Unified Reinforcement and Imitation Learning for Vision Language Models

Byung-Kwan Lee[1,2*] Ryo Hachiuma[1] Yong Man Ro[2]

Yu-Chiang Frank Wang[1,3] Yueh-Hua Wu[1]

[*]Work Done during Internship

1 NVIDIA

2 KAIST

3 National Taiwan University

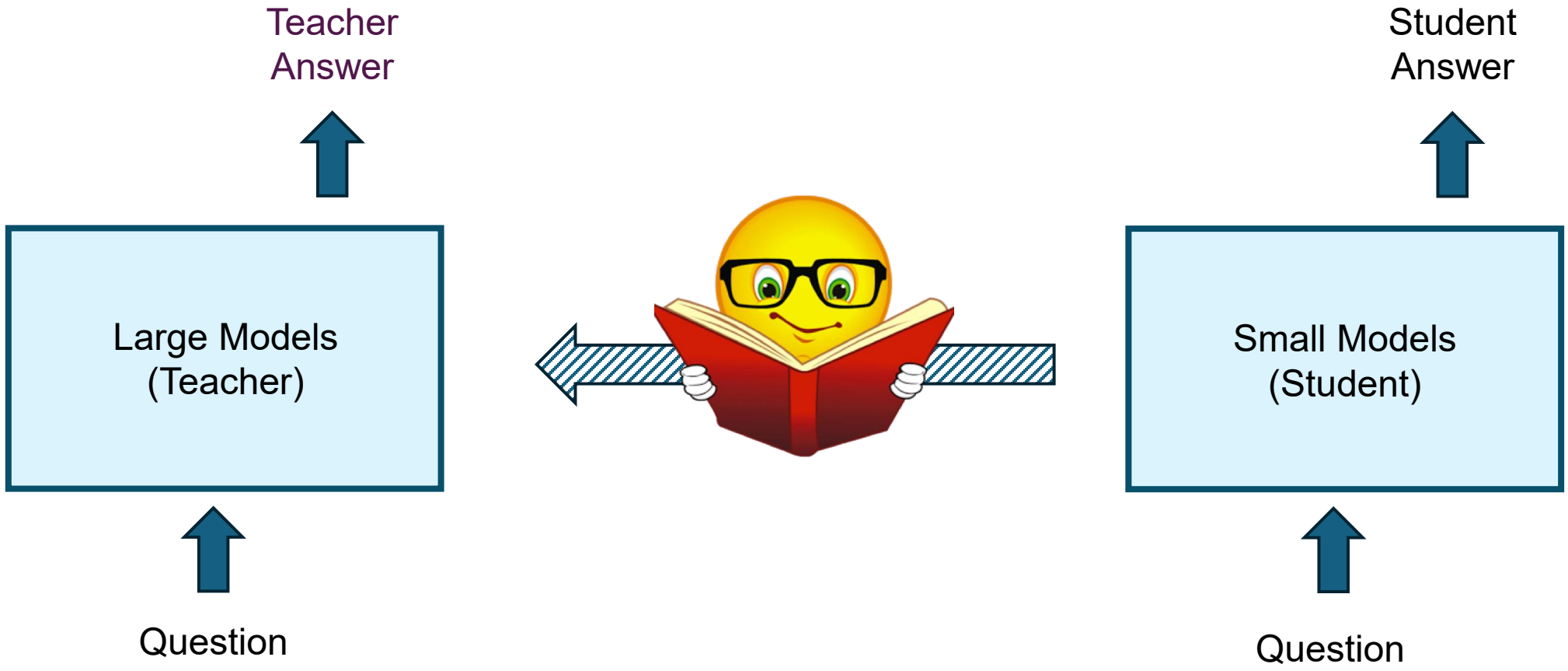
Motivation

- Limitations of Existing VLM Scaling and Architectural Changes
- Demand for Inference Efficiency and Lightweight Models

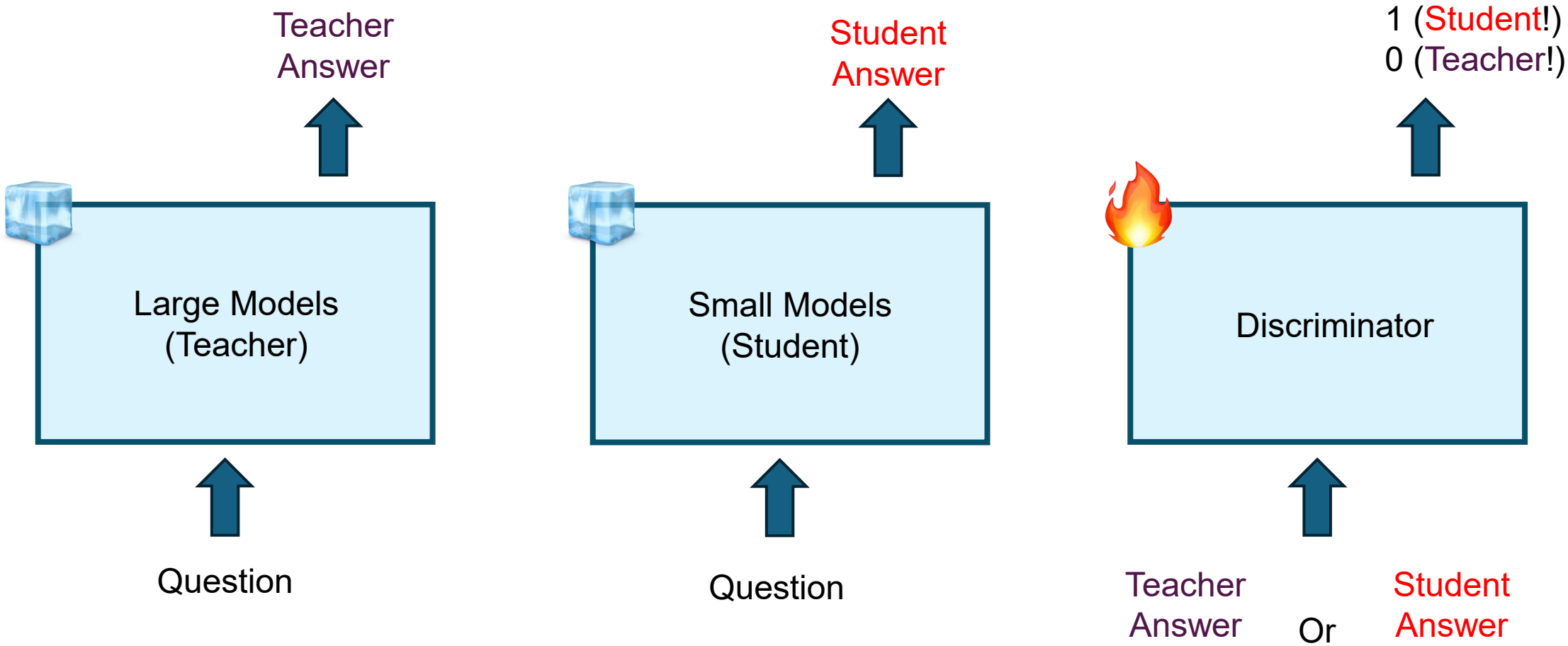
How can remove think-answer but get high-performance?

Think-Answer process is actually too much slow.

Motivation

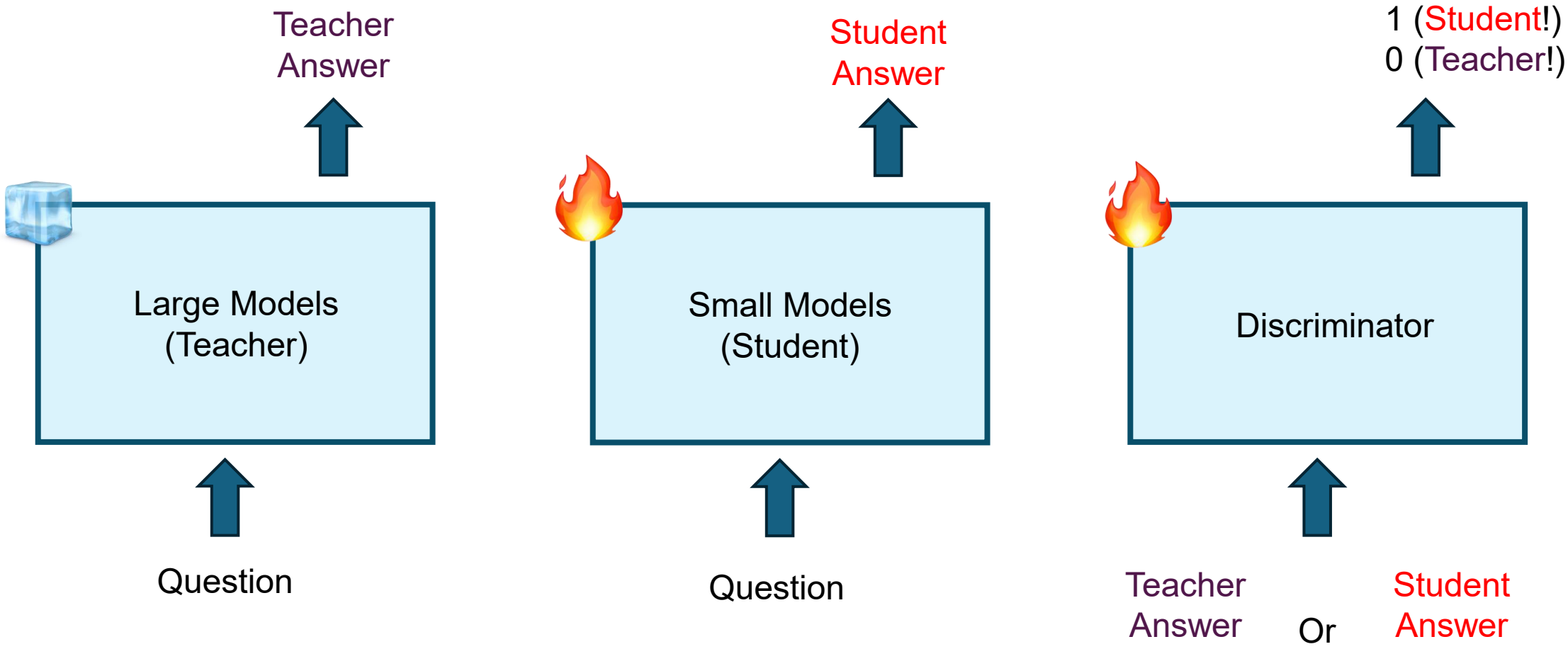

Teacher
Answer
Large Models
(Teacher)
Question
Student
Answer
Small Models
(Student)
Question

Proposed Method (RIL): Training Step1



$$\max_{\phi} \mathcal{L}(\phi) = \frac{1}{N} \sum_{i=1}^{N} \left\{ \log D_{\phi}(q, o_i^{(\mathrm{s})}) + \log(1 - D_{\phi}(q, o_i^{(\mathrm{t})})) \right\}$$

Proposed Method (RIL): Training Step2



Reward: $R_{\text{teacher similarity}} + R_{\text{answer}}$

$$\mathbb{1}(D_\phi(q, o_i) < 0.5) \qquad \text{LLM-as-a-Judge}(q, a, o_i)$$

Experimental Results

Algorithm 2 RIL for VLMs

Require: Pre-trained discriminator D_ϕ and Pre-trained student VLMs $\pi_{\theta_{\text{init}}}$

Require: Saved collection for the generated text responses $\{o^{(1)}\}$ from teacher VLMs

Set reference model $\pi_{\text{ref}} \leftarrow \pi_{\theta_{\text{init}}}$

Set the training model $\pi_\theta \leftarrow \pi_{\theta_{\text{init}}}$

for sample a batch $\mathcal{B}$ in Dataset **do**

  Copy and freeze model $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$

  Sample G outputs $\{o_i^{(\text{s})}\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot|q)$ for each question $q \in \mathcal{B}$

  Extract G outputs $\{o_i^{(\text{t})}\}_{i=1}^{G}$ in the saved collection for each question $q \in \mathcal{B}$

  for Discriminator Updating iteration $= 1, 2, \cdots, \mu$ **do**

    Update D_ϕ by using Equation 1

  end for

  Compute Rewards and Advantages for all $2G$ outputs by Discriminator and LLM-as-a-Judge

  for Student VLMs Updating iteration $= 1, 2, \cdots, \mu$ **do**

    Update π_θ by using Equation 2

  end for

end for

Experimental Results

VLMs	AI2D	ChartQA	MathVista	MMB	MMB^{CN}	MM-Vet	MM-Vet-v2	MMMU	MMMU-Pro	MMStar	BLINK	SEED	SEED2+	RWQA
Qwen2.5-VL-7B	83.9	87.3	67.8	83.5	83.4	71.8	63.7	55.0	38.3	63.9	56.4	77.0	70.4	68.5
w. RIL (Qwen2.5-VL-72B)	86.7	95.4	74.5	**86.8**	**87.2**	77.3	66.1	61.8	48.2	**71.1**	68.5	80.7	**73.0**	74.2
w. RIL (InternVL3-78B)	**86.8**	95.5	74.6	86.7	87.1	75.8	66.0	60.9	47.1	**71.1**	68.1	**80.8**	72.7	**75.4**
w. RIL (Both)	86.1	**95.6**	**79.7**	86.3	86.5	**80.4**	**71.1**	**65.7**	**48.5**	**71.1**	**70.0**	80.5	72.8	72.8