arXiv:2504.07934v1 [cs.CV] 10 Apr 2025

SoTA with Less: MCTS-Guided Sample Selection for Data-Efficient Visual Reasoning Self-Improvement

Xiyao Wang[1,2†], **Zhengyuan Yang**[2], **Chao Feng**[3], **Hongjin Lu**[1]
Linjie Li[2], **Chung-Ching Lin**[2], **Kevin Lin**[2], **Furong Huang**[1,‡], **Lijuan Wang**[2,‡]
[1]University of Maryland, College Park [2]Microsoft [3]University of Michigan
[†]xywang@umd.edu [‡]Equal advise

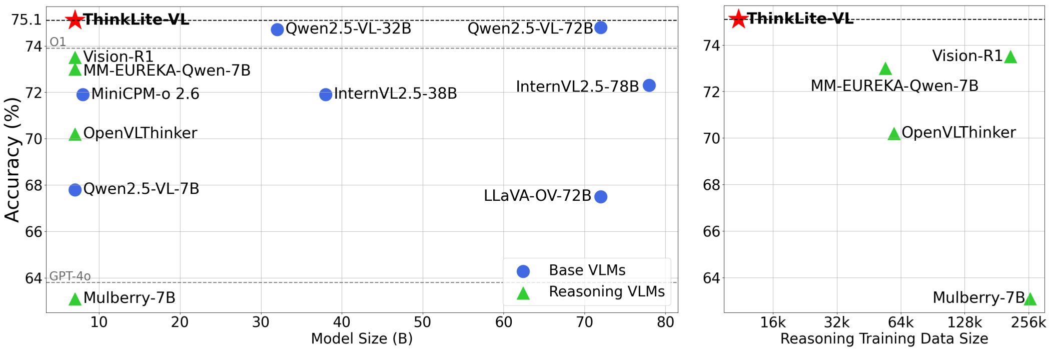

Figure 1: Recent "Reasoning VLMs" studies finetune "Base VLMs" with extra reasoning training data to improve visual reasoning. This paper presents a data-efficient self-improving method for better training reasoning VLMs. **(Left)** Comparison of VLMs with different parameter sizes on **MathVista**. Our model ThinkLite-VL-7B achieves the state-of-the-art (SoTA) accuracy of **75.1**, surpassing Qwen2.5-VL-72B-Instruct, GPT-4o, O1, and other 7B-level reasoning VLMs. **(Right)** Comparison of the reasoning training data size used by 7B-level reasoning models. Our model achieves SoTA performance using only 11k data, and without any additional knowledge distillation.

Abstract

In this paper, we present an effective method to enhance visual reasoning with significantly fewer training samples, relying purely on self-improvement with no knowledge distillation. Our key insight is that the difficulty of training data during reinforcement fine-tuning (RFT) is critical. Appropriately challenging samples can substantially boost reasoning capabilities even when the dataset is small. Despite being intuitive, the main challenge remains in accurately quantifying sample difficulty to enable effective data filtering. To this end, we propose a novel way of repurposing Monte Carlo Tree Search (MCTS) to achieve that. Starting from our curated 70k open-source training samples, we introduce an MCTS-based selection method that quantifies sample difficulty based on the number of iterations required by the VLMs to solve each problem. This explicit step-by-step reasoning in MCTS enforces the model to think longer and better identifies samples that are genuinely challenging. We filter and retain 11k samples to perform RFT on Qwen2.5-VL-7B-Instruct, resulting in our final model, ThinkLite-VL. Evaluation results on eight benchmarks show that ThinkLite-VL improves the average performance of Qwen2.5-VL-7B-Instruct by 7%, using only 11k training samples with no knowledge distillation. This significantly outperforms all existing 7B-level reasoning VLMs, and our fairly comparable baselines that use classic selection methods such as accuracy-based filtering. Notably, on MathVista, ThinkLite-VL-7B achieves the

SoTA accuracy of 75.1, surpassing Qwen2.5-VL-72B, GPT-4o, and O1. Our code, data, and model are available at https://github.com/si0wang/ThinkLite-VL.

1 Introduction

Leveraging long chain-of-thought reasoning with effective reflection during inference, large language models (LLMs) [24, 34] are capable of solving complex reasoning tasks such as math and coding. Recent studies [16] show that large-scale reinforcement fine-tuning (RFT) is a critical factor in enhancing model's reasoning performance. Notably, substantial reasoning performance improvements can be achieved solely through reinforcement fine-tuning in the post-training stage, even without the standard supervised fine-tuning (SFT) in post-training.

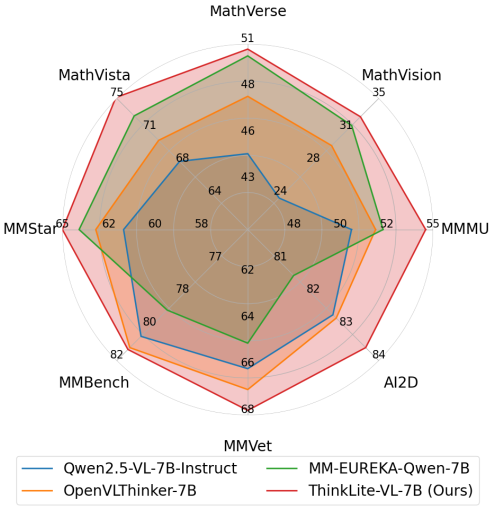

Figure 2: Performance comparison on 8 visual benchmarks. Our model significantly outperforms Qwen2.5-VL-7b-Instruct and other 7b-level reasoning models.

Despite the notable successes in enhancing LLM reasoning with large-scale RFT, similar progress in vision-language models (VLMs) remains limited, likely due to the mismatch between the text-focused pre-training and the multimodal nature of VLM post-training tasks. Recent attempts [22, 12, 53, 81] have employed knowledge-distillation via supervised fine-tuning before the RFT stage, to encourage more visual reasoning related responses being generated. Despite the performance improvement, the knowledge distillation stage is cumbersome, and inherently prevents base VLMs from self-improving themselves in achieving stronger intelligence.

In this paper, we demonstrate that high-quality, appropriately challenging training data is key factor to enable and self-improve visual reasoning ability. When visual reasoning training data aligns properly with the base VLM's skill level, large-scale RFT alone can significantly enhance VLM's reasoning ability without relying on knowledge distillation for format fine-tuning or base capability enhancement. Based on this insight, We introduce a data-efficient training pipeline that results in ThinkLite-VL, a reasoning VLM that achieves SoTA visual reasoning performance with less training samples.

The critical factor to ThinkLite-VL's success is the strategic selection of training samples with suitable difficulty. To achieve this, we repurpose Monte Carlo tree search (MCTS), a classic inference-time search algorithm, to accurately quantify the sample difficulty. Specifically, MCTS's explicit tree search enforces sufficient thinking compute in deciding the question difficulty, and provide a tight correlation between the question difficulty and the number of MCTS iterations needed to solve it. Our training pipeline begins with collecting 70k open-source samples from three key domains: mathematical reasoning, natural image understanding, and chart comprehension. We then implement MCTS-guided sample selection by applying the VLM itself to perform iterative reasoning on each of the 70k samples, using the number of iterations required to reach the correct solution as a difficulty measure. This rigorous filtering process results in a set of 11k challenging and high-quality samples tailored specifically for our base model. We then directly perform RFT with these selected samples, avoiding any additional supervised fine-tuning steps.

Using the Qwen2.5-VL-7B-Instruct model as our base, we develop our final model, ThinkLite-VL-7B. We evaluate ThinkLite-VL-7B on eight widely used VLM benchmarks. As shown in Figure 2, after RFT with the filtered 11k high-quality data, ThinkLite-VL-7B significantly improves the average performance of Qwen2.5-VL-7B-Instruct from 59.69 to 63.89. It also surpasses the fairly comparable baseline that RFT with the same amount of unfiltered data, from 60.89 to 63.89. Furthermore, compared with the most recent 7B-level reasoning VLMs, ThinkLite-VL-7B consistently demonstrates substantial performance advantages. Notably, on the MathVista benchmark, ThinkLite-VL-7B