample 100 error cases from each method across all datasets, manually classifying the underlying causes.

Tree search-based methods significantly reduce intermediate errors. Figure 7 presents the distribution of error types for various inference-time scaling methods. We can find that sampling-based methods exhibit a predominance of intermediate reasoning errors (Execution Error, Ineffective Operation, Invalid Reasoning). Among these, Execution Error is the most prevalent, which mostly occurs in VQA tasks. This is likely due to the difficulty expert models face in accurately operating images, coupled with a lack of error-handling logic in the code. Consequently, simple sampling methods often continue exploring flawed paths after encountering such errors, leading to the failure. In contrast, tree search-based methods significantly mitigate intermediate reasoning errors. By incorporating verifiers at each step, these methods strategically select optimal reasoning paths, effectively preventing the exploration of erroneous reasoning steps.

Effective verifiers are crucial for further error reduction. The reduction in intermediate errors leads to a relative increase of Wrong Decision, highlighting this error type as the new primary challenge. Further manual observation reveals that these errors primarily stem from limitations in the verifier. While effective at identifying blatant errors, the verifier struggles with more nuanced or complex reasoning steps. This underscores the critical importance of improving verifier effectiveness, particularly in discerning subtle errors, to fully unlock the potential of multi-modal thought.

Methods	Geometry3K	Maxflow	Connectivity
Qwen2-VL-72B-Instruct			
Direct	43.8	10.4	85.4
CoT	31.3	35.4	64.6
+Best-of-N	31.3	35.4	66.7
Visual Sketchpad	41.7	12.5	93.8
+Best-of-N	45.8	31.3	98.4
OPENAI-gpt-4o-mini			
Direct	35.4	28.9	52.3
CoT	39.6	30.9	53.9
+Best-of-N	43.8	39.8	60.9
Visual Sketchpad	33.3	81.3	85.9
+Best-of-N	43.8	91.4	91.4
OPENAI-gpt-4o			
Direct	52.1	50.0	71.1
CoT	54.2	55.5	76.6
+Best-of-N	70.8	67.2	90.6
Visual Sketchpad	62.5	93.8	98.4
+Best-of-N	77.1	96.9	99.2

Table 3: Accuracy on three datasets by employing different policy models.

3.5 Policy Model Impact

To assess the impact of the underlying policy model on the effectiveness of multi-modal thought, we extended our experiments beyond the previously used GPT-4o-mini to include a representative open-source LVLM Qwen2-VL-72B-Instruct (Wang et al., 2024d) and a more advanced closed-source LVLM GPT-4o[4]. The experiments are conducted on the Geometry3K, Maxflow, and Connectivity datasets, using Best-of-N as the inference-time scaling method.

Performance of multi-modal thought generalizes across models. As shown in Table 3, multi-modal reasoning consistently improves performance across all three evaluated LVLMs, suggesting that the benefits of multi-modal thought are not model-specific but rather a generalizable advantage across different architectures and training paradigms.

The stronger the model, the greater the performance gains. An intriguing trend also emerges from the results, i.e., the performance gains from multi-modal thought correlate with the strength of the policy model. Weaker models, such as Qwen2-VL, exhibit modest improvements, whereas the more powerful GPT-4o achieves significantly larger gains. This suggests that a model's capacity to effectively extract and integrate visual information is contingent upon its inherent capabilities. Stronger models, possessing superior reasoning abilities, are able to leverage multi-modal cues more efficiently, resulting in greater performance improvements under inference-time scaling.

[4] gpt-4o-2024-08-06

4 Discussion

The necessity of multi-modal thought. In this study, we empirically demonstrate that multi-modal thought possesses unique potential in representing cognitive processes that transcend the limitations of text alone. We argue that this approach can convey richer cognitive features, a perspective that aligns with the opinions from Chen et al. (2025). A representative example, illustrated in Figure 1, involves geometric reasoning. Here, visual-spatial elements (e.g., the use of auxiliary lines) provide critical insights that are inherently difficult to articulate through textual descriptions, thereby enhancing conceptual clarity. By integrating such multi-modal thought, the policy can engage in more informed decision-making processes, systematically improving problem-solving performance.

Reducing token costs for multi-modal thought to achieve efficient inference. Our analysis of token computations (Figure 4) demonstrates that multi-modal thought, while effective, incurs substantial token costs. This underscores the need for efficient compression techniques specifically designed for processing visual information in LVLMs. While prior works (Wang and Xuan, 2024; Huang et al., 2024; Zhu et al., 2024) have explored visual token compression, they have largely overlooked the unique challenges of multi-modal thought reasoning. Thus developing methods for efficient visual information handling in this context becomes a key direction for future research.

Constructing stronger multi-modal verifiers for better performance. Our error analysis (§3.4) highlights the critical role of the verifier in inference-time scaling, where its effectiveness directly influences error reduction and performance gains. However, research on multi-modal verifiers (Ouali et al., 2024; Xiyao et al., 2024; Yu et al., 2024; Zang et al., 2025) remains in its early stages, limiting the potential of multi-modal thought. Advancing more robust and accurate verifiers is therefore crucial for fully unlocking the benefits of inference-time scaling.

Exploring inference-time scaling for the chain of any-modal thought reasoning. Recent works

have extended CoT reasoning beyond text and images to more modalities, such as audio (Ma et al., 2025) and video (Fei et al., 2024). However, the effectiveness of inference-time scaling for these modalities remains unexplored. While our findings confirm the benefits of scaling for text-image reasoning, whether these advantages extend to other modality combinations remains an open question.

5 Conclusion

In this work, we present a systematic investigation of inference-time scaling with multi-modal thought. We conduct comprehensive experiments across 10 challenging tasks spanning diverse domains, investigating both sampling-based and tree search-based scaling methods. We demonstrate that multi-modal thought consistently outperform text-only reasoning, achieving progressively stronger performance with increased computational budgets. This empirical evidence positions multi-modal thought scaling as a promising direction for enhancing complex reasoning capabilities in multi-modal scenarios. However, we notice that performance gains are accompanied by higher computational costs and depend heavily on the effectiveness of the verifier, thus hindering real-world deployment. We hope this work will serve as a springboard for studying inference-time scaling with multi-modal thought, thereby advancing multi-modal reasoning.

Limitations

Due to the inability of current LVLMs to generate fine-grained images directly, we rely on generating image operation codes as a substitute, preventing a fully end-to-end multi-modal thought process. Additionally, we do not explore smaller-scale models, such as 7B, as their limited visual processing, instruction-following, and code generation abilities make them unsuitable for objectively evaluating multi-modal thought under inference-time scaling. Another limitation is the exclusion of training-based methods that require long CoT data, as current LVLMs struggle to generate high-quality long reasoning chains over multi-modal datasets, making data construction a significant challenge. Lastly, our study focuses solely on text and image modalities, leaving the generalizability of our findings to other modalities, such as audio and video, an open question.

Acknowledgments

The project was supported by National Key R&D Program of China (No. 2022ZD0160501), Natural Science Foundation of Fujian Province of China (No. 2024J011001), and the Public Technology Service Platform Project of Xiamen (No.3502Z20231043). We also thank the reviewers for their insightful comments.

References

Jinhe Bi, Danqi Yan, Yifan Wang, Wenke Huang, Haokun Chen, Guancheng Wan, Mang Ye, Xun Xiao, Hinrich Schuetze, Volker Tresp, and Yunpu Ma. 2025. Cot-kinetics: A theoretical modeling assessing lrm reasoning process. Preprint, arXiv:2505.13408.

Guoxin Chen, Minpeng Liao, Chengxi Li, and Kai Fan. 2024a. Alphamath almost zero: Process supervision without process. In Advances in Neural Information Processing Systems, volume 37, pages 27689–27724. Curran Associates, Inc.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. 2025. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. arXiv preprint arXiv:2503.09567.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 2024b. M3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8199–8221.

Zihui Cheng, Qiguang Chen, Jin Zhang, Hao Fei, Xiaocheng Feng, Wanxiang Che, Min Li, and Libo Qin. 2025. Comt: A novel benchmark for chain of multi-modal thought on large vision-language models. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, pages 23678–23686.

Ruomeng Ding, Chaoyun Zhang, Lu Wang, Yong Xu, Minghua Ma, Wei Zhang, Si Qin, Saravan Rajmohan, Qingwei Lin, and Dongmei Zhang. 2024. Everything of thoughts: Defying the law of penrose triangle for thought generation. In Findings of the Association for Computational Linguistics: ACL 2024, pages 1638–1662, Bangkok, Thailand. Association for Computational Linguistics.

Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. 2024. Video-of-thought: Step-by-step video reasoning from perception to cognition. In Forty-first International Conference on Machine Learning.

Deqing Fu, Ruohao Guo, Ghazal Khalighinejad, Ollie Liu, Bhuwan Dhingra, Dani Yogatama, Robin Jia,

and Willie Neiswanger. 2024. Isobench: Benchmarking multimodal foundation models on isomorphic representations. arXiv preprint arXiv:2404.01266.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2025. Blink: Multimodal large language models can see but not perceive. In European Conference on Computer Vision, pages 148–166. Springer.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. 2024. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. arXiv preprint arXiv:2406.09403.

Minbin Huang, Runhui Huang, Han Shi, Yimeng Chen, Chuanyang Zheng, Xiangguo Sun, Xin Jiang, Zhenguo Li, and Hong Cheng. 2024. Efficient multimodal large language models via visual token grouping. arXiv preprint arXiv:2411.17773.

Jaylen Jones, Lingbo Mo, Eric Fosler-Lussier, and Huan Sun. 2024. A multi-aspect framework for counter narrative evaluation using large language models. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 147–168, Mexico City, Mexico. Association for Computational Linguistics.

Daniel Kahneman. 2011. Thinking, fast and slow. Farrar, Straus and Giroux.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4015–4026.

Levente Kocsis and Csaba Szepesvári. 2006. Bandit based monte-carlo planning. In European conference on machine learning, pages 282–293. Springer.

Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, Lei Zhang, and Jianfeng Gao. 2024a. Segment and recognize anything at any granularity. In European Conference on Computer Vision, pages 467–484. Springer.

Shuangtao Li, Shuaihao Dong, Kexin Luan, Xinhan Di, and Chaofan Ding. 2025. Enhancing reasoning through process supervision with monte carlo tree search. arXiv preprint arXiv:2501.01478.

Zhiyuan Li, Hong Liu, Denny Zhou, and Tengyu Ma. 2024b. Chain of thought empowers transformers to solve inherently serial problems. In The Twelfth International Conference on Learning Representations.

Jiacheng Liu, Andrew Cohen, Ramakanth Pasunuru, Yejin Choi, Hannaneh Hajishirzi, and Asli Celikyilmaz. 2024a. Don't throw away your value model! generating more preferable text with value-guided monte-carlo tree search decoding. In First Conference on Language Modeling.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. 2024b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In European Conference on Computer Vision, pages 38–55. Springer.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv preprint arXiv:2310.02255.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-chun Zhu. 2021. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 6774–6786.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. Advances in Neural Information Processing Systems, 35:2507–2521.

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, et al. 2024. Improve mathematical reasoning in language models by automated process supervision. arXiv preprint arXiv:2406.06592.

Ziyang Ma, Zhuo Chen, Yuping Wang, Eng Siong Chng, and Xie Chen. 2025. Audio-cot: Exploring chain-of-thought reasoning in large audio language model. arXiv preprint arXiv:2501.07246.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. 2024. Compositional chain-of-thought prompting for large multimodal models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14420–14431.

Yassine Ouali, Adrian Bulat, Brais Martinez, and Georgios Tzimiropoulos. 2024. Clip-dpo: Vision-language models as a source of preference for fixing hallucinations in lvlms. In European Conference on Computer Vision, pages 395–413. Springer.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. Reflexion: Language agents with verbal reinforcement learning. Advances in Neural Information Processing Systems, 36.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of go with deep neural networks and tree search. nature, 529(7587):484–489.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. 2017. Mastering the game of go without human knowledge. nature, 550(7676):354–359.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. arXiv preprint arXiv:2408.03314.

I Sutskever. 2014. Sequence to sequence learning with neural networks. arXiv preprint arXiv:1409.3215.

Ye Tian, Baolin Peng, Linfeng Song, Lifeng Jin, Dian Yu, Lei Han, Haitao Mi, and Dong Yu. 2024a. Toward self-improvement of llms via imagination, searching, and criticizing. In Advances in Neural Information Processing Systems, volume 37, pages 52723–52748. Curran Associates, Inc.

Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas Griffiths, and Faeze Brahman. 2024b. MacGyver: Are large language models creative problem solvers? In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 5303–5324, Mexico City, Mexico. Association for Computational Linguistics.

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

Ziyu Wan, Xidong Feng, Muning Wen, Stephen Marcus McAleer, Ying Wen, Weinan Zhang, and Jun Wang. 2024. Alphazero-like tree-search can guide large language model decoding and training. In Forty-first International Conference on Machine Learning.

Ante Wang, Linfeng Song, Ye Tian, Baolin Peng, Dian Yu, Haitao Mi, Jinsong Su, and Dong Yu. 2024a. Litesearch: Efficacious tree search for llm. arXiv preprint arXiv:2407.00320.

Ante Wang, Linfeng Song, Ye Tian, Dian Yu, Haitao Mi, Xiangyu Duan, Zhaopeng Tu, Jinsong Su, and Dong Yu. 2025. Don't get lost in the trees: Streamlining llm reasoning by overcoming tree search exploration pitfalls. arXiv preprint arXiv:2502.11183.

Jun Wang, Meng Fang, Ziyu Wan, Muning Wen, Jiachen Zhu, Anjie Liu, Ziqin Gong, Yan Song, Lei Chen, Lionel M Ni, et al. 2024b. Openr: An open source framework for advanced reasoning with large language models. arXiv preprint arXiv:2410.09671.

Ke Wang and Hong Xuan. 2024. Llava-zip: Adaptive visual token compression with intrinsic image information. arXiv preprint arXiv:2412.08771.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024c. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9426–9439.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024d. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35:24824–24837.

Sean Welleck, Jiacheng Liu, Ximing Lu, Hannaneh Hajishirzi, and Yejin Choi. 2022. Naturalprover: Grounded mathematical proof generation with language models. Advances in Neural Information Processing Systems, 35:4913–4927.

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

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. 2024. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models. arXiv preprint arXiv:2408.00724.

Yonghui Wu. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.

Violet Xiang, Charlie Snell, Kanishk Gandhi, Alon Albalak, Anikait Singh, Chase Blagden, Duy Phung, Rafael Rafailov, Nathan Lile, Dakota Mahan, et al. 2025. Towards system 2 reasoning in llms: Learning how to think with meta chain-of-though. arXiv preprint arXiv:2501.04682.

Yuxi Xie, Anirudh Goyal, Wenyue Zheng, Min-Yen Kan, Timothy P Lillicrap, Kenji Kawaguchi, and Michael Shieh. 2024. Monte carlo tree search boosts

reasoning via iterative preference learning. arXiv preprint arXiv:2405.00451.

Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, James Xu Zhao, Min-Yen Kan, Junxian He, and Michael Xie. 2023. Self-evaluation guided beam search for reasoning. Advances in Neural Information Processing Systems, 36:41618–41650.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. 2024. Llava-critic: Learning to evaluate multimodal models. arXiv preprint arXiv:2410.02712.

Wang Xiyao, Yang Zhengyuan, Li Linjie, Lu Hongjin, Xu Yuancheng, Lin Chung-Ching Lin, Lin Kevin, Huang Furong, and Wang Lijuan. 2024. Scaling inference-time search with vision value model for improved visual comprehension. arXiv preprint arXiv:2412.03704.

Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. 2024. Llava-o1: Let vision language models reason step-by-step. arXiv preprint arXiv:2411.10440.

Haotian Xu. 2023. No train still gain. unleash mathematical reasoning of large language models with monte carlo tree search guided by energy function. arXiv preprint arXiv:2309.03224.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. 2024. Depth anything: Unleashing the power of large-scale unlabeled data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10371–10381.

Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, et al. 2024a. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. arXiv preprint arXiv:2412.18319.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024b. Tree of thoughts: Deliberate problem solving with large language models. Advances in Neural Information Processing Systems, 36.

Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. 2024. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. arXiv preprint arXiv:2405.17220.

Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Ziyu Liu, Shengyuan Ding, Shenxi Wu, Yubo Ma, Haodong Duan, Wenwei Zhang, et al. 2025. Internlm-xcomposer2. 5-reward: A simple yet effective multi-modal reward model. arXiv preprint arXiv:2501.12368.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. 2024a. Generative verifiers: Reward modeling as next-token prediction. arXiv preprint arXiv:2408.15240.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. 2024b. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In European Conference on Computer Vision, pages 169–186. Springer.

Zhuosheng Zhang, Aston Zhang, Mu Li, George Karypis, Alex Smola, et al. 2023. Multimodal chain-of-thought reasoning in language models. Transactions on Machine Learning Research.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. 2023. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. Advances in Neural Information Processing Systems, 36:5168–5191.

Qiji Zhou, Ruochen Zhou, Zike Hu, Panzhong Lu, Siyang Gao, and Yue Zhang. 2024. Image-of-thought prompting for visual reasoning refinement in multimodal large language models. arXiv preprint arXiv:2405.13872.

Yuke Zhu, Chi Xie, Shuang Liang, Bo Zheng, and Sheng Guo. 2024. Focusllava: A coarse-to-fine approach for efficient and effective visual token compression. arXiv preprint arXiv:2411.14228.

A Implementation Details of Multi-modal Thought

Recognizing the limitations of current generative models in fine-grained drawing, we leverage Visual Sketchpad (Hu et al., 2024) to enable multi-modal reasoning. Visual Sketchpad empowers LVLMs to generate sketches as intermediate reasoning artifacts, boosting problem-solving capabilities. It interprets multi-modal queries, devises sketching strategies, synthesizes programs for sketch generation, and analyzes the resulting sketches to formulate responses.

Specifically, at each step i, the model outputs a textual thought $\mathbf{t}_i$ and a visual operation o_i:

$$\mathbf{t}_i, o_i = \mathcal{M}\big(\mathbf{q}, \mathbf{I}, \mathbf{s}_{1:i-1}\big), \tag{5}$$

where $\mathbf{o}_i$ is executable code specifying visual operations.

A code executor $\mathcal{G}$ then updates the visual contexts by applying o_i to the previous images $\mathbf{I}$:

$$\mathbf{I}' = \mathcal{G}\big(\mathbf{I}, o_i\big), \tag{6}$$

Dataset	Description	Size
Maxflow	Find the maximum flow through a network.	128
Isomorphism	Determine structural equivalence of two graphs.	128
Connectivity	Determine path existence between two graph vertices.	128
Convexity	Determine whether a function is convex or concave.	256
Parity	Determine if a function is even, odd, or neither.	384
Winner ID	Determine chess game outcome from the final board state.	257
V^* Bench	Questions about small items in an image.	238
MMVP	Visual questions designed to reveal CLIP-based LM shortcomings.	300
BLINK	Visual perception tasks challenging for multi-modal LMs.	728

Table 4: The detailed introduction of the datasets in our experiments.

$$\mathbf{I} = \mathbf{I} \cup \mathbf{I}', \tag{7}$$

where $\mathbf{I}'$ represents the updated images. Thus the next reasoning step $\mathbf{s}_i = \{\mathbf{t}_i, o_i, \mathbf{I}'\}$.

Visual Sketchpad integrates various tools for sketch generation. For geometric and mathematical tasks, it uses Python libraries like matplotlib and networkx. For VQA, it incorporates specialized vision models and visual manipulations, including:

- **Detection:** Use Grounding-DINO (Liu et al., 2024b) to detect and label objects in images based on textual queries.
- **Segmentation:** Generate segmented images with labeled masks by employing SegmentAnything (Kirillov et al., 2023) and Semantic-SAM (Li et al., 2024a).
- **Depth Estimation:** Leverage DepthAnything (Yang et al., 2024) to produce depth maps from input images.
- **Visual Search:** Implements a sliding window approach to locate small objects based on textual queries.
- **Image Manipulation:** Include zoom/crop (region of interest extraction) and image overlay (alpha blending).

B Prompts for Verifiers

We use the following prompts for two kinds of verifiers:

- **Classification-Based:** *Verify the reasoning process above and provide the final judgment of 'yes' or 'no' on whether the reasoning is valid at last after "Final Decision:".*
- **Regression-Based:** *Verify the reasoning process above and provide a validation score from 0 (worst) to 1.0 (best) at last after "Final Score:".*

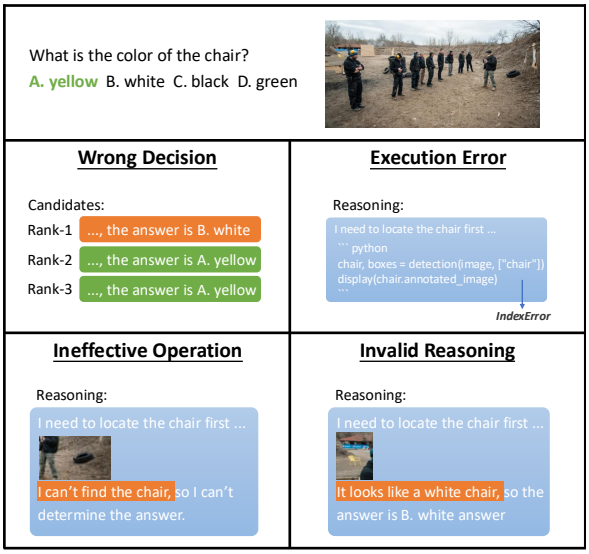

Figure 8: Examples of each error category.

C Datasets

The detailed introduction of datasets in our experiments are presented in Table 4.

D Baseline Introduction

Here we introduce the latest multi-modal reasoning baselines in our experiments:

- **Multimodal-CoT** (Zhang et al., 2023): A multi-modal extension of CoT that concate-

nates the reasoning process with the problem context for answering.

- **DDCoT** (Zheng et al., 2023): A method that decouples reasoning and recognition, and then integrate visual capabilities into the reasoning process.
- **CCoT** (Mitra et al., 2024) A framework that generates a scene graph before answering.

E Error Category Explanation

We define four categories of error of multi-modal thought under inference-time scaling methods as follows:

- **Wrong Decision:** The correct answer is present within the candidate reasoning chains but is not selected. This indicates an error in the decision-making process.
- **Execution Error:** The code generated during the reasoning process contains bugs, preventing the reasoning process from proceeding correctly.
- **Ineffective Operation:** The reasoning process is hampered by ineffective operations that produce unhelpful or misleading visual information.
- **Invalid Reasoning:** The error stems from a flawed reasoning process, leading to an incorrect conclusion.

Examples of each error category are provided in Figure 8.

F Additional Experiments

F.1 Test Results on M^{3}CoT

To validate the effectiveness on more challenging datasets, we conduct evaluations on M^{3}CoT (Chen et al., 2024b) datasets using GPT-4o, as presented in Table 5. These results demonstrate that the scaling of multi-modal thought continues to yield performance improvements in more complex scenarios.

F.2 The Influence of the Image Quality

To investigate the impact of image quality in the reasoning chain, we conduct additional experiments in which the resolution of intermediate images is artificially degraded by a scaling factor α (where a larger α corresponds to lower resolution and poorer quality). The results are reported in Table 6.

These results show that reduced image quality (i.e., increased α) leads to a clear degradation in reasoning performance and we also observe a qualitative reduction in inference latency with lower-resolution images. This suggests a potential trade-off between reasoning accuracy and computational efficiency, which merits further exploration.

Methods	M^{3}CoT
Directly	64.4
CoT	65.0
+Self-Consistency	68.2
+Best-of-N	68.8
Visual Sketchpad	66.4
+Self-Consistency	69.0
+Best-of-N	70.4

Table 5: Performance on M^{3}CoT using GPT-4o.

Methods	BLINK
Visual Sketchpad	58.2
+α=0.2	57.1
+α=0.4	54.8
+α=0.6	47.4

Table 6: Performance on BLINK using multi-modal thought with varying intermediate image resolution scale factors (α).