
Vision-EKIPL: 注入外部知识的视觉推理策略学习

Chaoyang Wang
Central South University

Zeyu Zhang
The Australian National University

Haiyun Jiang*
Shanghai Jiao Tong University

Abstract

视觉推理对于理解复杂的多模态数据和提升通用人工智能至关重要。现有的方法通过强化学习 (RL) 微调 (例如, GRPO) 来增强多模态大语言模型 (MLLMs) 的推理能力。然而, 目前的强化学习方法仅从政策模型本身采样动作组, 这限制了模型推理能力的上限并导致低效的训练。为了解决这些限制, 本文提出了一种新颖的 RL 框架, 称为 Vision-EKIPL。此框架的核心在于在 RL 训练过程中引入由外部辅助模型生成的高质量动作, 以指导政策模型的优化。从外部模型引入知识进行的政策学习显著扩展了模型的探索空间, 有效提高了推理边界, 并大幅加速了训练的收敛速度和效率。实验结果表明, 与最先进的技术 (SOTA) 相比, 本文提出的 Vision-EKIPL 在 Reason-RFT-CoT 基准上实现了高达% 的性能提升。这表明 Vision-EKIPL 能够克服传统 RL 方法的局限性, 显著增强 MLLMs 的视觉推理性能, 并为该领域的研究提供了一种新的有效范式。

1 介绍

视觉推理是一种核心的认知能力, 涉及基于视觉信息的解释、推断和逻辑思维, 它已经成为人工智能领域一个关键且极具挑战性的研究前沿。该能力是许多复杂 AI 应用的基础, 包括从图像识别和场景理解到自主机器人导航和自动驾驶, 强调了其日益增长的战略重要性。

为了有效增强机器的视觉推理能力, 研究界已经探索了多种技术方法。目前主流的研究范式大致可以分为三类: (1) 神经-符号方法 [? ? ? ? ?], 这种方法旨在将深度神经网络卓越的模式识别能力与符号系统固有的逻辑严密性和可解释性结合起来。(2) 多模态大语言模型 (MLLMs) [? ?] 的监督微调 (SFT), 该方法依赖于大规模标注的数据集进行端到端训练, 以直接优化特定视觉推理任务上的模型性能。(3) 基于强化学习 (RL) 的方法 [? ?], 以技术 (例如, 组相对策略优化 (GRPO) [? ?]) 为例。这类方法利用 RL 的奖励机制来引导和激发预训练基础模型中的潜在推理能力, 展示了良好的泛化能力, 特别是在处理涉及数学逻辑推导或代码理解的复杂视觉认知任务时, 因此受到越来越多的关注。

然而, 尽管基于 RL 的方法在一系列视觉推理任务中取得了显著的成功, 最近的研究 [?] 揭示了一个值得注意的现象: 经过 RL 训练的模型生成的推理路径在很大程度上似乎并没有超越预训练基础模型的固有能力范围。这表明, RL 训练带来的性能提升可能主要来源于其作为偏好优化器的作用。具体而言, RL 通过奖励信号强化模型的采样策略, 使其倾向于选择那些历史上产生高回报的已知推理路径, 从而更有效地生成正确答案。然而, 这一机制固有的一个潜在瓶颈是: 它可能过于偏向于利用已知的成功路径, 从而抑制了对新颖或更复杂推理路径的探索。一个潜在的后果是, 经过 RL 微调的模型与其拥有巨大潜力的基础模型相比, 其推理边界不仅可能未能扩展, 反而可能会收缩。此外, 现有的 RL 方法通常存在收敛速度慢和训练效率低的问题。

*Corresponding author.

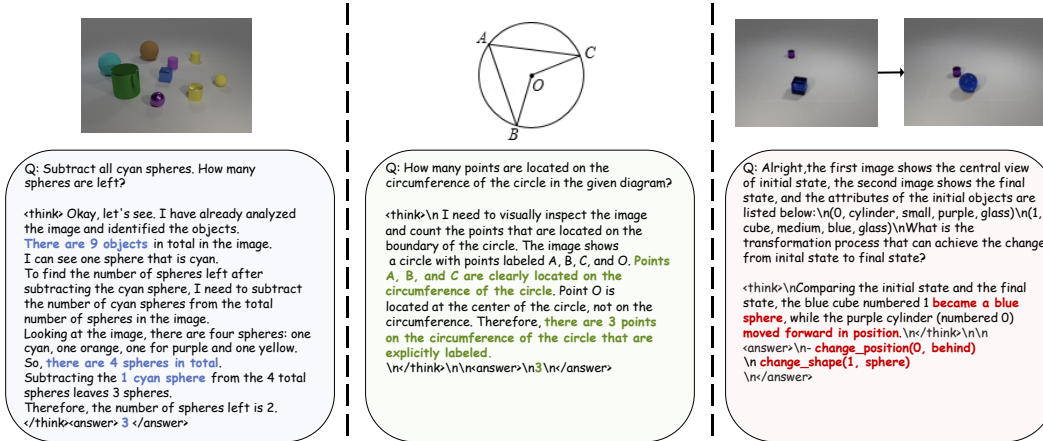


Figure 1: Vision-EKIPL 在三个视觉推理任务中的输出示例。

为克服当前强化学习方法在推理边界扩展和训练效率上的双重限制，本文引入了一种名为 Vision-EKIPL 的新型强化学习框架。其核心创新在于显著拓宽策略学习过程中的信息来源。在每个输入状态下，该框架不仅基于当前策略模型采样动作，还将多个外部辅助模型的动作纳入候选集合。随后，这些候选动作根据它们获得的奖励信号进行排序，并选择奖励最高的 k 个动作以形成高质量动作组。该组随后被用来指导策略模型的优化。通过这种机制，Vision-EKIPL 通过整合不同“专家”（即辅助模型）提供的潜在解决方案，有效拓宽了策略模型的探索空间，协助发现可能被单一策略模型忽视的有效推理路径。

从本质上讲，我们的方法旨在通过在优化过程中主动引入和整合外部知识（表现为辅助模型的高质量动作），显著提升策略模型的推理前沿，从而使其能够探索和学习更丰富、更复杂的推理策略，有效缓解与标准强化学习微调相关的潜在推理能力消退。同时，通过直接利用外部模型的高质量动作来引导策略模型的优化，我们的方法也大幅提高了训练过程的收敛速度和整体效率。从更广泛的视角来看，这一方法可以被概念化为监督微调（数据蒸馏）和强化学习的一种有前途的混合范式。当策略模型的初始推理能力相对有限时，模型展现出较大的倾向选择从外部模型中获取的知识进行监督学习；反之，随着策略模型自身能力的逐步提升，它逐渐倾向于自主探索更深层次的推理策略。与最先进的（SOTA）模型相比，Vision-EKIPL 在 Reason-RFT-CoT 基准上实现了高达 5% 的性能提升。Vision-EKIPL 的输出示例展示在图 1 中。虽然这项研究主要验证了该框架在视觉推理任务上的有效性，但所提出的框架具有广泛的普适性，理论上可以灵活应用于更广泛的人工智能领域，包括各种语言任务、视觉任务和多模态任务。

我们的主要贡献总结如下：

- 我们提出了 Vision-EKIPL，一种创新的强化学习框架，通过整合由外部模型生成的高质量动作，显著增强了多模态大语言模型（MLLMs）的视觉推理能力，从而辅助策略模型的优化。
- 我们证明，在策略优化过程中结合来自外部模型的高质量动作，能够有效扩展策略模型的动作探索空间，从而扩展其推理边界。
- 通过广泛的实验评估，我们验证了 Vision-EKIPL 框架的有效性，为推进视觉推理研究提供了宝贵的见解，并引入了一种可能有助于促进多模态学习研究的新范式。

2 相关工作

视觉推理 视觉推理是人工智能领域的一个关键研究方向，旨在使机器能够理解和解释视觉信息，并在此基础上执行复杂的认知任务，如逻辑推理、因果分析和问题解决。该技术具有广泛的应用前景，包括视觉计数 [??]、几何问题解决 [????]、视觉变换推理 [?]、科学研究 [??] 和机器人任务规划 [??]。早期的视觉推理工作依赖于程式化生成 [??] 或神经符号方法 [????]。近年来，由于多模态大语言模型（MLLMs）的快速发展，该领域取得了突破性进展。例如，LLaVA-CoT [?] 采用多阶段的链式思维（CoT）[?] 监督微调（SFT）策略，而 Insight-V [?] 结合了 SFT 和强化学习（RL）。DeepSeek-R1-Zero [?] 引入了一

种基于规则的 RL 方法，显著增强了推理能力。在 DeepSeek-R1 [?] 的基础上，我们提出了一种新的 RL 方法，大幅提高了模型的推理性能。

强化学习 强化学习 (RL) 已通过迭代、反馈驱动的优化在增强大型语言模型 (LLM) 的推理能力方面展示了显著的效果 [?????]。值得注意的方法包括来自人类反馈的强化学习 (RLHF) [?] 和来自 AI 反馈的强化学习 (RLAIF) [?]，两者都利用人类或 AI 生成的反馈来优化模型行为。在视觉-语言任务领域，RL 已成功用于对齐模型预测与人类偏好，并减少幻觉的发生 [????]。最近，诸如 DeepSeek-R1-Zero [?] 这样的进展引入了组相对策略优化 (GRPO) [?]，该技术利用基于规则的奖励来增强推理能力，而无需监督微调。GRPO 已被进一步改编用于专门的应用，例如 Visual-RFT [?] 将其用于视觉定位，Med-R1 [?] 将其应用于医学推理任务。Vision-R1 [?] 和 Reason-RFT [?] 采用两阶段训练范式——CoT 监督微调，随后是基于 GRPO 的强化微调，以增强多模态大型语言模型 (MLLM) 的推理性能。与众不同的是，我们的 Vision-EKIPL 首次利用外部模型生成的高质量动作来指导策略模型优化，从而注入新的推理知识并推动模型推理能力的边界。

视觉推理可以正式定义为通过共同分析视觉和文本信息来推断结论或答案的任务。给定一个视觉输入（例如，图像或视频）及其相关的文本描述或问题，目标是生成一个相应的答案。这个过程可以形式化为：其中表示视觉输入，以高度、宽度和通道数为特征。文本输入通常由自然语言查询或描述组成，而输出表示推断的答案，可以用自然语言或结构化格式表达。通过这种映射，设计视觉推理模型以有效整合和解释多模态信息以执行复杂推理任务。

群体相对策略优化 (GRPO)。 GRPO [?] 提出了一种新颖的强化学习框架，在诸如 DeepSeek R1 [?] 等模型中表现出了强劲的性能。GRPO 的基本目标是通过根据组内采样动作的相对表现迭代地优化其策略，从而增强模型的推理能力。

这个过程从当前策略 π_θ 开始，针对给定状态 s 。一个包含 N 个动作的组 $\{o_1, o_2, \dots, o_N\}$ ，从策略的输出分布 $\pi_\theta(o|s)$ 中被采样。该组中每一个采样的动作 o_i 随后都通过一个奖励函数 $R(o_i)$ 来进行评估，该函数量化动作的期望程度或有效性。

GRPO 的一个关键元素是为每个行动计算优势分数。动作 o_i 的优势 A_i 定义为：

$$A_i = \frac{R(o_i) - \text{mean}(\{R(o_1), R(o_2), \dots, R(o_N)\})}{\text{std}(\{R(o_1), R(o_2), \dots, R(o_N)\})} \quad (1)$$

。具有正优势的动作被认为优于群体平均水平，而具有负优势的动作则被视为劣于平均水平。在计算出优势 A_i 之后，GRPO 评估每个动作在更新后的策略 $\pi_{\theta_{\text{new}}}$ 和之前的策略 $\pi_{\theta_{\text{old}}}$ 下的概率比，记为 ratio_i 。

$$\text{ratio}_i = \pi_{\theta_{\text{new}}}(o_i | s) / \pi_{\theta_{\text{old}}}(o_i | s) \quad (2)$$

然后更新策略模型参数 θ ，以增加选择显示出正优势的动作的概率，并减少选择负优势动作的概率。这种更新通常使用基于梯度的优化方法进行。为了减少过度的策略更新并增强训练的稳定性，GRPO 将 ratio_i 限制在 $[1 - \delta, 1 + \delta]$ 的区间内。此外，为了鼓励所学习的策略保持在参考分布 π_{ref} 的附近，一个 Kullback-Leibler (KL) 散度惩罚（由系数 β 加权）被集成到优化目标中。最后，GRPO 的优化目标可以表述如下：

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{s \sim Q, \{o_i\}_{i=1}^N \sim \pi_{\text{old}}} \left[\frac{1}{N} \sum_{i=1}^N \min \left(\text{ratio}_i A_i, \right. \right. \\ \left. \left. \text{clip}(\text{ratio}_i, 1 - \epsilon, 1 + \epsilon) A_i \right) - \beta \mathbb{D}_{\text{KL}}[\pi_\theta || \pi_{\text{ref}}] \right] \quad (3)$$

其中， Q 表示候选问题集， $\mathbb{D}_{\text{KL}}$ 表示 KL 正则化。 π_{ref} 通常是一个冻结的预训练 MLLM。简而言之，GRPO 旨在最大化期望优势，常常将此 KL 散度作为惩罚项。

2.1 外部知识注入的策略学习

Vision-EKIPL 的整体框架如图 2 所示。Vision-EKIPL 是一个强化学习框架，旨在增强多模态大语言模型 (MLLMs) 的视觉推理能力。Vision-EKIPL 的关键洞察在于利用由外部辅助模型生成的高质量动作来指导策略模型的优化，从而注入新的推理知识，并进一步扩展模型的推理能力。

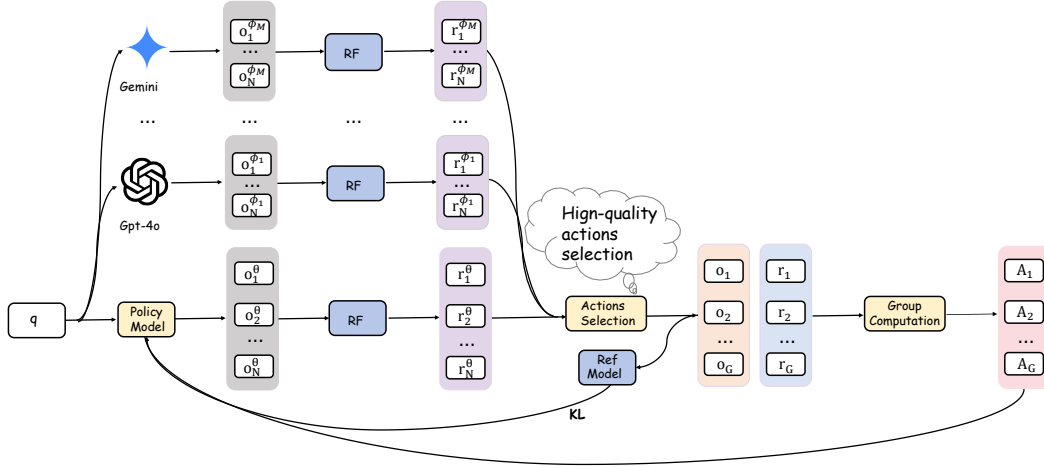


Figure 2: 所提出的 Vision-EKIPL 框架概述。Vision-EKIPL 从外部模型和策略模型的动作集合中，根据奖励函数（RF）的评估采样高质量的动作组，然后通过 GRPO 算法使用高质量动作组优化策略模型。

从策略模型之外采样动作组。我们引入总共 M 个辅助模型来支持学习过程。给定一个输入状态 $s = (x, q)$ ，其中 x 表示输入图像的视觉编码， q 表示问题的文本编码，GRPO 首先从当前策略 π_θ 中采样一组动作 $\{o_1^\theta, \dots, o_N^\theta\}$ 。此外，对于每个辅助模型 π^{ϕ_j} ，它会采样一个相应的动作组 $\{o_1^{\phi_j}, \dots, o_N^{\phi_j}\}$ 。采样过程如下：

$$o_i^\theta \sim \pi_\theta(o | x, q), \quad \text{for } i = 1, 2, \dots, N \quad (4)$$

所有这些采样的动作然后被结合成一个完整的动作组 O ：

$$O = \{o_i^\theta | i = 1, \dots, N\} \cup \bigcup_{j=1}^M \{o_i^{\phi_j} | i = 1, \dots, N\} \quad (5)$$

奖励计算。每个采样的动作 o_i 根据可验证的标准被赋予奖励 $R(o_i)$ 。在视觉推理任务的背景下，奖励函数 $R(o_i)$ [?] 整合了两个组成部分：格式奖励 $R_{\text{format}}(o_i)$ 和准确性奖励 $R_{\text{acc}}(o_i)$ 。格式奖励强制遵循结构化的响应格式，而准确性奖励评估输出的正确性，从而在结构化推理和事实准确性之间取得平衡。奖励函数形式定义为：

$$R(o_i) = R_{\text{format}}(o_i) + R_{\text{acc}}(o_i). \quad (6)$$

奖励计算遵循以下列出的标准：

- 如果回答提供了正确的最终答案，模型将获得 +1 的准确奖励。否则，模型将获得 0 奖励。
- 如果响应将其推理放在 `<think></think>` 标签中，并将最终答案放在 `<answer></answer>` 标签中，则模型会收到 +1 的格式奖励。否则，模型将收到 0 奖励。

动作选择与优势计算。行为组 O 中的动作根据它们的奖励值按降序排序，并选择靠前的 G 个动作以形成高质量动作组 $T: \{o_1, o_2, \dots, o_G\}$ ，以及它们对应的奖励组 $R: \{r_1, r_2, \dots, r_G\}$ 。

将采样奖励组 R 中的奖励值进行归一化，以计算相对优势 $\{A_1, A_2, \dots, A_G\}$ ，计算方法如方程 1 所示。在计算出动作组 T 的相对优势后，根据方程 3 更新策略模型。

在这篇论文中，我们使用 Reason-RFT-CoT 数据集 [?] 来评估我们的方法。实验按以下三个任务类别组织：(1) 视觉计数该任务通过整合语言、视觉和数学技能来解决 3D 基础积木场景中的算术问题，以评估多模态推理。(2) 结构感知这一视觉推理任务要求模型解读各种数学几何、医学成像、图表布局和建筑设计中的结构信息。(3) 空间变换这一空间-视觉推理任务通过分析从多个视角（例如，中心、左、右）呈现的 3D 场景的初始和最终视觉状态，评估模型推断单步或多步骤变换动作的能力。每个任务都包含域内测试集和域外测试集。具体信息可以在 [?] 中找到。

Algorithm 1 视觉-EKIPL: 外部知识注入的视觉推理策略学习

Input initial policy model $\pi_{\theta_{\text{init}}}$; reward function R ; input states $\mathcal{S}$; external models $\{\pi^{\phi_j}\}_{j=1}^M$

```
1: policy model  $\pi_{\theta} \leftarrow \pi_{\theta_{\text{init}}}$ 
2: for iteration = 1, ..., I do
3:   reference model  $\pi_{\text{ref}} \leftarrow \pi_{\theta}$ 
4:   for step = 1, ..., B do
5:     Sample a batch  $\mathcal{S}_b$  from  $\mathcal{S}$ 
6:     Update the old policy model  $\pi_{\theta_{\text{old}}} \leftarrow \pi_{\theta}$ 
7:     Sample  $N$  actions  $\{o_i^{\theta}\}_{i=1}^N \sim \pi_{\theta_{\text{old}}}(o | s)$  for each input state  $s \in \mathcal{S}_b$ 
8:     for j = 1, ..., M do
9:       Sample  $N$  actions  $\{o_i^{\phi_j}\}_{i=1}^N \sim \pi_{\theta}^{\phi_j}(o | s)$  for each input state  $s \in \mathcal{S}_b$ 
10:    Action set  $O \leftarrow \{o_i^{\theta} | i = 1, \dots, N\} \cup \bigcup_{j=1}^M \{o_i^{\phi_j} | i = 1, \dots, N\}$ 
11:    Compute rewards for each sampled action within  $O$  by running  $R$ 
12:    Select the top-G reward actions from  $O$  to form high-quality action set  $T$ 
13:    Compute  $A_i$  for the action  $o_i$  within  $T$  through group relative advantage estimation (Eqn. 1)
14:    for GRPO iteration = 1, ...,  $\mu$  do
15:      Update the policy model  $\pi_{\theta}$  by maximizing the GRPO objective (Eqn. 3)
```

Output π_{θ}

评估指标主要的评估指标是准确率 (Acc) [?]。对于数值答案, 通过验证与标准答案的数学等价性来确定正确性。在多项选择题的情况下, 字符串匹配是针对正确选项进行的。对于函数类型的序列, 采用逐步多层次评估方法来评估与正确解序列的一致性。

实施细节在我们的实验中, 我们使用 Qwen2-VL-2B 和 Qwen2-VL-7B [?] 作为策略模型。对于外部模型, 我们选择了 GPT-4o [?] 和 Gemini-1.5-Pro [?]。我们的实现基于开源框架 Open-R1 [?] 和 vLLM [?] 以确保结果的可重复性和系统的可拓展性。所有实验均在配备 8 个 A100 GPU 的服务器上进行。

用于比较的基线为评估不同训练策略的性能和泛化能力, 并遵循 [?] 中的设置, 本文比较的方法如下: (1) 基于 SFT 的方法——ANS-SFT, 针对答案生成进行微调, 以及 CoT-SFT, 使用链式思维 (CoT) 推理的监督学习。(2) 基于 RL 的方法——Reason-RFT-Zero, 直接对基础模型应用 RL 训练; Reason-RFT, 先使用部分链式思维 (CoT) 数据进行监督学习, 再进行 RL 训练; 以及 Vision-EKIPL, 将外部辅助模型整合到 RL 训练中。

为了进行全面的评估, 我们采用 Qwen2-VL-Instruct [?] 作为基础模型, 评估其 2B 和 7B 变体以研究模型规模的影响。此外, 最先进的开源模型 [?????] 和专有模型 [??] 被作为基准纳入, 以评估各种训练范式的性能。

为了评估 Vision-EKIPL 在视觉推理任务中相对于不同训练范式和基线模型的域内 (ID) 性能, 我们对三个任务的 2B/7B 模型进行了广泛的训练和评估。表 ?? 中展示的结果表明如下几点: (1) 在视觉计数任务中, 基于 RL 的方法在 2B 和 7B 模型中通常优于所有开源和专有基线模型, 以及基于 SFT 的方法, 其中 Vision-EKIPL 在 7B 模型中表现最佳; (2) 在结构感知任务中, 基于 RL 的方法在 7B 模型中超过了基于 SFT 的方法, 而 ANS-SFT 在 2B 模型中表现最佳。CoT-SFT 显示出有限的改善, 这可能是因为强制的推理监督阻碍了认知增强。此外, 7B 模型中的 Vision-EKIPL 超过了所有专有模型和大多数开源模型, 除了 InternVL-2.5-8B [?] 和 Llava-OneVision-7B [?]; (3) 在空间转换任务中, Vision-EKIPL 取得了最高性能, 超过了所有基线模型。与 Reason-RFT 不同, Vision-EKIPL 不需要监督微调就能激活其推理能力, 但仍优于 Reason-RFT。这表明结合外部模型的高质量动作可以有效提高模型的推理能力。

为了验证 Vision-EKIPL 在域外 (OOD) 性能相对于不同训练范式和视觉推理任务的基准模型, 我们对三项任务的 2B/7B 模型进行了全面的实验。结果如表 ?? 所示, 揭示了以下几点: (1) 在视觉计数任务中, 基于强化学习 (RL) 的视觉计数方法在 2B 和 7B 模型中均表现出比基于 SFT 方法更优越的泛化能力。具体来说, Vision-EKIPL 在 2B 模型中比 ANS-SFT 高出 13%, 在 7B 模型中高出 19%, 并且超越了所有开源和专有的基准。值得注意的是, 与传统的 RL 方法 (例如 Reason-RFT) 相比, Vision-EKIPL 显著扩展了模型的推理边界, 使得模型能够探索并发现 Reason-RFT 未能找到的复杂问题的正确推理路径。(2) 结构感知的 RL 方法始终优于基于 SFT 的方法, Vision-EKIPL 在 2B 和 7B 模型中均取得了最佳成绩 (比 Reason-RFT 在 2B 模型中高出 8%), 而 Reason-RFT 在 2B 模型中取得了可比较的性能。基于 SFT 的方法显示出有限的影响, 尤其是在 7B 模型中; (3) 在空间变换任务中, RL 方法在 2B 和 7B 模型中均优于 SFT 方法, 同时明显超越了所有基准模型。Vision-EKIPL (2B)

表现出卓越的 OOD 泛化能力，超越了 GPT-4o [?] 34 % 以及 Gemini-1.5-Pro [?] 47 %。总体而言，Vision-EKIPL 超越了所有开源和专有的基准模型以及其他训练方法，展示了在视觉推理泛化能力上的卓越表现。为了展示 Vision-EKIPL 在训练过程中的数据效率，我们在

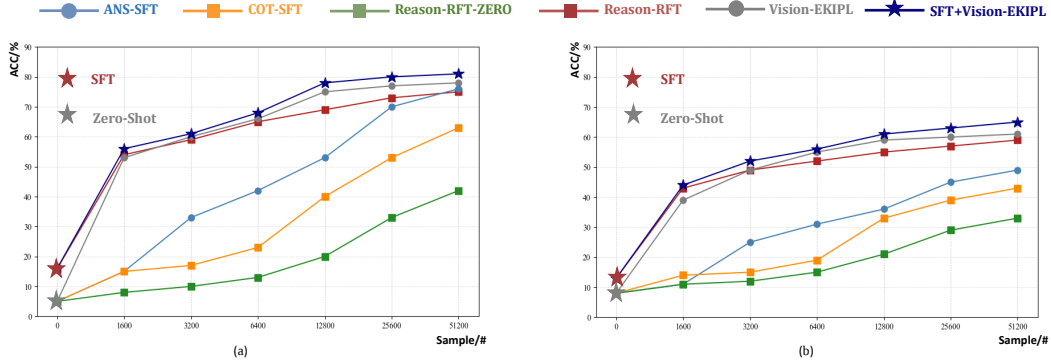


Figure 3: 不同方法在空间变换任务中的训练过程结果。(a) 2B 模型在 ID 任务上的评估结果，(b) 2B 模型在 OOD 任务上的评估结果。

TRANSE 数据集上训练所有方法，并记录中间结果和验证结果，如图 3 所示。Vision-EKIPL 在域内 (ID) 和域外 (OOD) 任务中均展示了出色的数据效率。主要结果包括：(1) 在 ID 任务中，Vision-EKIPL 仅使用 25 % 的训练数据 (12,800 个样本) 就超越了 Reason-RFT 的性能。此外，当 Vision-EKIPL 在 RL 训练之前使用 CoT 数据集进行 SFT 训练，并遵循 [?] 中的设置时，仅用 12 % 的训练数据就达到了 Reason-RFT 93 % 的性能。(2) 在 OOD 任务中，Vision-EKIPL 仅使用 12 % 的数据就达到了 Reason-RFT 的性能，展现了强大的泛化能力。

2.2 行动起源分析

如图 ?? 所示，我们跟踪了在 TRANSE 数据集上训练 2B 模型期间用于参数更新的动作组中，来自外部模型和策略模型的动作比例的动态演变。我们可以观察到，随着训练迭代次数的增加，来自外部模型的动作比例逐渐减少，而来自策略模型自身的动作比例在用于更新策略模型参数的动作组中逐步增加。

这一现象可以归因于训练的初期阶段：与辅助模型（或外部模型）相比，策略模型的推理能力相对较弱。因此，从策略模型中采样的动作通常比辅助模型提供的动作获得的奖励更低。为了有效地引导策略模型的优化方向，我们优先选择由辅助模型生成的动作。

然而，随着模型的进一步优化和更深入的训练，策略模型的推理能力显著提高，从其生成的动作中获得的奖励也相应增加。在这一阶段，为了充分利用策略模型自身的学习成果并加速其收敛，我们越来越多地选择由策略模型生成的动作来推动模型的优化。如图 ?? 所示，我们对不同 k 值下 Vision-EKIPL 和 Reason-RFT 的 Pass@ K 分数进行了比较分析。Pass@ K 指标衡量了模型在 K 个独立样本中至少生成一个正确答案的概率，这有效地反映了通过更广泛的探索可以实现的模型推理能力的上限。通过提供大量机会让模型尝试解决问题（即设置较大的 K 值），我们可以更全面地评估基础模型以及采用不同 RL 方法训练的模型所能达到的推理界限。这种评估方法为确定 RL 训练是否真正增强了模型的基本推理能力提供了一个关键而严谨的视角。

在较小的 k 值时，Vision-EKIPL 和 Reason-RFT 都优于未经过任何 RL 训练的基础模型。然而，一个值得注意的现象是，随着 k 值的不断增加，Reason-RFT 的 Pass@ K 得分开始低于基础模型。这一结果表明基础模型本身，通过足够多样化的采样探索，具备生成正确答案的能力，即使对于传统上认为需要 RL 训练的问题。这进一步表明，传统 RL 训练可能无法有效增强，甚至在一定程度上限制了模型本身的潜在推理范围，这些范围通过广泛的探索可以被发掘出来。

鲜明对比的是，在较大的 k 值下，Vision-EKIPL 的 Pass@ K 得分显著超过了基准模型和 Reason-RFT。这一观察结果强烈地强调了 Vision-EKIPL 的重要潜力：通过其设计所支持的大规模采样，Vision-EKIPL 能够揭示出基准模型和 Reason-RFT 难以覆盖的更广泛的正确推理路径，从而大幅扩展模型的推理边界。Vision-EKIPL 突破基准模型推理边界的能力主要在于其核心机制：巧妙地将外部模型生成的高质量动作作为蒸馏目标融入到策略模型的更新过程中。这一创新机制不仅显著提升了模型的采样效率，更重要的是，有效地注入了新的推理知识，并提供了更为丰富的决策视角。在探索复杂问题的空间时，这使得 Vision-EKIPL

能够展现出卓越的效能，最终使其成功地识别并利用基准模型难以企及的正确推理路径，从而提升模型的推理边界。

在本文中，我们提出了 Vision-EKIPL，这是一种新颖的强化学习框架，旨在增强视觉推理模型的泛化能力。通过巧妙地引入由外部辅助模型生成的高质量动作，以指导策略模型在 RL 训练过程中的优化。这种创新方法显著扩展了模型的探索空间，使模型能够有效地超越传统推理边界，同时也显著加快了训练收敛速度和整体效率。大量实验证明了 Vision-EKIPL 的有效性，为推进视觉推理研究提供了有价值的见解，并引入了多模态学习的新范式。