Mulberry: Empowering MLLM with o1-like Reasoning and Reflection via Collective Monte Carlo Tree Search

Huanjin Yao[2,3,*] **Jiaxing Huang**[1,*,✉] **Wenhao Wu**[3] **Jingyi Zhang**[1] **Yibo Wang**[2] **Shunyu Liu**[1] **Yingjie Wang**[1]
Yuxin Song[3] **Haocheng Feng**[3] **Li Shen**[4] **Dacheng Tao**[1]

Abstract

In this work, we aim to develop an MLLM that understands and solves questions by learning to create each intermediate step of the reasoning involved till the final answer. To this end, we propose Collective Monte Carlo Tree Search (CoMCTS), a new learning-to-reason method for MLLMs, which introduces the concept of collective learning into "tree search" for effective and efficient reasoning-path searching and learning. The core idea of CoMCTS is to leverage collective knowledge from multiple models to collaboratively conjecture, search and identify effective reasoning paths toward correct answers via four iterative operations including Expansion, Simulation and Error Positioning, Backpropagation, and Selection. Using CoMCTS, we construct Mulberry-260k, a multimodal dataset with a tree of rich, explicit and well-defined reasoning nodes for each question. With Mulberry-260k, we perform collective SFT to train our model, Mulberry, a series of MLLMs with o1-like step-by-step Reasoning and Reflection capabilities. Extensive experiments demonstrate the superiority of our proposed methods on various benchmarks. Code will be available at `https://github.com/HJYao00/Mulberry`

1. Introduction

"What I cannot create, I do not understand."
—Richard Feynman

Multimodal large language models (MLLMs) embody the essence of this dictum, which understand the world by learning to create expected responses to multimodal inputs such as images and text. While MLLMs have recently shown significant progress in straightforward tasks (Liu et al., 2024; Wang et al., 2024b), they often experience obviously increased failures on complex tasks requiring in-depth reasoning (Zhang et al., 2024d). Feynman's dictum might be the perfect metaphor of such failures of MLLMs, as we should only be able to work something out if we can create and have a firm understanding of each step of the reasoning involved. However, current MLLMs predominantly operate in a simple "direct prediction" mode (Xu et al., 2024), *i.e.*, generating brief, final answers to questions with little explicit and well-defined intermediate reasoning steps.

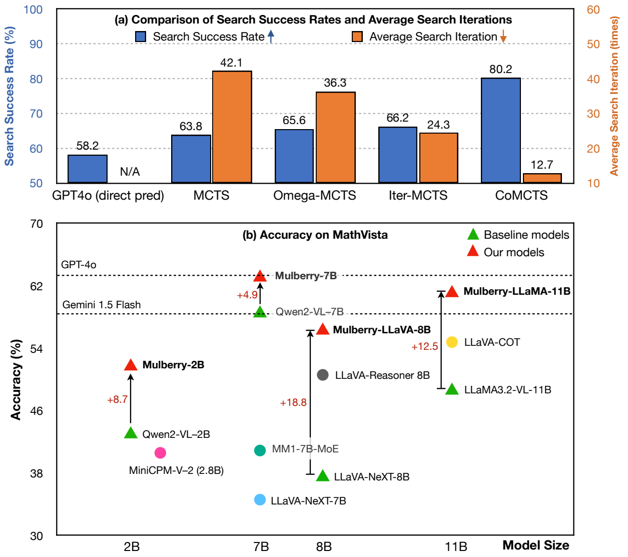


Figure 1: **(a)** Our CoMCTS shows great superiority in search effectiveness and efficiency against other tree search methods. **(b)** Our Mulberry, trained on CoMCTS-searched data, outperforms most open-sourced MLLMs and achieves competitive results against closed-source ones, showing outstanding abilities in step-by-step reasoning and reflection.

In this work, we aim to develop an MLLM that understands and solves questions by learning to create each intermediate step of the reasoning involved till the final answer. Recent advances in NLP, such as OpenAI o1 (OpenAI, 2024), have

---

[*] Equal Contribution. Correspondence to: Jiaxing Huang <jiaxing.huang@ntu.edu.sg>. [1] Nanyang Technological University; [2] Tsinghua University; [3] Baidu Inc.; [4] Sun Yat-sen University.

shown great potential in enabling LLM to learn to reason and tackle complex language tasks (Xie et al., 2024). The core design of these advances lies in AlphaGo-like "tree search": they employ tree search methods, like MCTS (Coulom, 2006), to bootstrap an LLM itself to build a tree of intermediate thoughts, explore effective reasoning paths, and leverage these paths to teach the model to reason step-by-step.

An intuitive idea is to directly apply these "tree search" methods to search effective reasoning paths for MLLMs, which, however, does not work well. As illustrated in Figure 1, we believe this is largely attributed to several observed search challenges for MLLMs. (1) *Search Effectiveness:* Traditional MCTS methods generally work by self-bootstrapping while current MLLMs are typically trained with little explicit and well-defined intermediate reasoning steps, making these search methods often trapped in homogeneous low-quality nodes within the reasoning space of a single MLLM, ultimately leading to low search success rates. (2) *Search Efficiency:* Traditional MCTS methods typically expand and explore only one subsequent reasoning node per search iteration, which advance a single step each time and demand massive iterations, making them inefficient for computation-intensive MLLMs.

To tackle these challenges, we propose Collective Monte Carlo Tree Search (CoMCTS), a new learning-to-reason method for MLLMs, which introduces the concept of collective learning into "tree search" for effective and efficient reasoning-path searching and learning. The core idea of CoMCTS is to leverage collective knowledge to collaboratively conjecture, search and identify effective reasoning paths toward correct answers. Specifically, CoMCTS searches effective reasoning paths iteratively, and in each iteration, it leverages collective knowledge from multiple MLLMs to jointly (a) expand diverse and complementary candidate subsequent reasoning nodes till the end from a given start node, (b) simulate reasoning outcomes, position error candidate nodes and prune them along with their child nodes, (c) backpropagate to update the score and visit count of each reasoning node in a bottom-up manner, and (d) select the leaf reasoning node with the highest Upper Confidence Bound value as next start node.

In this way, our CoMCTS achieves effective and efficient reasoning search. (1) The joint expansion mechanism enables CoMCTS to concatenate reasoning trajectories from multiple MLLMs via iterative search, ultimately constructing an unified reasoning tree comprising diverse and complementary reasoning nodes. Thus, it allows reasoning-path search not only within the reasoning space of a given MLLM itself but also among those of others, benefiting from the synergy of multiple MLLMs while avoiding being trapped in homogeneous low-quality nodes within the reasoning space of a single MLLM itself. (2) The joint simulation and error positioning mechanism enables CoMCTS to, in each search iteration, skip multiple intermediate steps and select the last correct step as the next start node, largely reducing search time while maintaining search effectiveness. Here, collective knowledge is also crucial as it is often challenging for a model to recognize and position errors made by itself while relatively easy by using other models.

Furthermore, we extend our CoMCTS for reflective reasoning-path search. Based on the unified reasoning tree constructed by CoMCTS, which provides both positive and negative reasoning nodes , we identify and integrate negative sibling nodes into effective reasoning paths to build the reflective reasoning path that includes a transition from a negative reasoning node to a positive one. By learning from reflective reasoning paths, MLLMs can perform appropriate step-wise reflection, dynamically calibrating their reasoning trajectory from an erroneous node toward a correct one during long-chain reasoning. Here, collective knowledge facilitates reflective reasoning-path search by providing a rich set of diverse positive and negative reasoning nodes.

Using our CoMCTS, we search effective and reflective reasoning paths for a set of multimodal inputs, and construct Mulberry-260k, a Multimodal learning-to-Reason-and-Reflect dataset with a tree of rich, explicit and well-defined reasoning nodes for each question. With Mulberry-260k, we perform collective supervised fine-tuning to train our model, Mulberry, a series of Multimodal LLMs with o1-like step-by-step Reasoning and Reflection capabilities.

The main contributions of this work are fourfold. ***First***, we introduce the concept of collective learning into MCTS, and propose CoMCTS which leverages collective knowledge to collaboratively conjecture, search and identify effective and reflective reasoning paths for MLLMs, showing great superiority in search effectiveness and efficiency. To the best of our knowledge, this is the first work that explores collective learning with MCTS for MLLMs. ***Second***, we construct Mulberry-260k that provides a valuable resource for advancing research in step-by-step reasoning and reflection in MLLMs. ***Third***, we develop Mulberry, a series of MLLMs with outstanding capabilities in step-by-step reasoning and reflection. ***Fourth***, extensive experiments demonstrate the superiority of our proposed methods on various benchmarks.

2. Related Works

2.1. Multimodal Large Language Model

MLLMs (Liu et al., 2024; Wang et al., 2024b; Lu et al., 2024a; Yao et al., 2024a) have made notable advancements in general vision-language understanding, enabling them to interpret visual semantics across various domains. Recent studies (Yue et al., 2024; Zhang et al., 2024d) explore MLLM reasoning and reveal that directly employing CoT prompt to derive the final answer may result in limited gains or even degradation. In addition, some studies (Mitra et al., 2024; Luan et al., 2024) introduce plan-based CoT prompting to guide models to generate intermediate information for predicting final answers. Recent advances (Xu et al., 2024) attempt structured reasoning with a planed flow of certain pre-defined stages, enhancing the CoT capabilities (Zhang et al., 2024c) of MLLMs. Differently, this paper, for the first time, introduces the concept of "tree search" into MLLM reasoning and proposes a novel CoMCTS technique to search effective and reflective reasoning paths to train our Mulberry, a series of MLLMs with outstanding capabilities in step-by-step reasoning and reflection.

2.2. Large Language Model Reasoning

LLM reasoning methods can be broadly categorized into three types, including prompt-based, plan-based and learning-based reasoning. Prompt-based methods, like Chain-of-Thought (CoT) (Wei et al., 2022), mimic human reasoning by providing a few hand-crafted, step-by-step solutions as references. Plan-based methods, such as Tree/Graph-of-thought (Yao et al., 2024b; Besta et al., 2024), predict multiple reasoning paths in a tree or graph manner and take consistent units of thought for thoughtful decision-making. Learning-based reasoning methods, represented by GPTo1, Star (Zelikman et al., 2022), Iter-MCTS (Xie et al., 2024) and ReST-MCTS (Zhang et al., 2024a), first employ tree search approaches, like MCTS, to bootstrap an LLM itself to build a tree of intermediate thoughts, explore effective reasoning paths, and leverage these paths to train the model to reason step-by-step.

2.3. Monte-Carlo Tree Search

Monte-Carlo Tree Search (MCTS) is a powerful search paradigm for complex decision making problems and has been extensively explored across diverse fields, including games (Silver et al., 2017; Ye et al., 2021), robotics (Best et al., 2019; Dam et al., 2022), theorem proving (Lample et al., 2022), matrices multiplication (Fawzi et al., 2022), etc. For instance, AlphaGo (Silver et al., 2017) introduces deep learning into MCTS, achieving superhuman performance in board and video games (Silver et al., 2017; Ye et al., 2021). Besides, (Pitanov et al., 2023; Yang, 2023) explore MCTS for path finding and train timetabling problems, while (Vagadia et al., 2024) integrates MCTS into physics-informed planning networks for robot control. In this work, we propose CoMCTS that enables effective and reflective reasoning-path searching and learning on MLLMs.

2.4. Collective Learning

Collective learning, also known as Co-training, aims to harness collective intelligence of multiple individuals to improve learning outcomes. This concept originates in early pioneering studies (Blum & Mitchell, 1998; Sun & Jin, 2011; Yu et al., 2011), which utilize collective knowledge to address data insufficiency issues in classification learning. Recent advances introduce collective learning into deep neural networks for efficient and effective deep learning. For example, (Qiao et al., 2018; Saito et al., 2018) employ collective knowledge from multiple classifiers to predict more accurate pseudo-labels for semi-supervised classification; (Cui et al., 2022) utilizes collective knowledge from multiple discriminators to enhance image discrimination and generation; and (Foerster et al., 2016) leverages the synergy of multiple models for reinforcement learning.

3. Methodology

We first present our proposed CoMCTS that introduces the concept of collective learning into "tree search" for effective and efficient reasoning-path searching and learning. We then illustrate the extension of CoMCTS for reflective reasoning-path search, and describe data construction (*i.e.*, Mulberry-260k) and model training (*i.e.*, Mulberry) using CoMCTS. More details to be elaborated in the ensuing subsections.

3.1. CoMCTS for effective reasoning

The core idea of CoMCTS is to leverage collective knowledge to collaboratively conjecture, search and identify effective reasoning nodes in an iterative manner, aiming to find effective reasoning paths leading to correct answers.

We denote a policy model as π, which is initialized by a pre-trained MLLM. We leverage collective knowledge from a group of MLLMs $\{\pi_1, \pi_2, ..., \pi_K\}$ to jointly search and learn effective reasoning paths. Given a multimodal input question Q (*e.g.*, a text task instruction with an image, $Q = \{\text{text}, \text{image}\}$), each model π can generate a sequence of intermediate reasoning states toward the final answer $(s_1, s_2, s_3, ..., s_M) \sim \pi_\theta(\cdot|Q)$ via autoregressive next token prediction. We define the intermediate reasoning state at step m as s_m and the state generated by model π_k at step m as s_m^k. Each reasoning step consists of one or a few sentences containing multiple word tokens.

CoMCTS algorithm begins at the root node, *i.e.*, either the start of a response or an incomplete response, and performs reasoning-path search via a certain number of iterations, where each iteration comprises four key operations: (a) Expansion, (b) Simulation and Error Positioning, (c) Backpropagation, and (d) Selection, as elaborated below.

(a) Expansion. The goal of this operation in CoMCTS

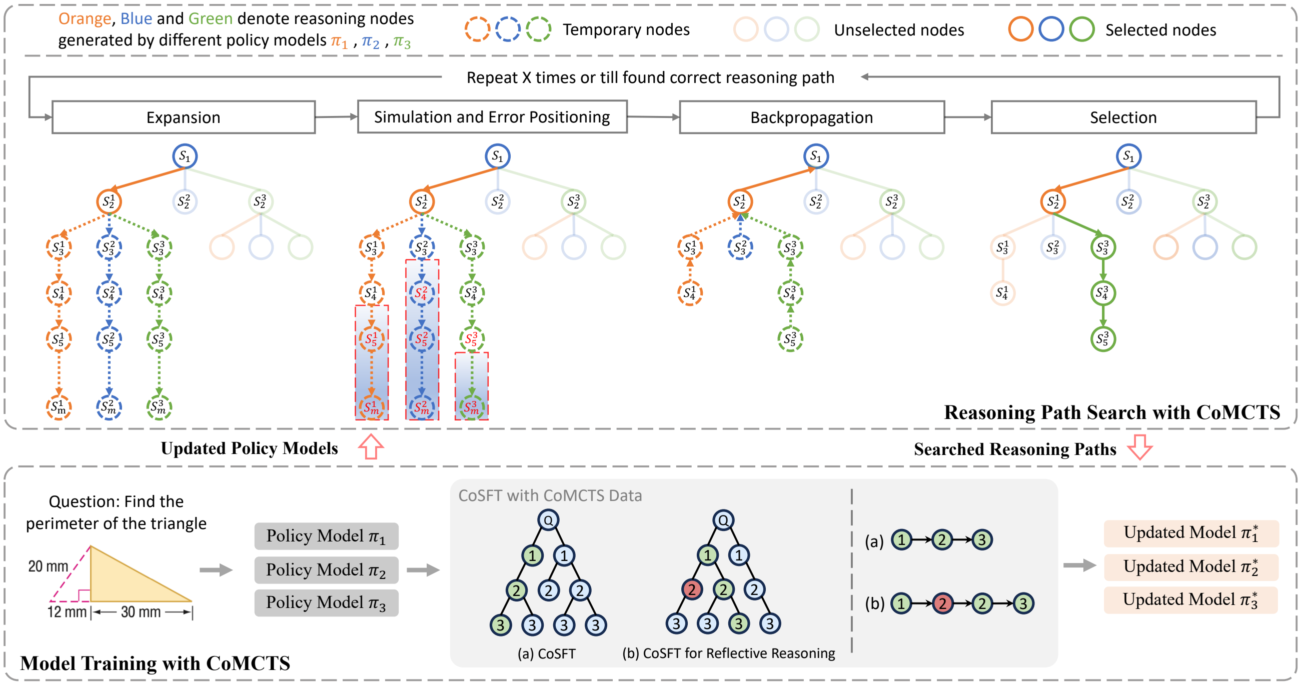

Orange, Blue and Green denote reasoning nodes generated by different policy models π_1, π_2, π_3
Temporary nodes
Unselected nodes
Selected nodes
Repeat X times or till found correct reasoning path
Expansion
Simulation and Error Positioning
Backpropagation
Selection
Reasoning Path Search with CoMCTS
Updated Policy Models
Searched Reasoning Paths
Question: Find the perimeter of the triangle
20 mm
12 mm
30 mm
Policy Model π_1
Policy Model π_2
Policy Model π_3
CoSFT with CoMCTS Data
(a) CoSFT
(b) CoSFT for Reflective Reasoning
Updated Model π_1^*
Updated Model π_2^*
Updated Model π_3^*
Model Training with CoMCTS

Figure 2: **Overview.** Our CoMCTS trains Mulberry with two alternating phases. *In top part,* CoMCTS searches reasoning paths iteratively, and in each iteration, it utilizes collective knowledge from multiple MLLMs to jointly (a) expand diverse and complementary candidate subsequent reasoning nodes till the end from a given start node, (b) simulate reasoning outcomes, position error candidate nodes and prune them along with their child nodes, (c) backpropagate to update the score and visit count of each reasoning node in a bottom-up manner, and (d) select the leaf reasoning node with the highest UCB value as next start node. *In bottom part,* we train the model to learn from the reasoning trees constructed by CoMCTS.

is to expand the current leaf reasoning node (if it is not a terminal node) to integrate new subsequent candidate reasoning nodes. Given the current leaf node s_m^k (*i.e.*, the node selected by Operation (d) Selection or the root node), CoMCTS utilizes collective knowledge from a group of MLLMs, $\{\pi_1, \pi_2, ..., \pi_K\}$, to jointly expand a set of diverse and complementary candidate reasoning paths $S_{\text{candidate}} = \cup_{j=1}^{K} S_{\text{candidate}}^j$ in parallel till the terminal node:

$$S_{\text{candidate}}^j \sim \pi_j(\cdot|Q, \text{Parent}(s_m^k), s_m^k), \tag{1}$$

where $\text{Parent}(s_m^k)$ returns all parent nodes of s_m^k and $(\text{Parent}(s_m^k), s_m^k)$ denotes the current reasoning path from the root node to s_m^k. $S_{\text{candidate}}^j = \{s_i^j\}$ stands for a potential reasoning path generated by model π_j starting from s_m^k.

(b) Simulation and Error Positioning. In this operation, CoMCTS utilizes collective knowledge from $\{\pi_1, \pi_2, ..., \pi_K\}$ to jointly estimate the potential value of child nodes $s_i^j \in S_{\text{candidate}}$ (added in Operation (a)), and considers low-score nodes as erroneous reasoning nodes, and positions and filters out them along with their child nodes:

$$R(s_i^j) = \frac{1}{K}\sum_{l=1}^{K} \pi_l(\cdot|\text{prompt}_{\text{eval}}, Q, \text{Parent}(s_i^j), s_i^j) \tag{2}$$

$$S_{\text{candidate}}^* = \{s_i^j \in S_{\text{candidate}} | R(s_i^j) >= t\} \tag{3}$$

where $R(s_i^j)$ denotes a reasoning node evaluation function that uses the prompt, $\text{prompt}_{\text{eval}}$, to request a group of MLLMs, $\{\pi_1, \pi_2, ..., \pi_K\}$, to jointly evaluate the candidate reasoning node s_i^j. t is a threshold and Discontinued reasoning nodes in $S_{\text{candidate}}^*$ are automatically removed following the error node removal in Eq.(3).

(c) Backpropagation. Given the new reasoning tree expanded and simulated using collective knowledge in Operations (a)-(b), CoMCTS performs a bottom-up update from the leaf nodes back to the root node. Each node s along the newly expanded path in the reasoning tree updates its statistics, including visit count N and node value V:

$$V(s) \leftarrow \frac{N(s) \cdot V(s) + \sum_{s_l \in \text{Child}(s)} R(s_l)}{N(s) + \text{CountChild}(S_{\text{candidate}}^*, s)}, \tag{4}$$

$$N(s) \leftarrow N(s) + \text{CountChild}(S_{\text{candidate}}^*, s), \tag{5}$$

where $\text{Child}(s)$ returns all the child nodes of s, and $\text{CountChild}(S_{\text{candidate}}^*, s)$ is a child node counting function that calculates the number of child nodes of s in $S_{\text{candidate}}^*$.

(d) Selection. Following Operations (a), (b) and (c), CoMCTS traverses the updated reasoning tree to select the next starting node. This selection is guided by the Upper Confidence Bound (UCB) value, which balances search exploration and exploitation. The UCB value of a node s is computed using the node reward value $V(s)$ and the visit cound $N(s)$. Among the candidate nodes $s \in S^*_{\text{candidate}}$, the one with the highest UCB value is chosen as the starting node $s_m^{k^*}$ for next search iteration:

$$s_m^{k^*} = \underset{s \in S^*_{\text{candidate}}}{\arg\max}\, V(s) + c \cdot \sqrt{\frac{\log N(\hat{s})}{1 + N(s)}} \tag{6}$$

where c stands for a constant which controls the level of exploration. $\hat{s}$ denotes the parent node of s.

CoMCTS. These four operations, *i.e.*, (a) Expansion, (b) Simulation and Error Positioning, (c) Backpropagation and (d) Selection, are repeated for a pre-defined number of iterations or until correct reasoning paths are found. This iterative process allows CoMCTS to construct a question-dependent reasoning tree S with the correct reasoning path Y, and ultimately form a multimodal learning-to-reason data triplet $\{Q, Y, S\}$. By applying our CoMCTS to a set of multimodal questions, we can construct a collection of multimodal learning-to-reason data triplets, which provide a tree of rich, explicit and well-defined reasoning nodes toward the final answer for each question and enable MLLMs to learn to reason step-by-step.

3.2. CoMCTS for reflective reasoning

In this subsection, we extend CoMCTS for reflective reasoning-path search. Based on the unified reasoning tree constructed by CoMCTS, *i.e.*, $\{Q, Y, S\}$, which provides both positive and negative reasoning nodes, we identify and integrate negative sibling nodes into effective reasoning paths to build the reflective reasoning path that includes a transition from a negative reasoning node to a positive one.

Identifying negative sibling node. Given the effective reasoning path Y, we identify the negative sibling reasoning node for $s \in Y$ using UCB:

$$s_{\text{neg}} = \underset{s_l \in \text{Sibling}(s)}{\arg\min}\ \text{UCB}(s_l) - \text{UCB}(s),\ \ \forall s \in Y, \tag{7}$$

where Sibling(s) returns all the sibling nodes of s, *i.e.*, the nodes on the same hierarchical level under the same parent node of s. UCB$(s) = V(s) + c \cdot \sqrt{\frac{\log N(\hat{s})}{1+N(s)}}$ as in Eq. 6.

Constructing reflective reasoning path. Based on Eq. 7, we randomly sample a reasoning node $s \in Y$ with its negative sibling node s_{neg}, and concatenate them with a reflection prompt to form a reflection trajectory, *i.e.*, $(s_{\text{neg}}, \text{prompt}_{\text{reflect}}, s)$. We then use a function Replace$(\cdot)$ that replaces $s \in Y$ with $(s_{\text{neg}}, \text{prompt}_{\text{reflect}}, s)$ to convert Y into the reflective reasoning path Y_{reflect}:

$$Y_{\text{reflect}} = \text{Replace}(Y, s, (s_{\text{neg}}, \text{prompt}_{\text{reflect}}, s)), \tag{8}$$

where prompt$_{\text{reflect}}$ denotes a reflection prompt, such as "The previous reasoning step is wrong and let's rethink it again." Then, we can integrate the reflective reasoning path Y_{reflect} into our data as a quadruplet $\{Q, Y, Y_{\text{reflect}}, S\} \in D$.

Algorithm 1 Training Mulberry with CoMCTS

```
Input: a set of policy models $\{\pi_1, \pi_2, ..., \pi_K\}$ initialized by different MLLMs; a set of multimodal questions $D_Q$
for i = 1 to MaxEpoch do
  Reasoning Tree Search using CoMCTS:
  for $Q \in D_Q$ do
    Collective Monte Carlo tree search:
    $\{Q, Y, S\} = \text{CoMCTS}(\{\pi_1, \pi_2, ..., \pi_K\}; Q)$
    if found an effective reasoning path then
      Search and find $Y_{\text{reflect}}$ from $S$
      Add $\{Q, Y, Y_{\text{reflect}}, S\}$ into $D$
      Remove $Q$ from $D_Q$
  Model Training with CoMCTS Reasoning Trees:
  for k = 1 to K do
    for $(Q, Y, Y_{\text{reflect}}, S) \in D$ do
      Supervised Fine-Tuning:
      Optimize $\pi_k$ via $\mathcal{L}_{\text{CoSFT}}(\pi_k)$ and $\mathcal{L}_{\text{CoSFT-Re}}(\pi_k)$
Output: Trained policy models $\{\pi_1, \pi_2, ..., \pi_K\}$
```

3.3. Training with Collective MCTS

Using CoMCTS, we search effective and reflective reasoning paths for a set of multimodal input questions, and construct Mulberry-260k, a multimodal learning-to-reason-and-reflect dataset with a tree of rich, explicit and well-defined reasoning nodes for each question, *i.e.*, a set of quadruplets $\{Q, Y, Y_{\text{reflect}}, S\} \in D$. To learn collective knowledge from Mulberry-260k, we perform collective SFT to train our model, Mulberry, a series of Multimodal LLMs with o1-like step-by-step Reasoning and Reflection capabilities.

Collective Supervised Fine-Tuning (CoSFT). Given $(Q, Y) \in \mathcal{D}$, we apply standard SFT objective to train our MLLM to learn from D constructed by CoMCTS:

$$\mathcal{L}_{\text{CoSFT}}(\pi_k) = \sum_{(Q,Y) \in \mathcal{D}} \log \pi_k(Y|Q), \tag{9}$$

where $Y = \{s\}$ denotes the effective reasoning path that includes a sequence of reasoning nodes collectively conjectured, searched and identified by a group of MLLMs.

CoSFT for reflective reasoning. Given a question and its reasoning tree $(Q, S) \in \mathcal{D}$ constructed by CoMCTS, we randomly sample a reflective reasoning path Y_{reflect} from S

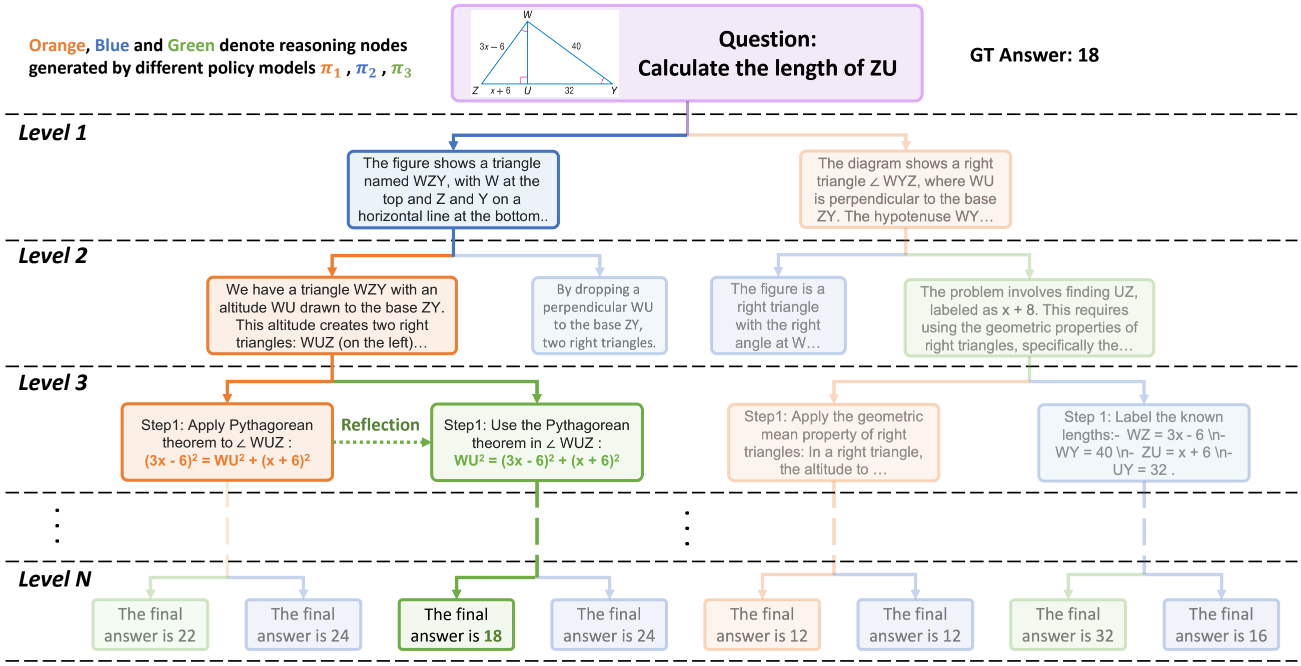

Orange, Blue and Green denote reasoning nodes generated by different policy models π_1, π_2, π_3
W
3x − 6
40
Z
x + 6
U
32
Y
Question: Calculate the length of ZU
GT Answer: 18
Level 1
The figure shows a triangle named WZY, with W at the top and Z and Y on a horizontal line at the bottom..
The diagram shows a right triangle ∠ WYZ, where WU is perpendicular to the base ZY. The hypotenuse WY…
Level 2
We have a triangle WZY with an altitude WU drawn to the base ZY. This altitude creates two right triangles: WUZ (on the left)…
By dropping a perpendicular WU to the base ZY, two right triangles.
The figure is a right triangle with the right angle at W…
The problem involves finding UZ, labeled as x + 8. This requires using the geometric properties of right triangles, specifically the…
Level 3
Step1: Apply Pythagorean theorem to ∠ WUZ : (3x - 6)² = WU² + (x + 6)²
Reflection
Step1: Use the Pythagorean theorem in ∠ WUZ : WU² = (3x - 6)² + (x + 6)²
Step1: Apply the geometric mean property of right triangles: In a right triangle, the altitude to …
Step 1: Label the known lengths:- WZ = 3x - 6 \n- WY = 40 \n- ZU = x + 6 \n- UY = 32 .
Level N
The final answer is 22
The final answer is 24
The final answer is 18
The final answer is 24
The final answer is 12
The final answer is 12
The final answer is 32
The final answer is 16

Figure 3: Qualitative illustration of reasoning tree searched by CoMCTS with rich, explicit, well-defined reasoning nodes.

as in Eqs.7-8, and conduct CoSFT for reflective reasoning:

$$\mathcal{L}_{\text{CoSFT-Re}}(\pi_k) = \sum_{(Q, Y_{\text{reflect}})\in\mathcal{D}} \log \pi_k(Y_{\text{reflect}}|Q), \qquad (10)$$

where $Y_{\text{reflect}} = \{s\}$ denotes the reflective reasoning path that includes an additional step-wise reflection trajectory.

The goal of $\mathcal{L}_{\text{CoSFT}}$ and $\mathcal{L}_{\text{CoSFT-Re}}$ is to maximize the log probability of effective and reflective reasoning path Y and Y_{reflect} over a tree of reasoning nodes S generated by CoMCTS. In addition, $\mathcal{L}_{\text{CoSFT-Re}}$ enables to leverage the negative information during CoMCTS search process by learning to calibrate negative reasoning nodes.

4. Experiment

In this section, we first introduce our CoMCTS-generated dataset, Mulberry-260K, including its sources, construction, and analysis in Section 4.1, and provide implementation details in Section 4.2. We then present the main results in Section 4.3, demonstrating the effectiveness of the searched data (*i.e.*, Mulberry-260K) and the trained models (*i.e.*, Mulberry). In Section 4.4, we perform comprehensive ablation studies on the impact of effective and reflective reasoning data and the contributions of collective knowledge sources. In final, Section 4.5 discuses the effectiveness and efficiency of tree search methods, explores different training strategies, and provides qualitative comparisons.

4.1. Dataset

The Sources of Raw Data. To construct a comprehensive and general-purpose tree-based reasoning dataset, we collect 260K raw multimodal input questions (*i.e.*, a text task instruction with an image as an input question) from a wide range of domains, covering General Multimodal Understanding, Mathematics, Figure Understanding, Realworld Understanding, Science, Medical Image Understanding, etc. The specific data sources are provided in the Appendix A.

Reasoning Data Construction. As detailed in Section 3 and Algorithm 1 and visually illustrated in Figures 2 and 3, we employ our CoMCTS to search effective and reflective reasoning paths for a set of raw multimodal input questions as collected from the mentioned "The Sources of Raw Data", ultimately constructing our dataset, Mulberry-260K. Note we only sample 15K data for reflective reasoning training to avoid overabundance of reflection data.

Reasoning Data Distribution. We analyze the CoMCTS-searched reasoning paths in Mulberry-260K by examining the distribution of reasoning steps, as shown in Figure 4. Specifically, Figure 4 shows that reasoning steps predominantly falls between 6 and 8, with an average of 7̃.5, for the entire Mulberry-260k. Meanwhile, for simple reasoning tasks, the chart-related subset of Mulberry-260k, reasoning steps typically ranges from 6 to 7, averaging x̃.x. For complex mathematical and logical reasoning tasks, such as the geometry-related subset of Mulberry-260k, the distribution shifts and largely falls between 7 and 10 steps, with an average of x̃.x. These observations highlight that the col-

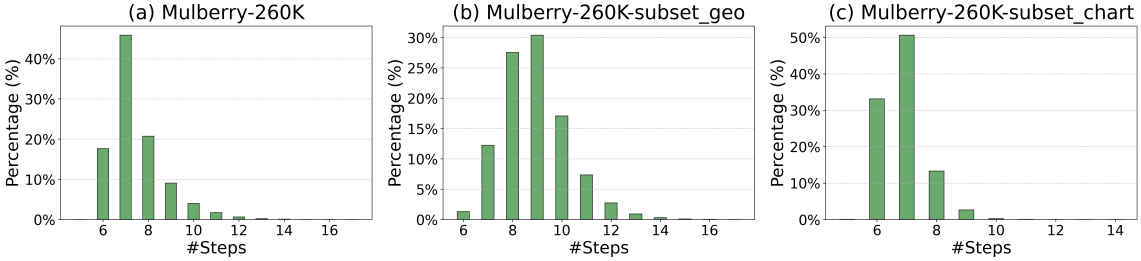


Figure 4: Distribution of reasoning steps in Mulberry-260K data.

lective tree search design in CoMCTS enables to generate effective reasoning trajectories with flexible numbers of reasoning steps, learning from which allows to train a powerful MLLM with great reasoning flexibility, *i.e.*, a model can "think less and faster" when handling simple questions (*i.e.*, allocate and generate fewer intermediate reasoning steps) and "think more and slower" when tackling complex tasks (*i.e.*, allocate and generate a greater number of intermediate reasoning steps).

4.2. Implementation Detail

In this paper, we implement the collective learning in CoMCTS by employing a group of four models, including GPT-4o, Qwen2-VL-7B, LLaMA-3.2-11B-Vision-Instruct, and Qwen2-VL-72B, to construct Mulberry-260K. In our CoMCTS, we set the maximum search iteration as 20. In each search iteration, we employ each model from the group to generate one subsequent candidate reasoning path to balance search exploration and exploitation. In Simulation and Error Positioning in CoMCTS, we simply set threshold t as 0. We adopt four popular MLLMs as baseline models, and conduct experiments on baselines Qwen2-VL-7B and LLaMA-3.2-11B-Vision-Instruct to examine the search effectiveness of our CoMCTS, and on baselines Qwen2-VL-2B and LLaVA-NeXT-8B to study the generalization of CoMCTS-searched data. The collective SFT experiments are conducted with a batch size of 128, a learning rate of 1e-5, and training over 2 epochs. For Qwen2-VL-7B, a smaller learning rate of 5e-6 is adopted to stabilize the training.

4.3. Main Results

To examine the effectiveness of the searched data (*i.e.*, Mulberry-260K) and the trained models (*i.e.*, Mulberry), we conduct extensive experiments with four powerful baseline models, and comprehensively benchmark our Mulberry with various state-of-the-arts, including general and reasoning-based MLLMs. The evaluation is performed on 8 widely used and challenging datasets (Huang & Zhang, 2024), covering the fields ranging from general and mathematical reasoning to hallucination and visual illusion, and multi-disciplinary understanding and reasoning, as shown in Table 1.

Comparison with baselines. We first conduct experiments on baselines Qwen2-VL-7B and LLaMA-3.2-11B-Vision-Instruct that are involved in collective learning of CoMCTS for joint reasoning-path conjecture, search and identification. We can observe that, trained with jointly-searched data (*i.e.*, Mulberry-260k), our Mulberry-7B and Mulberry-11B bring clear performance improvements against their baselines, *i.e.*, +4.2% over Qwen2-VL-7B and +7.5% over LLaMA-3.2-11B-Vision-Instruct averaged on 8 benchmarks, validating the search effectiveness of our CoMCTS. On the other hand, we examine the generalization of our Mulberry-260k by applying it to train other models that are not involved in collective tree search in CoMCTS, such as Qwen2-VL-2B and LLaVA-NeXT-8B. It can be observed that, trained with Mulberry-260k, our models (*i.e.*, Mulberry-2B and Mulberry-8B) enhance Qwen2-VL-2B and LLaVA-NeXT-8B with +5.4% and +11.0% gains averaged on 8 benchmarks, demonstrating the generalization of CoMCTS-searched data.

Comparison with reasoning-response models. We then benchmark our Mulberry with various state-of-the-art reasoning-response models. It shows that, using the same base model LLaVA-NeXT-8B (Li et al., 2024), our Mulberry outperforms LLaVA-Reasoner-8B and Insight-V-8B by +5.7% and +6.5% on mathematical benchmark MathVista, and by +3.0% and +1.0% on multi-disciplinary benchmark MMMU, respectively. Besides, Mulberry-11B surpasses LLaVA-COT-11B by +6.3% on reasoning-intensive benchmark MathVista under the same baseline LLaMA-3.2-11B-Vision-Instruct. The great superiority of Mulberry is largely attributed to our CoMCTS that conducts tree search and provides rich, explicit and well-defined reasoning nodes with flexible numbers of steps.

Comparison with state-of-the-arts. In final, we benchmark our Mulberry with popular state-of-the-arts included both open-source and closed-source ones. The results in Table 1 show that our Mulberry, trained on CoMCTS-searched data, outperforms most open-sourced MLLMs and achieves competitive results against closed-source ones, demonstrating outstanding abilities in step-by-step reasoning and reflection.

Method	MathVista	MMStar	MMMU	ChartQA	DynaMath	HallBench	MM-Math	MME$_{sum}$	AVG
Closed-Source Model									
GPT-4o (Hurst et al., 2024)	63.8	63.9	69.1	85.7	63.7	55.0	31.8	2329	64.5
Claude-3.5 Sonnet (Anthropic, 2024)	67.7	62.2	68.3	90.8	64.8	55.0	-	1920	-
Open-Source Model									
DeepSeek-VL-7B (Lu et al., 2024a)	36.1	37.1	35.4	59.1	21.5	-	-	-	-
Cambrain-1-8B (Tong et al., 2024)	49.0	-	42.7	73.3	-	-	-	-	-
MM-1.5-7B (Zhang et al., 2024b)	47.6	-	41.8	78.6	-	-	-	1861	-
Idefics3-LLaMA3-8B (Laurençon et al., 2024)	58.4	55.9	46.6	74.8	-	-	-	1937	-
InternVL2-8B (Chen et al., 2024)	58.3	**61.5**	51.8	83.3	39.7	-	-	2210	-
MiniCPM-Llama-V-2.5-8B (Yao et al., 2024c)	54.3	51.8	45.8	-	-	42.4	-	2025	-
MiniCPM-V-2.6-8B (Yao et al., 2024c)	60.6	57.5	49.8	-	-	48.1	-	2348	-
DeepSeek-VL2-MOE-4.5B (Wu et al., 2024)	62.8	61.3	51.1	86.0	-	-	-	2253	-
Reasoning Model									
LLaVA-CoT-11B (Xu et al., 2024)	54.8	57.6	-	-	-	47.8	-	-	-
LLaVA-Reasoner-8B (Zhang et al., 2024d)	50.6	54.0	40.0	83.0	-	-	-	-	-
Insight-V-8B (Dong et al., 2024)	49.8	57.4	42.0	77.4	-	-	-	2069	-
LLaVA-NeXT-8B (Li et al., 2024)	37.5	42.1	41.7	69.5	22.7	33.4	0.6	1957	39.7
Mulberry-LLaVA-8B	56.3	54.5	43.0	79.5	34.1	47.5	18.9	2021	$50.7^{11\uparrow}$
Llama-3.2-11B-V-Ins. (Dubey et al., 2024)	48.6	49.8	41.7	83.4	34.3	40.3	4.1	1787	45.8
Mulberry-Llama-11B	61.1	58.5	45.6	83.5	37.2	48.9	18.7	2035	$53.3^{7.5\uparrow}$
Qwen2-VL-2B (Wang et al., 2024b)	43.0	48.0	41.1	73.5	24.9	41.7	1.0	1872	42.5
Mulberry-2B	51.7	51.3	42.0	77.7	30.0	44.9	13.9	2013	$47.9^{5.4\uparrow}$
Qwen2-VL-7B (Wang et al., 2024b)	58.2	60.7	54.1	83.0	42.1	50.6	5.9	2327	54.7
Mulberry-7B	**63.1**	61.3	**55.0**	**83.9**	**45.1**	**54.1**	**23.7**	**2396**	$\mathbf{58.9}^{4.2\uparrow}$

Table 1: **Main Results.** To examine the effectiveness of the searched data (*i.e.*, Mulberry-260K) and the trained models (*i.e.*, Mulberry), we conduct extensive experiments with four powerful baseline models, and comprehensively benchmark our Mulberry with various state-of-the-arts, including general and reasoning-based MLLMs.

Direct Pred	CoMCTS				S.S.R.
GPT-4o	GPT-4o	Qwen2-VL-7B	LLama3.2-11B	Qwen2-VL-72B	
✓					58.2
	✓				63.8
	✓	✓			66.2
	✓	✓	✓		69.7
	✓	✓	✓	✓	80.2

Table 2: **Ablation Study on CoMCTS.** We study how each model in CoMCTS collective learning contribute to overall tree search performance in Search Success Rate (S.S.R.).

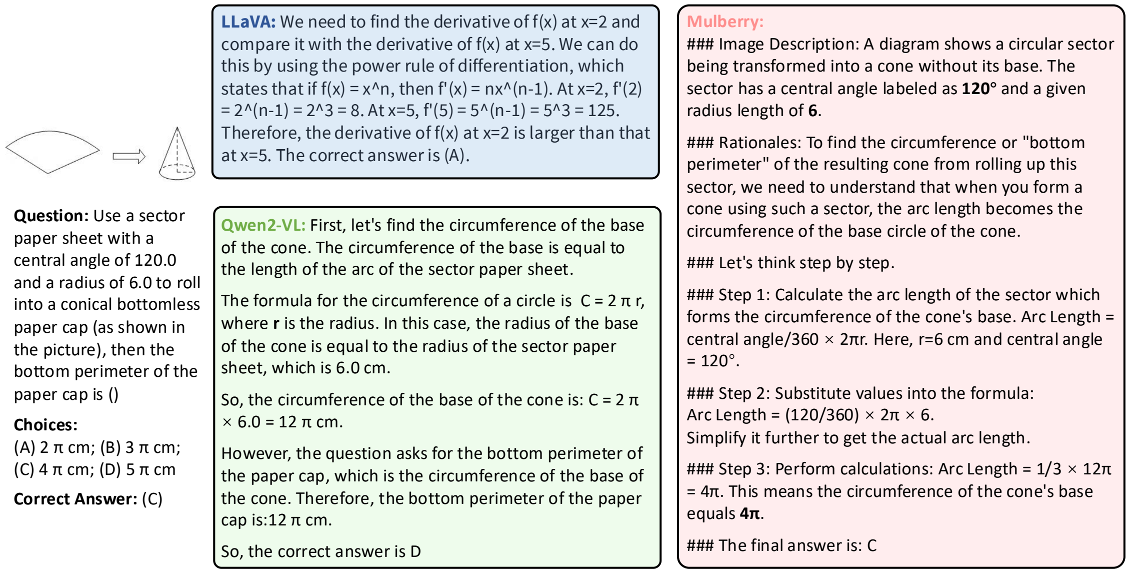
4.4. Ablation Study

Ablation Study on CoMCTS. We conduct ablation studies with the powerful GPT-4o as the baseline over 1K samples from Geo3K (Lu et al., 2021a) and GeoQA-Plus (Chen et al., 2021), as shown in Table 2. As the core of our proposed CoMCTS, we examine how each model in the collective learning group contribute to the overall tree search performance. Table 2 reports the Search Success Rates (S.S.R.). The baseline GPT-4o works not very well without tree search. It shows that CoMCTS with only GPT-4o improves the performance to 63.8%, largely becuase our tree search designs like expansion, simulation and error positioning can work even without using collective knowledge. Besides, progressively involving more models into CoMCTS consistently improves the search performance, even for including small models like Qwen2-VL-7B (*i.e.*, +2.4%), demonstrating the effectiveness of CoMCTS in capturing useful collective knowledge not only with large models but also from small models. In final, the inclusion of all four models in the proposed CoMCTS performs clearly the best, *i.e.*80.2%, validating the effectiveness of collective learning on reasoning tree search.

Benchmark	w/o Reflection Data	w/ Reflection Data
MathVista	50.9	51.7

Table 3: **Ablation Study on Mulberry.** As Mulberry is trained with effective and reflective reasoning data searched by CoMCTS, we study their respective contributions.

Ablation Study on Mulberry. We train Mulberry with effective and reflective reasoning data searched by CoMCTS, and study their respective contributions to overall reasoning performance. Table 3 presents the results on MathVista, which show that incorporating reflection data enhances the performance by 0.8%, demonstrating the complementarity of effective and reflective reasoning data searched by CoMCTS.



Question: Use a sector paper sheet with a central angle of 120.0 and a radius of 6.0 to roll into a conical bottomless paper cap (as shown in the picture), then the bottom perimeter of the paper cap is ()

Choices:
(A) 2 π cm; (B) 3 π cm; (C) 4 π cm; (D) 5 π cm

Correct Answer: (C)

LLaVA: We need to find the derivative of f(x) at x=2 and compare it with the derivative of f(x) at x=5. We can do this by using the power rule of differentiation, which states that if f(x) = x^n, then f'(x) = nx^(n-1). At x=2, f'(2) = 2^(n-1) = 2^3 = 8. At x=5, f'(5) = 5^(n-1) = 5^3 = 125. Therefore, the derivative of f(x) at x=2 is larger than that at x=5. The correct answer is (A).

Qwen2-VL: First, let's find the circumference of the base of the cone. The circumference of the base is equal to the length of the arc of the sector paper sheet.

The formula for the circumference of a circle is C = 2 π r, where **r** is the radius. In this case, the radius of the base of the cone is equal to the radius of the sector paper sheet, which is 6.0 cm.

So, the circumference of the base of the cone is: C = 2 π × 6.0 = 12 π cm.

However, the question asks for the bottom perimeter of the paper cap, which is the circumference of the base of the cone. Therefore, the bottom perimeter of the paper cap is:12 π cm.

So, the correct answer is D

Mulberry:
Image Description: A diagram shows a circular sector being transformed into a cone without its base. The sector has a central angle labeled as **120°** and a given radius length of **6**.

Rationales: To find the circumference or "bottom perimeter" of the resulting cone from rolling up this sector, we need to understand that when you form a cone using such a sector, the arc length becomes the circumference of the base circle of the cone.

Let's think step by step.

Step 1: Calculate the arc length of the sector which forms the circumference of the cone's base. Arc Length = central angle/360 × 2πr. Here, r=6 cm and central angle = 120°.

Step 2: Substitute values into the formula:
Arc Length = (120/360) × 2π × 6.
Simplify it further to get the actual arc length.

Step 3: Perform calculations: Arc Length = 1/3 × 12π = 4π. This means the circumference of the cone's base equals **4π**.

The final answer is: C

Figure 5: Qualitative Comparison. Our Mulberry, trained with CoMCTS-searched reasoning data, creates rich, explicit and well-defined reasoning steps with comprehensive understanding, ultimately arriving at the correct answer.

Methods	Search Success Rate ↑	Average Search Iteration ↓
GPT4o (direct)	58.2	-
MCTS	63.8	42.1
ReST-MCTS	65.6	36.3
Omega-MCTS	66.2	24.3
CoMCTS	80.2	12.7

Table 4: **Comparison with other tree search methods.** "GPT-4o (direct)" refers to the baseline without tree search. Our CoMCTS shows great superiority in search effectiveness and efficiency.

4.5. Discussion

Comparison with other tree search methods. We compare our CoMCTS with other tree search methods in search effectiveness and efficiency, including the baseline "GPT-4o direction prediction", "traditional MCTS (Coulom, 2006)", "ReST-MCTS (Zhang et al., 2024a)" that enhances MCTS by introducing partial search, and "Omega-MCTS (Luo et al., 2024)" that improves MCTS by designing binary search. Table 4 shows the results in search success rate and average search iteration that indicate search effectiveness and efficiency respectively. We can observe that existing tree search methods improve GPT-4o with limited gains. One main reason lies in that traditional MCTS methods generally work by self-bootstrapping and often get trapped in homogeneous low-quality nodes within the reasoning space of a single MLLM. On the other hand, our CoMCTS shows great superiority in search effectiveness and efficiency, largely thanks to the joint expansion mechanism in CoMCTS that allows reasoning-path search not only within the reasoning space of a given MLLM itself but also among those of others, benefiting from the synergy of multiple MLLMs while avoiding being trapped within the reasoning space of a single MLLM.

Qualitative comparison. We provide qualitative comparison of LLaVA-NeXT-8B (Li et al., 2024), Qwen2-VL-7B (Wang et al., 2024b), and Mulberry-7B in Figure 5. It shows that LLaVA-NeXT-8B and Qwen2-VL-7B generate relatively short predictions without thorough thinking, leading to incorrect answers. On the contrary, our Mulberry, trained with CoMCTS-searched reasoning data, creates rich, explicit and well-defined reasoning steps with comprehensive understanding, ultimately arriving at the correct answer.

5. Conclusion

This paper presents CoMCTS, a new learning-to-reason approach for MLLMs, which introduces the concept of collective learning into "tree search" for effective and efficient reasoning-path searching and learning. Based on the proposed CoMCTS, we search effective and reflective reasoning paths for a set of multimodal inputs, and construct Mulberry-260k, a multimodal learning-to-reason-and-reflect dataset with a tree of rich, explicit and well-defined reasoning nodes for each question. Using Mulberry-260k, we train our model, Mulberry, a series of Multimodal LLMs with o1-like step-by-step Reasoning and Reflection capabilities. Furthermore, we conduct extensive experiments, ablation studies and discussion, which demonstrate the superiority of our proposed methods on various benchmarks. We hope that CoMCTS along with Mulberry-260k and Mulberry will provides valuable resources and offer new insights for multimodal MCTS search and reasoning.

References

Anthropic. Claude 3.5 sonnet, 2024. URL `https://www.anthropic.com/news/claude-3-5-sonnet`.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Best, G., Cliff, O. M., Patten, T., Mettu, R. R., and Fitch, R. Dec-mcts: Decentralized planning for multi-robot active perception. *The International Journal of Robotics Research*, 38(2-3):316–337, 2019.

Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Podstawski, M., Gianinazzi, L., Gajda, J., Lehmann, T., Niewiadomski, H., Nyczyk, P., et al. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17682–17690, 2024.

Blum, A. and Mitchell, T. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pp. 92–100, 1998.

Chen, J., Tang, J., Qin, J., Liang, X., Liu, L., Xing, E. P., and Lin, L. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. *arXiv preprint arXiv:2105.14517*, 2021.

Chen, J., Li, T., Qin, J., Lu, P., Lin, L., Chen, C., and Liang, X. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. *arXiv preprint arXiv:2212.02746*, 2022.

Chen, Z., Wang, W., Tian, H., Ye, S., Gao, Z., Cui, E., Tong, W., Hu, K., Luo, J., Ma, Z., et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.

Coulom, R. Efficient selectivity and backup operators in monte-carlo tree search. In *International conference on computers and games*, pp. 72–83. Springer, 2006.

Cui, K., Huang, J., Luo, Z., Zhang, G., Zhan, F., and Lu, S. Genco: Generative co-training for generative adversarial networks with limited data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 499–507, 2022.

Dam, T., Chalvatzaki, G., Peters, J., and Pajarinen, J. Monte-carlo robot path planning. *IEEE Robotics and Automation Letters*, 7(4):11213–11220, 2022.

Dong, Y., Liu, Z., Sun, H.-L., Yang, J., Hu, W., Rao, Y., and Liu, Z. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. *arXiv preprint arXiv:2411.14432*, 2024.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Fawzi, A., Balog, M., Huang, A., Hubert, T., Romera-Paredes, B., Barekatain, M., Novikov, A., R Ruiz, F. J., Schrittwieser, J., Swirszcz, G., et al. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, 610(7930):47–53, 2022.

Foerster, J., Assael, I. A., De Freitas, N., and Whiteson, S. Learning to communicate with deep multi-agent reinforcement learning. *Advances in neural information processing systems*, 29, 2016.

Gao, J., Pi, R., Zhang, J., Ye, J., Zhong, W., Wang, Y., Hong, L., Han, J., Xu, H., Li, Z., et al. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*, 2023.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Gurari, D., Li, Q., Stangl, A. J., Guo, A., Lin, C., Grauman, K., Luo, J., and Bigham, J. P. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3608–3617, 2018.

Huang, J. and Zhang, J. A survey on evaluation of multimodal large language models. *arXiv preprint arXiv:2408.15769*, 2024.

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., and Girshick, R. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

Kafle, K., Price, B., Cohen, S., and Kanan, C. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5648–5656, 2018.

Kahou, S. E., Michalski, V., Atkinson, A., Kádár, Á., Trischler, A., and Bengio, Y. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*, 2017.

Kazemi, M., Alvari, H., Anand, A., Wu, J., Chen, X., and Soricut, R. Geomverse: A systematic evaluation of large models for geometric reasoning. *arXiv preprint arXiv:2312.12241*, 2023.

Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., and Farhadi, A. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pp. 235–251. Springer, 2016.

Kembhavi, A., Seo, M., Schwenk, D., Choi, J., Farhadi, A., and Hajishirzi, H. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pp. 4999–5007, 2017.

Lample, G., Lacroix, T., Lachaux, M.-A., Rodriguez, A., Hayat, A., Lavril, T., Ebner, G., and Martinet, X. Hypertree proof search for neural theorem proving. *Advances in neural information processing systems*, 35:26337–26349, 2022.

Lau, J. J., Gayen, S., Ben Abacha, A., and Demner-Fushman, D. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.

Laurençon, H., Marafioti, A., Sanh, V., and Tronchon, L. Building and better understanding vision-language models: insights and future directions. In *Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models*, 2024.

Li, B., Zhang, K., Zhang, H., Guo, D., Zhang, R., Li, F., Zhang, Y., Liu, Z., and Li, C. Llava-next: Stronger llms supercharge multimodal capabilities in the wild, May 2024. URL `https://llava-vl.github.io/blog/2024-05-10-llava-next-stronger-llms/`.

Li, Z., Wang, X., Stengel-Eskin, E., Kortylewski, A., Ma, W., Van Durme, B., and Yuille, A. L. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14963–14973, 2023.

Lindström, A. D. and Abraham, S. S. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*, 2022.

Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., and Wang, L. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2023.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

Lu, H., Liu, W., Zhang, B., Wang, B., Dong, K., Liu, B., Sun, J., Ren, T., Li, Z., Yang, H., et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024a.

Lu, P., Gong, R., Jiang, S., Qiu, L., Huang, S., Liang, X., and Zhu, S.-C. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *The 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021a.

Lu, P., Qiu, L., Chen, J., Xia, T., Zhao, Y., Zhang, W., Yu, Z., Liang, X., and Zhu, S.-C. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*, 2021b.

Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., and Kalyan, A. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022a.

Lu, P., Qiu, L., Chang, K.-W., Wu, Y. N., Zhu, S.-C., Rajpurohit, T., Clark, P., and Kalyan, A. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*, 2022b.

Lu, S., Li, Y., Chen, Q.-G., Xu, Z., Luo, W., Zhang, K., and Ye, H.-J. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*, 2024b.

Luan, B., Feng, H., Chen, H., Wang, Y., Zhou, W., and Li, H. Textcot: Zoom in for enhanced multimodal text-rich image understanding. *arXiv preprint arXiv:2404.09797*, 2024.

Luo, L., Liu, Y., Liu, R., Phatale, S., Lara, H., Li, Y., Shu, L., Zhu, Y., Meng, L., Sun, J., et al. Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv:2406.06592*, 2024.

Masry, A., Long, D. X., Tan, J. Q., Joty, S., and Hoque, E. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.

Mathew, M., Karatzas, D., and Jawahar, C. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2200–2209, 2021.

Mathew, M., Bagal, V., Tito, R., Karatzas, D., Valveny, E., and Jawahar, C. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1697–1706, 2022.

Methani, N., Ganguly, P., Khapra, M. M., and Kumar, P. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1527–1536, 2020.

Mitra, C., Huang, B., Darrell, T., and Herzig, R. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14420–14431, 2024.

OpenAI. Introducing openai o1, 2024. URL `https://openai.com/o1/`.

Pitanov, Y., Skrynnik, A., Andreychuk, A., Yakovlev, K., and Panov, A. Monte-carlo tree search for multi-agent pathfinding: Preliminary results. In *International Conference on Hybrid Artificial Intelligence Systems*, pp. 649–660. Springer, 2023.

Qiao, S., Shen, W., Zhang, Z., Wang, B., and Yuille, A. Deep co-training for semi-supervised image recognition. In *Proceedings of the european conference on computer vision (eccv)*, pp. 135–152, 2018.

Saito, K., Watanabe, K., Ushiku, Y., and Harada, T. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3723–3732, 2018.

Schwenk, D., Khandelwal, A., Clark, C., Marino, K., and Mottaghi, R. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pp. 146–162. Springer, 2022.

Seo, M., Hajishirzi, H., Farhadi, A., Etzioni, O., and Malcolm, C. Solving geometry problems: Combining text and diagram interpretation. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 1466–1476, 2015.

Shi, W., Hu, Z., Bin, Y., Liu, J., Yang, Y., Ng, S.-K., Bing, L., and Lee, R. K.-W. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models. *arXiv preprint arXiv:2406.17294*, 2024.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., and Rohrbach, M. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.

Sun, S. and Jin, F. Robust co-training. *International Journal of Pattern Recognition and Artificial Intelligence*, 25(07): 1113–1126, 2011.

Tong, S., Brown, E., Wu, P., Woo, S., Middepogu, M., Akula, S. C., Yang, J., Yang, S., Iyer, A., Pan, X., et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024.

Vagadia, H., Chopra, M., Barnawal, A., Banerjee, T., Tuli, S., Chakraborty, S., and Paul, R. Phyplan: Compositional and adaptive physical task reasoning with physics-informed skill networks for robot manipulators. *arXiv preprint arXiv:2402.15767*, 2024.

Wang, K., Pan, J., Shi, W., Lu, Z., Zhan, M., and Li, H. Measuring multimodal mathematical reasoning with math-vision dataset. *arXiv preprint arXiv:2402.14804*, 2024a.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

Xie, Y., Goyal, A., Zheng, W., Kan, M.-Y., Lillicrap, T. P., Kawaguchi, K., and Shieh, M. Monte carlo tree search boosts reasoning via iterative preference learning. *arXiv preprint arXiv:2405.00451*, 2024.

Xu, G., Jin, P., Hao, L., Song, Y., Sun, L., and Yuan, L. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024.

Yang, F. An integrated framework integrating monte carlo tree search and supervised learning for train timetabling problem. *arXiv preprint arXiv:2311.00971*, 2023.

Yao, H., Wu, W., Yang, T., Song, Y., Zhang, M., Feng, H., Sun, Y., Li, Z., Ouyang, W., and Wang, J. Dense connector for mllms. *arXiv preprint arXiv:2405.13800*, 2024a.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., and Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 2024b.

Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Li, H., Zhao, W., He, Z., et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024c.

Ye, W., Liu, S., Kurutach, T., Abbeel, P., and Gao, Y. Mastering atari games with limited data. *Advances in neural information processing systems*, 34:25476–25488, 2021.

Yu, S., Krishnapuram, B., Rosales, R., and Rao, R. B. Bayesian co-training. *The Journal of Machine Learning Research*, 12:2649–2680, 2011.

Yu, T., Yao, Y., Zhang, H., He, T., Han, Y., Cui, G., Hu, J., Liu, Z., Zheng, H.-T., Sun, M., et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13807–13816, 2024.

Yue, X., Zheng, T., Ni, Y., Wang, Y., Zhang, K., Tong, S., Sun, Y., Yu, B., Zhang, G., Sun, H., et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024.

Zelikman, E., Wu, Y., Mu, J., and Goodman, N. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

Zhang, D., Zhoubian, S., Hu, Z., Yue, Y., Dong, Y., and Tang, J. Rest-mcts*: Llm self-training via process reward guided tree search. *arXiv preprint arXiv:2406.03816*, 2024a.

Zhang, H., Gao, M., Gan, Z., Dufter, P., Wenzel, N., Huang, F., Shah, D., Du, X., Zhang, B., Li, Y., et al. Mm1. 5: Methods, analysis & insights from multimodal llm fine-tuning. *arXiv preprint arXiv:2409.20566*, 2024b.

Zhang, J., Huang, J., Jin, S., and Lu, S. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024c.

Zhang, R., Zhang, B., Li, Y., Zhang, H., Sun, Z., Gan, Z., Yang, Y., Pang, R., and Yang, Y. Improve vision language model chain-of-thought reasoning. *arXiv preprint arXiv:2410.16198*, 2024d.

Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., and Xie, W. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.

Zhao, Y., Li, Y., Li, C., and Zhang, R. Multihiertt: Numerical reasoning over multi hierarchical tabular and textual data. *arXiv preprint arXiv:2206.01347*, 2022.

A. The Sources of Raw Data

To construct a comprehensive and general-purpose tree-based reasoning dataset, we collect 260K raw multimodal input questions spanning varouis domain, including

- 55K Mathematical Data: From GLLaVA (Gao et al., 2023), GEOS (Seo et al., 2015), UniGeo (Chen et al., 2022), GeoQA Plus (Chen et al., 2021), Geo3K (Lu et al., 2021a), MathVision (Wang et al., 2024a), GeoMverse (Kazemi et al., 2023), and MathV360K (Shi et al., 2024).
- 116K Figure Understanding data: From DVQA (Kafle et al., 2018), DocVQA (Mathew et al., 2021), FigureQA (Kahou et al., 2017), PlotQA (Methani et al., 2020), ChartQA (Masry et al., 2022), InfoVQA (Mathew et al., 2022), MultiHiertt (Zhao et al., 2022), and LRV-Chart (Liu et al., 2023).
- 41K Math Word Problem Data: From IconQA (Lu et al., 2021b), TabMWP (Lu et al., 2022b), CLEVR (Johnson et al., 2017), CLEVR-Math (Lindström & Abraham, 2022), and Super-CLEVR (Li et al., 2023).
- 2K Mdeical Data: From VQA-RAD (Lau et al., 2018), and PMC-VQA (Zhang et al., 2023).
- 17K Sience Data: From TQA (Kembhavi et al., 2017), AI2D (Kembhavi et al., 2016), and ScienceQA (Lu et al., 2022a).
- 24K Nature World QA Data: From VQA-AS (Antol et al., 2015), A-OKVQA (Schwenk et al., 2022), TextVQA (Singh et al., 2019), Vizwiz (Gurari et al., 2018), and VQA2.0 (Goyal et al., 2017).