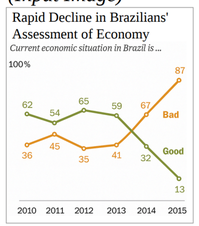
Empowering Small VLMs to Think with Dynamic Memorization and Exploration

Jiazhen Liu · Yuchuan Deng · Long Chen

The Hong Kong University of Science and Technology

Introduction

(Input Image)



Large-scale VLM

Small-scale VLM

After CoT SFT training

Which year shows a greater divergence of opinions about Brazil's economy, 2010 or 2012? Output the **thinking process** and then give the final answer **in <answer> </answer> tag**.

We **locate** *2010* and *2012*. Then, **extract** 2010 has *62* and *36*, 2012 has *65* and *35*. **Comparing** $62-36=26<65-35=30$, the year with greater divergence is **2012**. <answer>**2012**</answer>

To answer the question, we first **locate** the relevant years in the chart. Then, we **extract** the values for each year. **Compare** the differences, the year with greater divergence is **2010**. <answer>**2010**</answer>

LVLMs: vision preserved → **Grounded thinking traces**

SVLMs: vision compromised → **Pseudo thinking traces (fail)**

During RLVR training

Which year has the most divergent opinions about Brazil's economy? Output the **thinking process** and then give the final answer **in <answer> </answer> tag**.

To answer the question, we examine the chart and compare the value differences for each year. In 2015, the values are 87 and 13, showing the most significant divergence. **<answer>2015</answer>**

2015 has the greatest divergence.

SVLMs poorly follow instructions → **Advantage collapsing (fail)**

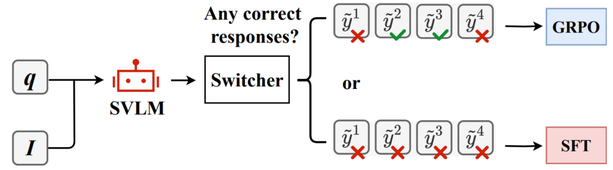
SFT or RL struggles to train reasoning abilities in small VLMs.

- Long texts overwhelm the model's capacity.
- Small models have poor instruction following.
- SFT handles memorization, while RL explores diverse modes.

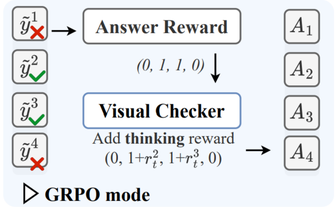
SFT and RL can complement each other.

Method I

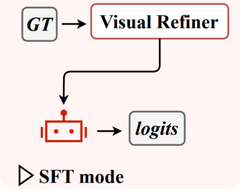
Dynamically switch between SFT and RL modes



▷ Overall pipeline



▷ GRPO mode



▷ SFT mode

***Dy**namic **M**emorize–**E**xplore*

Method II

Visual supervision

Given visual facts I_c

From the figure, the values represented by the three bars can be directly observed. Taking their mean yields an overall average of approximately **8.3**, which provides a concise summary of the distribution reflected

Visual Checker: 0

For GRPO

Visual Refiner

To answer the question, ..., **step1** is to extract the values for **Latvia, Portugal, and Romania: 3.0, 8.9, and 13.0,** respectively. **Step 2** is to calculate the average of these values: **(3.0 + 8.9 + 13.0) / 3 = 8.3. Therefore,** the average is **8.3**.

For SFT

Visual Checker: 1.0

Preliminary Study

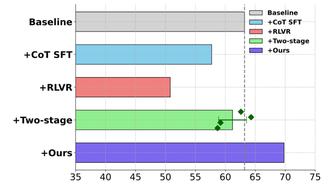
SFT and RL can provide equivalent gradient forms.

$$\nabla_\theta \mathcal{L}_{\mathrm{SFT}}(\theta) = -\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\nabla_\theta \log p_\theta(y \mid x)\right].$$

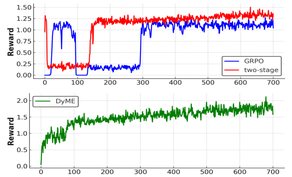
$$\nabla_\theta \mathcal{L}_{\mathrm{GRPO}}(\theta) = -\mathbb{E}_{\substack{x\sim\mathcal{D},\\ \tilde{y}\sim p_{\mathrm{old}}(\cdot|x)}}\left[r_\theta(x,\tilde{y})\, A(\tilde{y})\, \nabla_\theta \log p_\theta(\tilde{y} \mid x)\right].$$

Experiments

Better Performance



Stable Training Process



Lower data construction costs

Method	Ext. Model	Time	Acc.
GRPO (Baseline)	Qwen2.5-14B†	14.8s	60.8
Pure DyME	Qwen2.5-14B†	14.0s	64.9
Pure DyME	GPT-4o†	19.1s	**68.5**
Full DyME	Qwen2.5-7B	21.2s	66.8
Full DyME	Qwen2.5-14B	23.4s	67.5

† Used for offline data construction only.