207A+

通过游戏玩来概括：通过游戏学会推理

Yunfei Xie[1], Yinsong Ma[2], Shiyi Lan[3], Alan Yuille[2], Junfei Xiao[2*], Chen Wei[1†]

[1]Rice University, [2]Johns Hopkins University, [3]NVIDIA

Website https://yunfeixie233.github.io/ViGaL

Code & Model & Data https://github.com/yunfeixie233/ViGaL

Abstract

在多模态大型语言模型（MLLMs）中开发可泛化的推理能力仍然具有挑战性。受认知科学文献中关于游戏玩法促进可转移认知技能的启示，我们提出了一种新颖的后训练范式，称为视觉游戏学习（ViGaL），其中 MLLMs 通过玩类街机游戏开发多模态推理的域外泛化。具体而言，我们展示了在通过强化学习（RL）训练一个拥有 70 亿参数的 MLLM 时，通过简单的类街机游戏（例如贪吃蛇），它在多模态数学基准测试如 MathVista 上显著提高了其下游性能，并且在多学科问题如 MMMU 上同样表现优异，而在 RL 过程中并未见过任何解题过程、方程或图示，这表明其捕获了可转移的推理技能。值得注意的是，我们的模型在多模态推理基准测试中表现优于在多模态推理数据上调优的专业模型，同时保留了基础模型在通用视觉基准测试中的性能，而这是专业模型常常难以达到的挑战。我们的发现表明了一种新的后训练范式：合成的、基于规则的游戏可以作为可控且可扩展的预文本任务，释放 MLLMs 中的可泛化多模态推理能力。

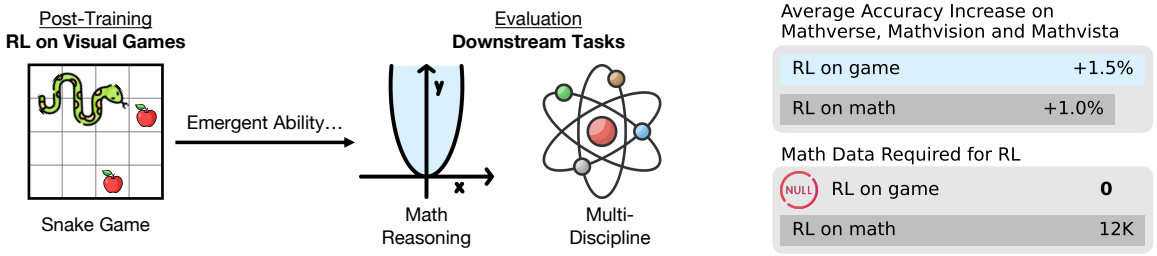


Figure 1 | ViGaL 的概述。左图：我们提出了一种新颖的训练后范式，通过强化学习调优多模态学习模型（MLLMs），使其能够玩诸如贪吃蛇 [25] 这样的街机游戏。我们证明了这种游戏后训练使多模态学习模型能够实现域外泛化，在不使用域内数学或多学科数据的情况下，提升其在需要数学、空间和多学科推理的下游多模态推理任务中的表现。右图：我们的 ViGaL （游戏上的强化学习）在三个多模态数学基准上较之 MM-Eureka [38] （数学上的强化学习）取得更高的平均准确率提升。这一点尤其值得注意，因为 MM-Eureka 使用了基于大规模、精心策划的数学数据集的强化学习，而 ViGaL 只使用游戏数据。详情见表 1 。

*Project Lead; †Corresponding Author

www.xueshuxiangzi.com

1. 介绍

游戏除了具有娱乐价值外，还为发展和研究一般推理和问题解决能力提供了丰富多样的结构化环境。人类自幼通过多种类似游戏的活动获取基础认知技能，如排列物品、导航空间和操控工具。这些经验促进了抽象思维的基本构件，如模式识别、空间推理和因果推断。在认知科学中，游戏被用作实验平台，以揭示人类思维的归纳偏见，如游戏“四连棋”中的计划深度，或者通过游戏“虚拟工具”揭示工具使用的认知基础。

AI 智能体也从类似人类游戏的环境中受益。这些环境鼓励探索、对稀疏奖励的鲁棒性以及从多模态输入中学习。例如，在通过捉迷藏 [4] 训练的智能体中观察到了工具使用的涌现，而 Atari 游戏已被纳入通用智能体的训练中 [41] 。通过在这些环境中学习，AI 系统发展了鲁棒且可转移的推理能力。

在这项工作中，我们特别研究了在后训练多模态大型语言模型的背景下使用游戏玩法以有效地进行推理。最近的研究表明，使用强化学习（RL）的后训练可以从其基础模型中解锁推理行为 [11, 39] 。这些 RL 训练的模型能够成功地在“思考后再说”，在输出最终答案之前生成内部的思维链追踪。

更重要的是，越来越多的证据表明，与另一种广泛使用的后训练方法——监督微调（SFT）相比，强化学习（RL）对分布外样本的泛化通常更加强健。例如，在 CLEVR [24] 上用 RL 训练的模型能够泛化到更具挑战性的 Super-CLEVR 基准 [32] ，在数学问题上训练的模型可以将推理扩展到物理问题 [38] ，以及在一个环境中训练的代理可以成功适应新的位置 [9] 。在每种情况下，RL 训练的模型始终优于其 SFT 对应模型。

虽然这些结果显示了分布外泛化的潜力，但它们通常仍然局限于单一领域。源任务和目标任务仍属于同一类别，例如 STEM 问题 [38] 或空间导航 [9] 。在这项工作中，我们探讨了一种更强形式的领域外泛化的潜力：从一个领域转移到完全不同的领域，具体而言，从游戏玩法转移到数学问题。

如图 1 所示，我们证明了对一个拥有 70 亿参数的多模态模型 Qwen2.5-VL-7B [3] 进行后训练，使其能够玩简单的街机风格游戏如 Snake [25] ，可以：（1）推广到解决从未见过的不同分布的 Atari 游戏（见第 1.1 节），（2）在多模态数学基准测试如 MathVista [35] 和多学科问答如 MMMU [56] 中获得增强的域外能力。尽管在强化学习过程中从未见过任何解题步骤、方程或图解，我们的模型不仅优于大型工业系统如 GPT-4o [23] ，同时也超越了那些在领域数据集上进行后训练的专用模型（见表 1 和 2 ）。此外，我们的模型在多模态推理基准测试中取得了进步，而没有牺牲其整体视觉能力，这是领域专用模型面临的挑战（见表 3 ）。有趣的是，最近的研究质疑在强化学习中是否需要领域内问题的真实标签 [43, 60] ，而我们的方法则表明，可能领域内的问题本身就是不必要的。

为什么有效？我们假设游戏玩法鼓励可推广的认知原始体或技能，这些技能可以转移到多模态推理基准测试中，例如空间理解和顺序规划。与在数学问题上的 SFT 或 RL 不同，那些可能会加强对训练数据的记忆 [9, 58] ，游戏玩法训练可能会激励更加灵活的表征和策略。支持这一观点，我们的消融研究显示，无论是提示还是奖励设计都在实现有效学习中起到了关键作用（见第 **??** 节）。我们还发现，不同的游戏强调不同的推理技能：贪吃蛇，一个玩家操纵“蛇”以避免碰撞和到达苹果的 2D 网格游戏，提升了多模态数学问题中关于 2D 坐标的表现，而旋转，一个识别 3D 物体旋转角度的拼图游戏，在角度和长度相关问题上表现更佳（见图 5 ）。此外，同时对这两个任务进行训练比单独训练任一游戏在下游多模态推理基准中表现出一致的更好表现，暗示了游戏的可扩展可能性（见表 1 ）。

这些结果表明了一种新的后训练范式。除了收集特定领域的数据，我们还可以设计可扩展且可控的前文本游戏，以激发可转移到下游任务的推理行为。合成游戏环境提供了结构化的、基于规则的奖励信号，具有高度的可控性，通过难度调度实现稳定的强化学习。当前有些研究正在利用游戏环境研究推理模型的属性，利用其可控性 [45] ，而我们则强调其跨域泛化能力。在这些环境中扩展数据也比收集人工标注数据容易得多。总之，这些发现表明一种利用合成任务

www.xueshuxiangzi.com

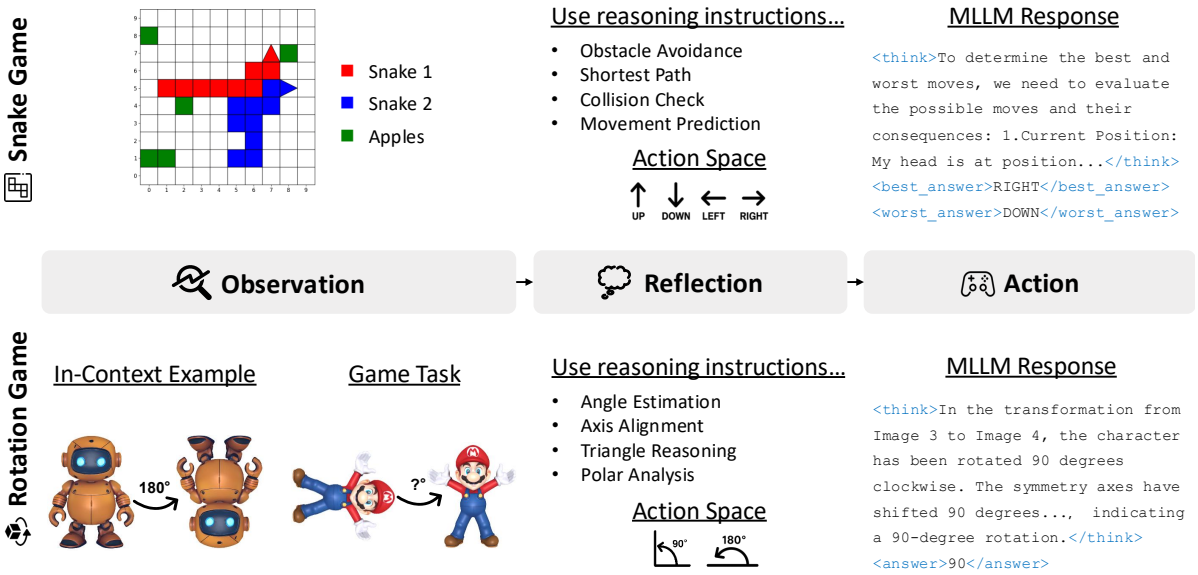


Figure 2 | 通过玩游戏来进行强化学习以训练后多模态语言模型的推理能力。我们提出通过玩视觉游戏进行强化学习后训练多模态语言模型。我们用两个游戏演示了这一点：经典街机游戏“贪吃蛇”[25] 和 “Rotation”，一个用于研究空间推理的自设计任务。在每场比赛中，模型接收多模态输入并遵循推理指导，例如，在 “贪吃蛇” 中的路径规划，在 “Rotation” 中的角度估计。模型通过反思来选择一个动作，输出其推理过程和决定，例如最佳/最差移动或预测角度，并接收奖励。通过玩游戏，模型获得了推理能力，可以迁移到下游的多模态推理任务，例如数学和多学科问答（图 3 ）。

（如游戏）的后训练范式的潜力，这让人想起视觉和语言中自监督学习的兴起 [13, 20, 40] ，通过在合成但有原则的前文本任务上进行预训练，可以实现广泛的泛化。

以下部分的组织如下：在 Sec. ?? 中，我们专注于游戏任务，介绍如何在游戏上进行强化学习的后期训练，并展示在未见过的游戏上的改进，以说明分布外泛化。在 Sec. 2 中，我们专注于领域外泛化评估，进一步表明在视觉游戏上的训练可以提升在未见过的视觉推理任务上的领域外泛化能力。在 Sec. ?? 中，我们总结了最近在 MLLMs 中关于强化学习和泛化的进展，并突出我们的 ViGaL 如何通过利用简单的游戏来实现更强的泛化能力。

在这一节中，我们介绍了一种新颖的后训练范式 ViGaL ，旨在增强泛化能力。第 ?? 节描述了用于训练和评估的 Snake 和 Rotation 游戏环境。第 ?? 节概述了我们框架中采用的强化学习算法。第 ?? 节展示了实现细节，并对分布内和分布外的游戏进行了全面评估。

如图 2 所示，在我们的 ViGaL 范式下，模型在一个游戏环境中进行训练，在该环境中，它从游戏环境接收状态，输出下一步的动作，并从环境中获得奖励作为反馈。形式上，给定指令 I ，每个任务可以被表示为一个部分可观测的马尔可夫决策过程（POMDP）：$(\mathcal{S}, \mathcal{A}, \mathcal{O}, T, R, \Omega)$ ，其中 $\mathcal{S}$ 是可能的环境状态集合，$\mathcal{O}$ 是模型可用的观察集合，$\mathcal{A}$ 代表模型在该游戏环境中可以做的动作。$T : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ 是状态转移函数，而 R 是来自环境的二进制奖励，表示动作的正确性。由于部分可观测性，智能体仅感知到观察 $o = \Omega(s)$ 。

我们设计了两个不同的游戏，Snake 和 Rotation，以研究所提出的范式（图 ~ 2 ），每个游戏针对不同的 MLLM 能力。Snake 游戏的灵感来自于竞争如何能够激发 MLLM 的推理能力 ~ [14] 。它通过挑战模型在与另一条蛇竞争时选择合适的行动来强调战略决策。与此同时，Rotation 游戏的灵感来自于自监督学习中的作为监督预设任务的旋转角度预测 ~ [19] 。该游戏评估 MLLM 的视觉感知能力，特别是理解复杂 3D 空间变换能力。通过这些互补的游戏，我们可以系统地探索和改进推理和感知这两种 MLLM 能力中的不同且基本的方面。

我们基于 SnakeBench [25] 建立了一个双蛇游戏。每个模型独立控制一条蛇。每条蛇的目标是吃到苹果、获取积分并击败对方。在时间点 t ，环境状态 s^t 包含蛇 $i \in \{1, 2\}$ 的坐标 $(x_{s_i}^t, y_{s_i}^t)$

www.xueshuxiangzi.com

，苹果的坐标 (x_a^t, y_a^t)，以及每条蛇选择的最后一次移动 A_i^{t-1}。这些元素被放置在一个 10×10 游戏棋盘上。在每个回合 t 中，每条蛇从 {up, down, left, right} 中选择其下一步动作 A_i^t。如果蛇与自身、另一条蛇或棋盘边界相撞，则会死亡。如果一条蛇死亡，另一条蛇获胜。如果两条蛇同时死亡，得分较高的蛇获胜。与只使用文本来表示游戏状态的 SnakeBench [25] 不同，我们使用游戏棋盘的图像和文本描述作为观察 $o^t = \Omega(s^t)$，以增强表示。

旋转游戏。 我们设计了一个旋转游戏来研究 MLLMs 的空间推理能力。我们向模型展示同一个 3D 物体的两个视图：一个初始视图 I_{init} 和旋转后的视图 I_{rot}。旋转后的视图是通过围绕朝向观察者的 z 轴，将 3D 物体从其初始方向旋转 90° 或 180° 而创建的。任务是确定应用了哪一个角度，90° 或 180°，以将物体从初始方向变为旋转后的方向。为了指导模型的推理过程，我们提供一个上下文示例，由另一对已知旋转角度的图片组成。类似于蛇游戏，我们同时提供带有图像和文本的观测。

我们应用基于规则的强化学习直接对针对视觉游戏的大型多模态语言模型进行后期训练，而不依赖监督学习作为热身。算法描述如下：

奖励设计 与其依赖于结果或过程为基础的奖励模型，我们遵循之前的方法 [22, 61]，使用简单的基于规则的奖励函数来避免奖励作弊 [18]，并帮助模型有效地学习如何玩游戏。

这个奖励函数由两个部分组成：一个准确性奖励和一个格式奖励。总奖励 r 被计算为准确性奖励和格式奖励 $r = r_{\text{accuracy}} + r_{\text{format}}$ 的和。如果答案正确，准确性奖励 r_{accuracy} 为 1，否则为 0。

格式奖励 r_{format} 检查响应是否遵循任务特定的格式：如果响应格式正确，则为 $r_{\text{format}} = 0.1$，否则为 $r_{\text{format}} = 0$。对于贪吃蛇游戏，期望的格式是：

```
<think>...</think><best_answer>...</best_answer><worst_answer>...</worst_answer>.
```

。正如格式所示，我们鼓励模型预测出既包括朝向苹果的积极移动，也包括导致失败的消极移动。这样的奖励机制鼓励对比性的决策，这不仅提高了模型的游戏能力，还提升了在视觉数学基准测试上的推理表现。我们在 Tab. 4b 中进行了效果消融。对于旋转任务，所需的格式仅仅是 `<think>...</think><answer>...</answer>.` 。

优势估计和策略更新。 在我们的强化学习训练阶段，我们采用了 REINFORCE Leave-One-Out (RLOO) 算法 [1, 28]。根据 Group Policy Gradient [10] 提出的技术，我们的实现中没有结合 KL 散度正则化。在没有限制策略变动幅度的 KL 约束下，模型可以更自由地探索解决方案空间，潜在地发现更好的推理策略。这样的设计选择使得我们的模型在强化学习阶段能够更灵活地适应。

在模型以图像作为输入来理解游戏的当前状态的同时，我们设计了一个结构化文本提示框架以提供游戏指导。我们的游戏提示由两部分组成：（1）游戏设置和（2）推理指导。（1）为了帮助模型理解游戏环境，我们在输入图像之外，以文本形式描述背景、当前游戏状态、规则、目标、行为空间等。（2）在推理指导部分，我们提供具体的思维指引，因为游戏可以通过各种思维链来进行。为了鼓励更广泛的思考，我们实施了不同类型的推理指导来引导决策过程。具体来说，我们使用了 GPT-4o [23] 来合成对贪吃蛇游戏的数学思维指导，例如“`finding the nearest apple by calculating Manhattan distances`”，以及对旋转游戏的空间思维指导，例如“`identify major symmetry axes in the original image`”。如图 4a 所示，这些推理指导有助于模型延长反应或内部思维链条。通过对游戏的推理指导，获得的推理能力在视觉数学问题的下游评估中具有广泛的适用性（表 4a）。文本提示设计的详细信息，包括使用的推理指导，见附录章节 A.1。

感谢使用合成游戏数据引擎，我们可以灵活地生成大规模训练数据，并精确控制难度级别。这完全消除了以前基于规则的强化学习工作中在特定领域数据（如数学 [2, 38]）的训练中使用的广泛数据过滤策略的需求，在这些情况下，难度定义困难且过滤可能显著减少数据集大小。在贪吃蛇游戏中，我们根据蛇的长度定义难度，较长的蛇创造了更复杂的游戏情况和更多受限的移动选项，这与人类玩贪吃蛇时感知难度的方式密切相关。在旋转游戏中，难度由两幅图像

www.xueshuxiangzi.com

之间的旋转角度决定，较小的角度差异带来更大的感知挑战。根据实证结果，我们为强化学习训练建立了最佳难度参数，并在表 4c 中进行消融实验。这个难度的控制进程使学习轨迹更加有效。

1.1. 在游戏中的实施与评估

我们采用 Qwen2.5-VL-7B-Instruct [3] 作为我们的基础模型。我们遵循 DeepSeek-R1 [11]，使用基于规则的格式奖励和准确性奖励的组合，核心 RL 算法为 RLOO [1, 28]。我们在基于 OpenRLHF [21] 的多模态输入 RL 框架内实现我们的训练。对于超参数，我们采用 MM-Eureka [38] 的默认设置，包括全局批大小为 128，展开批大小为 128，展开温度为 1.0，学习率为 $1e^{-6}$。训练使用 6 个 A100-80G GPU。

游戏训练数据。 我们构建了定制的游戏环境，以收集用于实验的训练数据。对于贪吃蛇游戏，我们利用 SnakeBench [25] 作为我们的数据引擎。这个环境允许我们输入动作以控制蛇的移动并生成游戏轨迹。为了创造有意义的游戏数据，我们基于近端策略优化（PPO）实现了一个具有线性输出层的策略网络。这个网络持续为两条试图收集苹果并避免死亡的蛇生成动作，使能够自动捕捉多样化的游戏轨迹用于强化学习训练。对于旋转游戏，我们使用 Hunyuan3D [48]，它是一个基于图像或文本指令生成 3D 网格的模型。我们从不同方向将每个网格渲染成 2D 图像，创建出具有关联旋转角作为真实标签的图像对，用于强化学习训练数据。

我们全面的数据生成流程能够在任何所需规模下以完全自定义的设置生成训练样本。在我们的实验中，我们合成了 36K 个 Snake 样本和 36K 个 Rotation 样本，这被证明足以实现收敛。数据合成的更多细节在附录第 **??** 节。为了评估 ViGaL 模型的游戏能力，我们在训练时未出现的各种状态下初始化这些环境。对于 Snake（见表 **??**），我们随机初始化游戏 10 次，并让两个模型相互竞争，以直接测量每个模型的胜场次数。对于 Rotation（见表 **??**），我们在由训练时未见的 3D 物体网格组成的全面验证集上测量旋转角度预测的准确性。我们的 7B 参数模型在 Snake 和 Rotation 游戏中 consistently 超过专有模型。这些结果证实，强化学习有效地释放了小型 7B 模型在需要环境理解、推理、规划和互动决策的视觉游戏中出色表现的能力。

在 Atari 游戏中的分布外泛化。 为了评估分布外泛化能力，我们在 Atari-GPT [53] 上测试了 ViGaL，它是一个评估 MLLMs 作为 Atari 视频游戏决策代理的基准，如图 6 所示。该基准由七种不同的 Atari 游戏组成，详细设置见附录第 B.1 节。我们遵循了大多数来自 Atari-GPT 的设置和提示，进行了小的修改以提供明确的 JSON 输出，确保所有模型的格式正确。根据 Atari-GPT [53]，我们以 1K 步的累积奖励作为评估指标，较高的奖励表示更好的表现。如表 **??** 所示，尽管只在贪吃蛇和旋转游戏上进行了训练，ViGaL 在 Atari 游戏上展示了显著的累积奖励提升。这一点尤其值得注意，因为 Atari 游戏在视觉外观和游戏策略上与我们的训练游戏完全不同。这些结果表明，我们基于规则的 RL 训练方法能够实现对完全未见过的游戏环境的强分布外泛化。

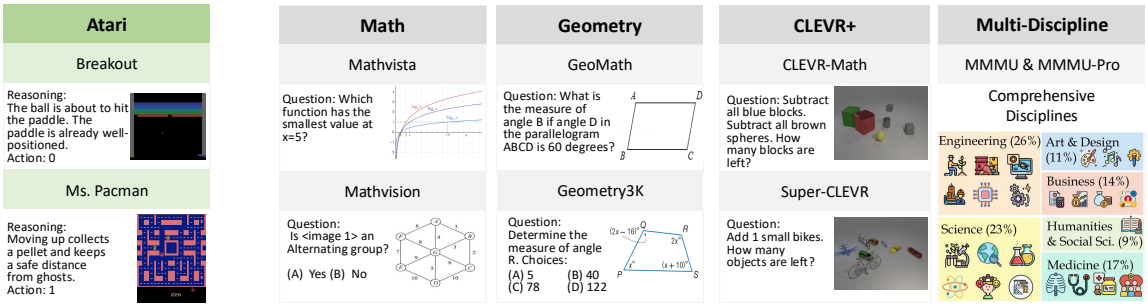
2. 视觉推理泛化

评估集合。 为了更清楚地了解多模态大模型（MLLM）性能的各个方面，我们遵循之前的研究 [31, 49]，系统地并仔细地将现有的基准测试分为两大类：(i) 以推理为导向的基准测试，其中需要多步骤或数学推理来解决问题，以及 (ii) 通用感知基准测试，主要评估广泛的视觉理解和感知能力。

对于以推理为导向的基准测试，我们通过在需要高级视觉推理技能的多样化任务集合中进行游戏，全面评估 RL 的视觉推理泛化能力。这些任务包括以数学为中心的任务，如数学和几何，还有超越数学的其他综合推理基准，如 CLEVR+ 和多学科。3b 图展示了每个基准的具体示例。

- Math 使用广泛使用的数据集来评估多模态数学推理：MathVista (testmini) [35]，MathVerse (testmini) [59]，以及 MathVision (test) [52]。MathVista 提供了涵盖 VQA、逻

www.xueshuxiangzi.com



(a) 分布外游戏。 (b) 非领域任务。

Figure 3 | 来自我们泛化推理基准的样本。我们通过两种类型的泛化来评估提出的 ViGaL：(a) 分布外泛化，即在我们的视觉游戏上训练的模型在未见过的 Atari 游戏 [53] 上进行测试；以及 (b) 域外泛化，即仅在游戏任务上训练的模型在各种多模态推理任务上进行评估，包括数学推理、几何问题解决、CLEVR+ 上的 3D 理解和 MMMU 系列上的多学科推理。

辑、代数和几何的多样化问题；MathVerse 强调代数和几何图像理解；MathVision 测试抽象的视觉推理。

- 几何学评估跨越数学图表、医学图像、图表和建筑布局的结构解释能力。它使用数据集 GeoMath（Geo170K [17]，Math360K [44]）和 Geometry3K [36]，包括选择题和非选择题。根据 Reason-RFT [46]，我们使用 820 个 GeoMath 样本和 800 个 Geometry3K 样本进行测试。
- CLEVR+ 通过复杂的 3D 基于积木的场景中的具有挑战性的算术问题来评估数学和空间推理技能的整合，包括关于 CLEVR-Math [33] 和 Super-CLEVR [32] 的子任务。根据 Reason-RFT [46]，我们使用来自 CLEVR-Math 和 Super-CLEVR 的每个子任务的 1K 测试样本。
- 多学科评估通过六个学科：艺术 & 设计、商业、科学、健康 & 医学、人文学 & 社会科学和技术 & 工程，来衡量大学水平的专家知识。我们遵循 MMMU [56] 验证集（900 个问题）和 MMMU-Pro [57] 总体得分（标准 10 选项和仅视觉设置的平均值）的评估设置。

对于通用的感知基准，我们系统地评估了全面的视觉能力。根据以往的工作，这些基准分为三种不同类型：通用、视觉中心和 OCR & 图表。具体来说，对于通用，我们评估用于多图像理解的 MuirBench [51] 和用于关系理解的 CRPE [26]。对于视觉中心基准，我们评估 MMVP [50]，RealWorldQA [54]，MMStar [7]，MME [15] 和 BLINK [16]，以全面评估感知、真实世界理解和多模态能力。对于 OCR & 图表理解，我们特别使用 AI2D [27] 来进行图表理解，使用 SEED-Bench-2-Plus [30] 来进行丰富文本的视觉理解，使用 DocVQA [37] 来进行文档理解，并使用 OCRBench [34] 来进行全面的 OCR 评估。

2.1. 主要结果

我们的方法即便在 RL 后训练期间没有直接接触领域内的训练数据时，在数学和其他推理任务上表现出显著的泛化能力。如表 1 所示，我们的方法明显优于专门在数学任务上进行 RL 训练的模型。例如，ViGaL Snake + Rotation 在数学任务上的准确率比 MM-Eureka-Qwen-7B [38] 高出 0.5 %，在几何任务上高出 28.7 %，即使 MM-Eureka-Qwen-7B 是专门在高质量的数学和几何数据集上进行训练的。

这一强大的泛化能力不仅限于数学。表格 2 显示，ViGaL 蛇 + 旋转在跨多学科推理的 MMMU 系列基准测试中，平均比 R1-OneVision-7B [55] 高出 5.4 %。这尤其值得注意，因为 R1-OneVision-7B 经由精心策划的数据集进行训练，涵盖多个学科。

这些实证结果表明，基于游戏的后期训练比直接在各种任务特定数据集上进行 RL 训练能更有效地发展基本的推理能力。此外，游戏环境似乎鼓励一般性问题解决策略，这些策略能够很好地推广到域外任务。

www.xueshuxiangzi.com

Model	Avg.	Math				Geometry		
		Avg.	MathVista	MathVerse	MathVision	Avg.	GeoMath	Geo3K
Proprietary Model								
GPT-4o [23]	47.5	48.1	61.4	50.2	30.4	46.8	50.2	43.5
Gemini-2.0-Flash [47]	55.4	56.4	73.4	54.6	41.3	54.4	55.3	53.5
General Multimodal Language Model								
InternVL2.5-8B [8]	48.2	41.2	64.4	39.5	19.7	55.2	63.0	47.3
Llava-OV-7B [29]	–	–	63.2	26.2	–	60.7	77.6	43.7
Qwen2.5-VL-7B [3]	46.3	47.7	68.0	49.0	26.0	44.8	44.0	45.6
Multimodal Reasoning Model Post-Trained on Qwen2.5-VL-7B								
R1-Onevision-7B [55]	40.9	46.8	64.1	46.4	29.9	35.0	45.4	24.5
R1-VL-7B [6]	40.9	42.7	63.5	40.0	24.7	39.0	42.0	36.1
MM-Eureka-Qwen-7B [38]	39.3	50.1	73.0	50.3	26.9	28.4	53.1	3.8
Reason-RFT-Zero-7B [46]	46.5	38.1	60.7	35.3	18.3	54.9	55.0	54.8
VLAA-Thinker-7B [5]	51.3	48.7	68.0	51.7	26.4	53.9	51.1	56.6
OpenVLThinker-7B [12]	52.1	47.8	70.2	47.9	25.3	56.4	49.2	63.5
ViGaL Snake	51.6	49.4	70.7	51.1	26.5	55.0	49.9	60.0
ViGaL Rotation	52.8	49.3	71.2	50.4	26.3	57.9	51.7	64.1
ViGaL Snake + Rotation	53.9	50.6	71.9	52.4	27.5	57.1	51.0	63.3

Table 1 | 关于多模态数学基准测试的主要结果。我们主要与基于 Qwen2.5VL7B [3] 上数学数据进行后训练的多模态推理模型进行比较。采用数学数据后训练的推理模型的结果是 不强调 ，而我们的 ViGaL 模型则仅使用视觉游戏进行后训练。每个 “Avg.” 列中后训练模型的最佳分数以粗体显示。

如表格 1 所示，在 Snake 上进行后训练在 CLEVR+ 基准测试中实现了最佳性能，而在 Rotation 上训练则在几何推理上表现更强。它们的比较优势在图 5 中进行了进一步的说明。值得注意的是，将模型同时在 Snake 和 Rotation 游戏上进行训练，使其能够从每个环境中学习互补的技能，将整体基准平均提高到 63.1 %。这些发现表明，结合多样化的游戏环境可以推动显著的性能提升。这表明 Visual Gaming Learning 作为一种有前景的训练范式，在无需大量特定领域数据的情况下，可以增强可推广的推理能力。在训练过程中扩大游戏的多样性可以始终如一地在广泛的视觉推理任务中提高性能。

在增强推理的同时保留一般视觉能力。 为了全面检查在推理任务上的泛化是否会导致一般视觉能力的下降，我们在更广泛的 MLLM 基准上评估了 ViGaL Snake + Rotation。如表 3 所示，与 RL 调整之前的 Qwen2.5-VL-7B 相比，我们的模型在实现更强的数学推理结果的同时，保持了可比的一般视觉性能。相比之下，其他通过 RL 后期训练提升数学表现的模型通常表现出一般视觉能力的大幅下降。这些结果表明，我们基于游戏的方法实现了数学的泛化，同时不损害其他视觉能力。

我们探讨了在 Snake 环境中的关键设计选择，对每个变体在下游基准上进行评估，并在表 4 和图 4 中报告结果。对应的 Rotation 环境的消融实验在附录第 B.2 节中提供。

我们在文本提示中使用诸如“`finding the nearest apple by calculating Manhattan distances`”之类的推理指令来指导模型的思考链。完整的文本提示见附录 A.1 节。在表 4a 中，我们展示了推理指令在三个域外基准测试中，将 Snake 的平均准确率从 59.5 % 显著提升了 1.9 %，达到 61.4 %。图 4a 显示，训练过程中整合推理指令显著增加了响应长度。这些结果凸显了在文本提示中添加推理指令的有效性，有助于 RL 训练以及推广到下游基准。

www.xueshuxiangzi.com

Model	Avg.	CLEVR + Avg.	CLEVRM	SCLEVR	MultiDiscipline Avg.	MMMU val	MMMUPro overall
Proprietary Model							
GPT4o [23]	55.9	51.2	68.1	34.3	60.5	69.1	51.9
Gemini2.0Flash [47]	–	46.3	64.9	27.6	–	71.9	–
General Multimodal Language Model							
InternVL2.58B [8]	54.8	64.4	93.5	35.3	45.2	56.0	34.3
LlavaOV7B [29]	42.9	49.4	69.7	29.1	36.5	48.8	24.1
Qwen2.5VL7B [3]	50.3	54.9	74.6	35.2	45.7	54.3	37.0
Multimodal Reasoning Model PostTrained on Qwen2.5VL7B							
R1Onevision7B [55]	53.7	65.1	75.5	54.7	42.3	51.9	32.6
R1VL7B [6]	53.9	68.0	87.4	48.6	39.7	50.0	29.4
MMEurekaQwen7B [38]	62.8	79.3	98.4	60.1	46.4	55.8	36.9
ReasonRFTZero7B [46]	58.6	76.2	99.4	53.0	40.9	51.2	30.6
VLAAThinker7B [5]	61.7	83.4	94.7	72.1	40.1	48.2	31.9
OpenVLThinker7B [12]	60.4	82.4	93.8	71.0	38.5	54.8	22.1
ViGaL Snake	64.4	82.6	92.6	72.6	46.2	55.8	36.6
ViGaL Rotation	63.3	80.7	93.0	68.3	45.9	54.1	37.7
ViGaL Snake + Rotation	64.7	81.7	91.9	71.4	47.7	58.0	37.4

Table 2 | 关于多模态空间和多学科推理基准的主要结果。我们将评估扩展到非数学推理任务，比较采用基于 Qwen2.5VL7B 进行域特定数据后训练的多模态推理模型。CLEVRM 表示 CLEVRMath，SCLEVR 代表 SuperCLEVR。对应的领域内数据后训练的推理模型的结果为 弱化 ，而我们的 ViGaL 模型仅通过视觉游戏进行后训练。每个“平均”列中后训练模型的最佳得分以粗体显示。

前文本游戏的奖励设计对下游任务很重要。 我们展示了强化学习的奖励设计对于游戏的下游任务起着关键作用。正如表格 4b 所示，我们首先让模型只预测最佳的下一步行动，这被定义为向最近的苹果移动同时避免死亡。在我们改进的奖励设计中，我们要求模型同时预测最佳和最差的下一步行动，其中最差的行动会直接导致游戏失败。正如图 4b 所示，同时预测最佳和最差动作可以提高推理长度，意味着更好的思维能力。更重要的是，这导致所有下游任务的改善，平均提升了 1.8 %。这些结果表明，在前置文本游戏中适当的奖励设计不仅可以提高游戏能力，还可以推广到下游任务。

此外，受到一些不需要标记奖励 [60] 或使用随机标签 [43] 来提高模型性能的早期工作的启发，我们还提供了一个随机奖励消融实验。在这个实验中，我们依然要求模型预测最佳和最差的动作，但使用随机动作作为标签。我们在表 4b 的最后一行报告了结果。在我们的游戏设置中，使用随机标签的强化学习平均报告 49.4 %，与基础模型相比并无显著提升，这与先前研究 [43] 的结论不同。可能的原因在于数据域和基础模型的差异，其中其他工作将随机标签应用于仅文本的数学数据，而我们的工作将随机标签应用于视觉游戏数据。

控制游戏难度以稳定地提高推理能力。 进行强化学习后训练的游戏玩法提供了一个独特的机会，可以轻松控制任务本身的难度。我们提出了一项关于难度控制重要性的消融研究。我们根据蛇的长度定义难度，其中蛇较长的状态被认为更难。对于我们控制的难度方法，我们使用蛇长度在 1 到 5 之间的中等范围内的状态来收集训练数据。如图 4c 所示，采用难度控制策略训练的模型在整个训练过程中保持了相对稳定的趋势，响应的长度逐渐增加。相反，不使用难度控制且包含难样本的模型在游戏玩法中经历了困难。如表 4c 所示，使用难度控制的方法达到

www.xueshuxiangzi.com

Model	Avg.	General			Vision-Centric						OCR & Chart				
		Avg.	Muir-Bench	CRPE $_{rel.}$	Avg.	MMVP	Real-WorldQA	MMStar	BLINK $_{val}$	MME $_{p}$	Avg.	AI2D $_{w.M.}$	SEED-Bench-2+	DocVQA val	OCR-Bench
Proprietary Model															
GPT-4o [23]	74.8	72.3	68.0	76.6	69.4	–	75.4	64.7	68.0	1614	82.6	84.6	72.0	91.1	736
General Multimodal Language Model															
Qwen2.5-VL-7B [3]	72.4	68.0	59.6	76.4	65.8	74.3	68.5	63.9	56.4	1698	83.3	83.9	70.4	95.7	864
Multimodal Reasoning Model Post-Trained on Qwen2.5-VL-7B															
R1-Onevision-7B [55]	–	66.8	46.3	87.3	56.5	61.3	58.0	57.8	48.7	1504	–	–	–	–	–
R1-VL-7B [6]	67.4	63.3	54.1	72.4	59.6	70.3	61.4	55.6	51.0	1657	79.2	81.7	66.4	89.4	81.0
MM-Eureka-Qwen-7B [38]	71.8	68.9	61.1	76.7	65.1	74.3	66.1	65.9	54.0	1626	81.5	84.3	68.2	92.0	87.0
Reason-RFT-Zero-7B [46]	68.4	66.9	58.5	75.2	58.5	58.0	65.3	59.1	51.6	1653	79.8	83.3	68.0	88.1	82.0
VLAA-Thinker-7B [5]	69.7	65.9	57.1	74.6	62.6	71.6	65.4	60.4	53.0	1593	80.6	83.4	67.4	90.9	84.5
OpenVLThinker-7B [12]	–	64.3	52.8	75.8	50.4	32.3	60.2	59.1	49.9	1513	–	–	–	–	–
ViGaL Snake + Rotation	72.2	68.6	60.5	76.7	65.7	74.6	67.3	65.4	55.6	1685	82.2	84.8	69.1	92.7	86.6

Table 3 | 在多模式语言基准上，主要结果针对更为一般和全面的视觉能力。我们与在 Qwen2.5-VL-7B [3] 上进行后训练的模型进行比较。表现最好的类别平均值用粗体突出显示。请注意，由于规模差异，MME $_{p}$ 未包含在以视觉为中心的类别平均准确度中。

61.4 % 的总体准确率，而不使用难度控制的方法为 60.6 %。这些发现表明我们的游戏引擎可以轻松生成具有合适难度的数据，以稳定强化学习训练，并有助于防止优化过程中模型崩溃。

强化学习在游戏上显示了数据可扩展性。 得益于游戏引擎的使用，我们可以以高度灵活性生成任意规模的数据。为了展示视觉游戏中强化学习的数据可扩展性，我们分别使用 16k 和 32k 蛇游戏样本进行实验。正如在表格 4d 中所示，将数据从 16k 扩展到 32k 平均在所有领域带来了 1.3 % 的性能提升。这表明所提出的 ViGaL 范式有潜力通过轻松扩展训练数据来提高下游性能，这与领域特定的人工标注数据面临的数据扩展挑战形成对比，这些需要大量人工努力。

文本和视觉都对更好的视觉推理有贡献。 为了分隔文本和视觉模态的贡献，我们进行了一项仅文本设定的消融研究。在这个设置中，我们仅使用文本描述来表示游戏状态——包括蛇的位置、苹果的位置和边界约束——在强化学习训练中。模型在仅文本输入的 Snake 游戏中训练后，在所有多模态基准测试中显示出显著的改进，平均表现从 49.1 % 提高到 59.6 %。结合视觉输入会额外增加 1.8 % 的性能提升。这些结果表明，多模态强化学习增强了视觉推理能力，文本和视觉模态都提供了互补贡献。图 4d 表明包括视觉模态在强化学习训练过程中导致响应长度增加，暗示了更详细的推理过程。

为了评估 ViGaL 的跨域泛化能力，我们将其与使用相同视觉游戏数据的监督微调（SFT）进行比较。表 4f 显示，使用贪吃蛇游戏数据的 SFT 显著降低了基础模型在数学推理和几何任务上的性能，分别下降了 9.7 % 和 12.7 %。虽然 SFT 在 CLEVR+ 上产生了适度的改进，但这些提升显著小于 RL 所取得的成果。总体而言，RL 提升了 12.3 % 的性能，而 SFT 则使性能下降了 1.9 %。这种鲜明的对比表明，RL 能够更好地保留并扩展模型在新领域的推理能力。

不同的游戏有利于不同的数学子领域。 我们假设，游戏过程可以促进诸如空间建模和顺序规划等基本技能的发展，这些技能可以转移到视觉数学问题上。不同的游戏可能会增强不同的推理能力。为验证该假设，我们分析了在如图 5 所示的 MathVerse [59] 子类别中，经由 Snake 或 Rotation 训练的 ViGaL 模型之间的准确性差异。Snake 游戏的训练显著提高了在表达式和坐标子类别上的表现。这两项任务涉及代数函数和图形表示的坐标级解释，与 Snake 的二维网格环境中的空间推理紧密相关。相比之下，Rotation 训练显著提高了角度和长度问题上的表现，这与 Rotation 对 3D 物体旋转角度进行推理的要求一致。这些结果表明，不同的游戏获得了与其独特游戏机制相对应的专门推理技能。此外，联合训练这两款游戏可以在所有推理类别中取得提升（见附录 Sec. **??** ）。我们还在附录 Sec. C 中包括了定性分析，展示了经过强化学习后的数学推理能力的提高。

强化学习（RL）在大语言模型（LLMs）中的应用越来越超越了监督微调（SFT）的界限，增强了推理能力。诸如 DeepSeek-R1 之类的纯文本模型表明，尤其是在规则基础奖励的情况下，RL 在复杂推理方面的效果显著。这一范式目前正积极扩展到多模态 LLM（MLLMs）中。最近的 MLLM 研究探索了用于提高视觉推理的 RL，汲取了 LLM 成功的灵感。例如，各种研究探

www.xueshuxiangzi.com

(a) 文本提示设计。

prompt	Avg.	Math	CLEVR+	Geo.
base model	49.1	47.7	54.9	44.8
w/o reasoning instr.	59.5	48.0	80.4	50.1
w/ reasoning instr.	62.3	49.4	82.6	55.0

(b) 奖励设计。

reward	Avg.	Math	CLEVR+	Geo.
基础模型	49.1	47.7	54.9	44.8
best moves	59.6	48.2	80.4	50.2
best & worst moves	62.3	49.4	82.6	55.0
w/ random label	49.4	47.5	55.4	47.5

(c) 难度控制。

difficulty control	Avg.	Math	CLEVR+	Geo.
base model	49.1	47.7	54.9	44.8
w/o difficulty control	60.6	48.8	81.4	51.8
w/ difficulty control	62.3	49.4	82.6	55.0

(d) 数据可扩展性。

training samples	Avg.	Math	CLEVR+	Geo.
base model	49.1	47.7	54.9	44.8
16K	60.1	48.9	81.2	50.3
36K	62.3	49.4	82.6	55.0

(e) 输入模态。

input modality	Avg.	Math	CLEVR+	Geo.
base model	49.1	47.7	54.9	44.8
text	59.6	48.5	80.1	50.3
vision & text	62.3	49.4	82.6	55.0

(f) SFT 与 RL。

post-training	Avg.	Math	CLEVR+	Geo.
base model	49.1	47.7	54.9	44.8
SFT	47.2	38.0	71.5	32.1
RL	62.3	49.4	82.6	55.0

Table 4 | 消融研究。我们在 Snake 环境中对 ViGaL 的不同方面进行了消融，并在下游基准测试中进行评估。在附录的 Sec. B.2 中有类似的旋转评估。每个基准测试由几个子任务组成（Tab. 1 和 Tab. 2），我们报告它们的平均值。基础模型是 Qwen2.5-VL-7B，其结果在 灰色 中。默认设置在 Tab. 1 和 Tab. 2 中以 蓝色 高亮显示。

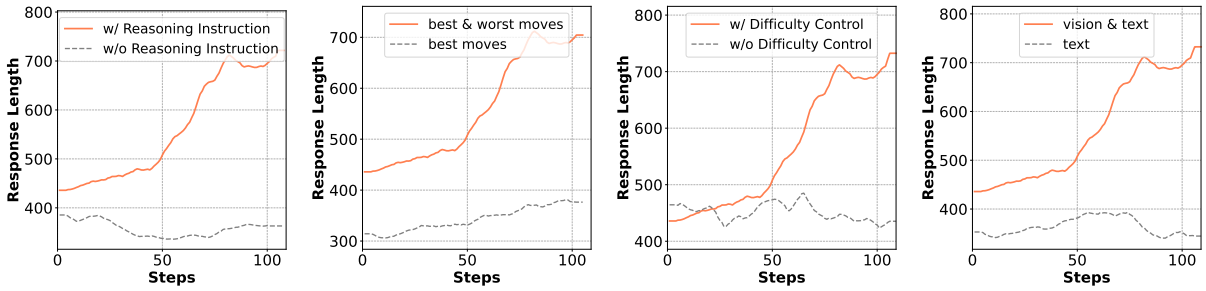
讨了多阶段训练、路径监督或基于规则的 RL，针对特定视觉子领域如几何和计数。其他研究则关注不同的 RL 算法，如流程奖励模型（PRMs），通常超越基于 SFT 的思维链生成。许多努力倾向于使用较为简单的规则奖励，以避免复杂奖励模型易受攻击的局限性。与那些依赖昂贵的、领域特定推理数据集的训练方法不同，我们的 ViGaL 范式通过扩展规则基础 RL 应用于简单的、合成视觉游戏，证明这些游戏可作为可扩展、成本效益高的预文本任务。

实现对新任务、分布和领域的稳健泛化是 MLLMs 开发中的一个核心目标。与 SFT 相比，RL 在提高分布外（OOD）泛化方面显示出希望，而开发像 CoT 这样的多步骤推理本身就是一种泛化形式。泛化通常通过在大量多样的指令跟随数据集上进行训练或通过显式训练一般推理能力来实现。尽管这些方法推进了 OOD 泛化，但它们通常在与训练数据相同的复杂视觉推理的广泛领域内操作。然而，我们的 ViGaL 范式探讨了一种更强的域外泛化形式。我们展示了从简单的合成游戏中学习的基本技能可以零样本迁移，以提高在完全不同的复杂领域（如视觉数学和多学科问题）上的表现，无需接触对应的领域特定数据。

我们引入了视觉游戏学习（ViGaL），这是一种新颖的后训练范式，其中 MLLMs 通过玩简单的街机风格游戏来学习可转移的推理能力。我们的核心发现是，在没有任何领域内数学数据的情况下，在像《Snake》和《Rotation》这样的游戏上进行强化学习，显著提升了 MLLM 在数学和多学科基准测试中的表现，超越了专业模型，甚至是大型专有系统。消融实验证实了游戏设计、奖励结构的重要性，并且强化学习的表现优于顺序微调（SFT），而不同的游戏解锁了不同的技能。我们认为游戏能够灌输基本的认知原语，提出了一种使用可扩展、可控的合成游戏作为强大前置任务的新途径，以解锁通用的推理能力。这项工作为探索基于游戏的学习以实现稳健的人工智能打开了大门。未来的方向包括研究不同游戏之间的协同效应以及深入理解迁移机制。

我们感谢 Haoqin Tu 和 Yuxuan Cheng 对本文稿的宝贵反馈。我们也感谢论文 Cambrian-1 [49] 的作者提供的模板参考。

www.xueshuxiangzi.com



(a) 文本提示。 (b) 奖励设计。 (c) 难度控制。 (d) 输入模态。

Figure 4 | 设计选择对响应长度影响的消融研究。实线橙色线表示完整配置，而虚线灰色线表示消融后的对照。图表展示了 (a) 加入推理指令，(b) 设计奖励时同时考虑最佳和最差的动作，(c) 实施难度控制，以及 (d) 使用多模态输入，随着训练的进行，都对响应长度的增加有贡献，这意味着更好的推理能力。

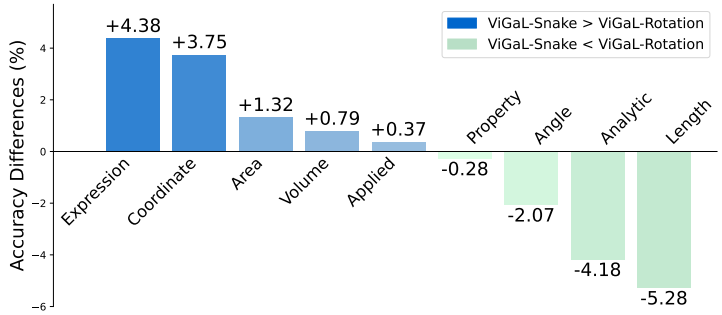


Figure 5 | Snake vs. Rotation: MathVerse 中的子领域差异。正值表示 ViGaL Snake 的结果更好，负值则表示 ViGaL Rotation 的性能更好。值得注意的是，Snake 在表达式和坐标上的增强最大，这些任务与 Snake 的二维网格对齐。Rotation 改善了角度和长度推理，反映了其对三维对象旋转的关注。

References

[1] Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce-style optimization for learning from human feedback in llms. In *ACL*, 2024.

[2] Sanghwan Bae, Jiwoo Hong, Min Young Lee, Hanbyul Kim, JeongYeon Nam, and Donghyun Kwak. Online difficulty filtering for reasoningoriented reinforcement learning. *arXiv:2504.03380*, 2025.

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. QwenVL: A versatile visionlanguage model for understanding, localization, text reading, and beyond. *arXiv:2308.12966*, 2023.

[4] Bowen Baker, Ingmar Kanitscheider, Todor Markov, Yi Wu, Glenn Powell, Bob McGrew, and Igor Mordatch. Emergent tool use from multi-agent autocurricula. In *ICLR*, 2019.

[5] Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, et al. SFT or RL? an early investigation into training R1Like reasoning large visionlanguage models. *arXiv:2504.11468*, 2025.

[6] Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. R1V: Reinforcing super generalization ability in visionlanguage models with less than $3. https://github.com/Deep-Agent/R1-V, 2025.

[7] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv:2403.20330*, 2024.

www.xueshuxiangzi.com

[8] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, et al. InternVL: Scaling up vision foundation models and aligning for generic visuallinguistic tasks. In *CVPR*, 2024.

[9] Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, et al. SFT memorizes, RL generalizes: A comparative study of foundation model posttraining. *arXiv:2501.17161*, 2025.

[10] Xiangxiang Chu, Hailang Huang, Xiao Zhang, Fei Wei, and Yong Wang. GPG: A simple and strong reinforcement learning baseline for model reasoning. *arXiv:2504.02546*, 2025.

[11] DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv:2501.12948*, 2025.

[12] Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and KaiWei Chang. OpenVLThinker: An early exploration to complex visionlanguage reasoning via iterative selfimprovement. *arXiv:2503.17352*, 2025.

[13] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, 2015.

[14] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *ICML*, 2023.

[15] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv:2306.13394*, 2023.

[16] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer, 2024.

[17] Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, et al. GLLaVA: Solving geometric problem with multimodal large language model. *arXiv:2312.11370*, 2023.

[18] Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. *arXiv:2210.10760*, 2022.

[19] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv:1803.07728*, 2018.

[20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

[21] Jian Hu, Xibin Wu, Zilin Zhu, Xianyu, Weixun Wang, Dehao Zhang, and Yu Cao. OpenRLHF: An easytouse, scalable and highperformance RLHF framework. *arXiv:2405.11143*, 2024.

[22] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, et al. VisionR1: Incentivizing reasoning capability in multimodal large language models. *arXiv:2503.06749*, 2025.

[23] Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, et al. GPT4o system card. *arXiv:2410.21276*, 2024.

www.xueshuxiangzi.com

[24] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li FeiFei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017.

[25] Greg Kamradt. Snake Bench: Competitive snake game simulation with LLMs. https://github.com/gkamradt/SnakeBench, 2025.

[26] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014.

[27] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016.

[28] Wouter Kool, Herke van Hoof, and Max Welling. Buy 4 REINFORCE samples, get a baseline for free! In *ICLR Workshop on Deep Reinforcement Learning Meets Structured Prediction*, 2019.

[29] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, et al. LLaVAOneVision: Easy visual task transfer. *arXiv:2408.03326*, 2024.

[30] Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*, 2024.

[31] Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, et al. A survey on benchmarks of multimodal large language models. *arXiv preprint arXiv:2408.08632*, 2024.

[32] Zhuowan Li, Xingrui Wang, Elias StengelEskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan L. Yuille. SuperCLEVR: A virtual benchmark to diagnose domain robustness in visual reasoning. In *CVPR*, 2023.

[33] Adam Dahlgren Lindström and Savitha Sam Abraham. ClevrMath: A dataset for compositional language, visual and mathematical reasoning. In *IJCLR*, 2022.

[34] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xucheng Yin, Cheng lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: On the hidden mystery of ocr in large multimodal models. *arXiv:2305.07895*, 2023.

[35] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, et al. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv:2310.02255*, 2024.

[36] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv:2105.04165*, 2021.

[37] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, 2021.

[38] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, et al. MMEureka: Exploring the frontiers of multimodal reasoning with rulebased reinforcement learning. *arXiv:2503.07365*, 2025.

[39] OpenAI. Introducing OpenAI o1. https://openai.com/o1/, 2024.

[40] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. *OpenAI*, 2018.

www.xueshuxiangzi.com

[41] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *arXiv:2205.06175*, 2022.

[42] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv:1707.06347*, 2017.

[43] Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei Du, Nathan Lambert, Sewon Min, Ranjay Krishna, Yulia Tsvetkov, Hannaneh Hajishirzi, Pang Wei Koh, and Luke Zettlemoyer. Spurious rewards: Rethinking training signals in rlvr. https://rethink-rlvr.notion.site/Spurious-Rewards-Rethinking-Training-Signals-in-RLVR-1f4df34dac1880948858f95aeb88872f, 2025. Notion Blog.

[44] Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, SeeKiong Ng, Lidong Bing, and Roy KaWei Lee. MathLLaVA: Bootstrapping mathematical reasoning for multimodal large language models. *arXiv:2406.17294*, 2024.

[45] Parshin Shojaee*, Iman Mirzadeh*, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity, 2025.

[46] Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. ReasonRFT: Reinforcement finetuning for visual reasoning. *arXiv:2503.20752*, 2025.

[47] Gemini Team. Gemini: A family of highly capable multimodal models. *arXiv:2312.11805*, 2023.

[48] Tencent Hunyuan3D Team. Hunyuan3D 2.0: Scaling diffusion models for highresolution textured 3d assets generation. *arXiv:2501.12202*, 2025.

[49] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

[50] Shengbang Tong, Zhuang Liu, Yuexiang Zhu, Xingjian Chen, Ruoyu Zhang, Bo Li, et al. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *arXiv preprint arXiv:2401.06209*, 2024.

[51] Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*, 2024.

[52] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with MathVision dataset. In *NeurIPS*, 2024.

[53] Nicholas R. Waytowich, Devin White, M.D. Sunbeam, and Vinicius G. Goecks. AtariGPT: Investigating the capabilities of multimodal large language models as lowlevel policies for atari games. *arXiv:2408.15950*, 2024.

[54] X.AI. Grok-1.5 vision preview. https://x.ai/blog/grok-1.5v, 2024.

[55] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, et al. R1OneVision: Advancing generalized multimodal reasoning through crossmodal formalization. *arXiv:2503.10615*, 2025.

www.xueshuxiangzi.com

[56] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.

[57] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024.

[58] Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? *arXiv:2504.13837*, 2025.

[59] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, et al. MathVerse: Does your multimodal LLM truly see the diagrams in visual math problems? *arXiv:2403.14624*, 2024.

[60] Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. Learning to reason without external rewards. *arXiv preprint arXiv:2505.19590*, 2025.

[61] Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and ChoJui Hsieh. R1Zero's "aha moment" in visual reasoning on a 2b nonsft model. *arXiv:2503.05132*, 2025.

www.xueshuxiangzi.com

A.

附录 内容

www.xueshuxiangzi.com

本节提供了关于游戏训练数据合成过程的额外实施细节，并扩展了第 1.1 节中概述的方法。

对于贪吃蛇游戏，该环境由一个 10×10 网格游戏板组成，初始状态下有两条长度为 1 格的蛇。在每个时间步 t 中，每条蛇分别接收一个动作来移动，从而产生一个新的游戏状态 s_{t+1} 。

为了生成有意义的动作，实现收集更多苹果同时保持生存的目标，我们基于近端策略优化（PPO） [42] 实施了一个策略网络。观察空间表示为一个 10×10 网格，不同的数值表示空单元格（0）、苹果（1）、代理的自身身体（2）和其他代理的身体（3）。为了融合时间信息，这些观察会在 4 个时间步上叠加，形成一个输入张量 $\boldsymbol{X} \in \mathbb{R}^{10\times10\times4}$ 。

策略网络架构由两个带有 3×3 内核的卷积层组成，随后是全连接层。第一个卷积层有 $C_1 = 16$ 个输出通道，而第二个有 $C_2 = 32$ 个输出通道，两者之后都接有 ReLU 激活函数。扁平化后，这些特征通过一个拥有 256 个单元的全连接层，然后输出四个可能运动方向（右、左、上、下）的动作 logits，这些 logits 随后通过 softmax 转换成概率分布 $\boldsymbol{\pi}(a|s)$ 。价值函数遵循类似的架构但产生一个标量输出 $V(s)$ 。

为了防止蛇轻易死亡，我们通过对危险动作（例如，移动到墙壁或其他蛇体）的逻辑值进行屏蔽，来加入阻止自杀性移动的动作先验。该模型使用带有熵正则化系数 $\beta = 0.01$ 的标准 PPO 目标来鼓励探索，同时配合一个价值函数系数 $\lambda = 0.5$ 和裁剪参数 $\varepsilon = 0.2$ 。在训练过程中，我们使用大小为 2048 的缓冲区和大小为 32 的小批量，采用 Adam 优化器，学习率为 $\eta = 10^{-3}$ 。

代理在收集苹果时获得 $r = +1$ 的奖励，而在被弄死时受到 $r = -1$ 的惩罚。这种奖励结构结合 PPO 算法，使代理能够学习复杂的行为，如避开障碍物、追逐苹果和多步轨迹规划。利用这个策略网络，我们持续从蛇游戏中收集数据，生成多样的游戏场景，作为下游强化学习训练的示例。

对于旋转游戏，训练数据包括合成生成的视觉拼图，重点是 3D 空间推理，特别是对象旋转的理解。我们总共使用了一个多样化的 540 种独特 3D 对象网格集合，其中 408 个网格来源于 Hunyuan3D 2.0 [48] ，还有 132 个网格来源于 Hunyuan3D 2.5。Hunyuan3D 是一个大规模 3D 资产生成系统，能够生成高分辨率纹理对象，为我们的游戏提供各种形状和纹理。我们的定制数据生成管道为每个网格生成了一对图像 $(I_{\text{init}}, I_{\text{rot}})$，表示对象在定义旋转前后。

每对 $(I_{\text{init}}, I_{\text{rot}})$ 的生成遵循一个精确的顺序。首先，为了为 I_{init} 建立一个多样化的初始视点，3D 对象经过基本定向：绕其 x 轴 0° 旋转，绕其 y 轴选择 $\{0°, 45°, \ldots, 315°\}$ 集合中的一个角度，绕其 z 轴 0° 旋转。为了进一步增强视觉多样性并防止从标准姿势中学习到简单的变换，随后施加了从 $\{0°, 30°, \ldots, 330°\}$ 选择的附加 z 轴旋转。这些复合初始变换后的对象渲染生成了 I_{init} 图像。随后，通过对 I_{init} 所示对象状态应用目标旋转生成了 I_{rot} 图像。这个目标旋转仅在 z 轴上进行，旋转角度为 90° 或 180° 之一，这也作为样本的真实标签。我们的坐标系统定义为 x 轴指向右，y 轴指向上，z 轴从屏幕向外指向观察者；因此，所有目标旋转都发生在图像平面中。

所有物体均以 512×512 像素分辨率渲染，使用提供正面视图的一致性透视相机，并在标准化光照条件下进行。不包括坐标轴的可视化结果在渲染图像中。此过程从分配用于生成测试实例的 537 个网格池中派生出大约 32k 个唯一 $(I_{\text{init}}, I_{\text{rot}})$ 对。正如在第 **??** 节中详细说明的，每个呈现给 MLLM 的训练实例包括四幅图像——一个示例对 $(I^{\text{ex}}_{\text{init}}, I^{\text{ex}}_{\text{rot}})$ 和一个任务对 $(I^{\text{task}}_{\text{init}}, I^{\text{task}}_{\text{rot}})$。示例对使用一组专用的 3 个网格单独生成，以确保情境示例中的物体与任何给定提示的测试部分中使用的物体不同。示例对和测试对均通过上述方法生成。

www.xueshuxiangzi.com

A.1. 视觉游戏学习中的训练提示

Prompt for Snake Game

Your role is to guide a snake within a Snake game featuring multiple apples.

This game is played on a board of size 10 by 10. The board uses a standard Cartesian coordinate system, where (0,0) represents the bottom-left position and (9,9) is the top-rightmost coordinate.

Apples at: { apple_position }

Direction of Your Last Action: { last_action }

Rules:
1) If you move onto an apple, you grow and gain 1 point.
2) If your head moves to a position where its coordinates (x, y) are outside the board boundaries (meaning $x < 0$, $x > 9$, $y < 0$, or $y > 9$), or into a space occupied by another snake's body, or into a space occupied by your own body, you die. That's the worst move.

3) The goal is to prioritize snake not die, then efficiently collecting apples. First avoid the worst move, then for each apple, find the nearest apple by calculating Manhattan distances. But only choose best next move to get closer the nearest apple if you can confirm best next move will not run outside the range of the listed coordinates, run into the position of another snake, or yourself. Otherwise it will be the worst move.

Your snake with the ID { snake_id } in { snake_color } has its head now positioned at { snake_position } , and its body extends to { body_position } You should avoid your next move into your own snake's position.

Enemy snakes in { enemy_color } positions: { enemy_position } .

Decreasing your x coordinate is to the LEFT, increasing your x coordinate is to the RIGHT.
Decreasing your y coordinate is DOWN, increasing your y coordinate is UP.

Read out another snake's position and apple position. Try to predict another snake's next move and avoid colliding with it.

Best answer is one of next move that is the closest to the apple and not lead to your death. Worst answer is all of next moves 1. makes your head's coordinates (x, y) are outside the board boundaries, meaning $x < 0$, $x > 9$, $y < 0$, or $y > 9$. 2. moves into a position occupied by another snake's body. 3. moves into a position occupied by body of yourself.

Check all the next moves to list out all the worst moves in `<worst_answer>` tag. If no worst answer, return None for worst answer, e.g., "`<worst_answer>None</worst_answer>`"

The best answer and the worst answer are mutually exclusive and different.

You need first to give your reasoning process then to choose one of best next move and worst next move from ['UP', 'DOWN', 'LEFT', 'RIGHT'].

The reasoning process and answer are enclosed within <think> </think>, <best_answer> </best_answer> and <worst_answer> </worst_answer> tags, respectively, i.e., "<think> reasoning process here </think><best_answer> one best move here </best_answer><worst_answer> all worst moves here </worst_answer>"

www.xueshuxiangzi.com

Prompt for Rotation Game

I'm showing you 4 images. Images 1-2 are an example pair, and Images 3-4 are the test pair. In each pair, the first image shows the initial orientation, and the second shows the object after rotation.

EXAMPLE OF ROTATION # #

Example: Image 1 shows the initial view and Image 2 shows the object after a 180 degree rotation.

YOUR TASK # #
Now, considering the transformation from Image 3 (initial) to Image 4 (rotated)
. Determine the angle of rotation from Image 3 to Image 4 on the plane
Analyze the rotation carefully using the example pair (Images 1-2) as a reference.

1. Coordinate System Transformation:
- Draw an x-y coordinate system on both original and rotated images with origin at center
- Identify a distinct feature point and note its coordinates in both images
- Apply rotation matrix equations to verify the transformation

Example: A star icon at coordinates (3,1) in the original image appears at (-1,3) in the rotated image. Testing with the 90ř clockwise rotation matrix [cos(90ř), sin(90ř); -sin(90ř), cos(90ř)] confirms the transformation from (3,1) to (-1,3), verifying a 90ř clockwise rotation.

2. Angular Displacement Measurement:
- Mark the image center as the origin in both images
- Draw a straight line from center to a distinctive feature in both images
- Measure the angle between these two lines using counterclockwise as positive

Example: A line from center to a red dot makes a 30ř angle with horizontal in the original image. In the rotated image, this line makes a 210ř angle with horizontal. The difference (180ř) indicates a clockwise 180ř rotation.

3. Symmetry Axis Tracking:
- Identify major symmetry axes in the original image
- Locate the same symmetry axes in the rotated image
- Calculate the angular displacement between original and rotated axes

Example: A rectangular logo has vertical and horizontal symmetry axes. After rotation, the vertical axis now points right and horizontal points down. This 90ř shift of both axes confirms a clockwise 90ř rotation.

4. Triangle Configuration Analysis:
- Select three non-collinear distinct points forming a triangle in both images
- Compare the orientation of this triangle in both images using vector cross products
- Determine rotation angle from the triangle's orientation change

Example: Three points form a right triangle with vertices clockwise arranged. After rotation, the same triangle has its vertices arranged in counterclockwise order while maintaining the same shape. This inversion indicates a clockwise 180ř rotation.

5. Polar Coordinate Comparison:
- Convert key points to polar coordinates (r, θ) relative to image center
- Compare θ values of the same features in original and rotated images
- Calculate consistent angular difference across multiple points

Example: A feature at polar angle 45ř in the original image appears at 135ř in the rotated image. Another feature shifts from 10ř to 100ř. Both show a +90ř shift in polar angle, confirming a clockwise 90ř rotation.

Choose the rotation angle from this list: ['counter clockwise 90', '180']

The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., "<think> reasoning process here </think><answer> answer here </answer>."

www.xueshuxiangzi.com

我们提供了关于使用 Snake 和 Rotation 游戏进行 RL 训练中的训练提示的更多详细信息。我们的游戏提示包含两个主要部分：游戏设置和推理指令。对于 Rotation 游戏，为了促进多样的问题解决方法，每个训练实例都包括一个从预定义的五个空间推理策略提示集中随机选择的提示：1. 坐标系变换, 2. 角位移测量, 3. 对称轴跟踪, 4. 三角形配置分析, 或 5. 极坐标比较。粗体文本表示由 GPT-4o [23] 合成的推理指令。

www.xueshuxiangzi.com

B. 评估

B.1. Atari 游戏的评估细节

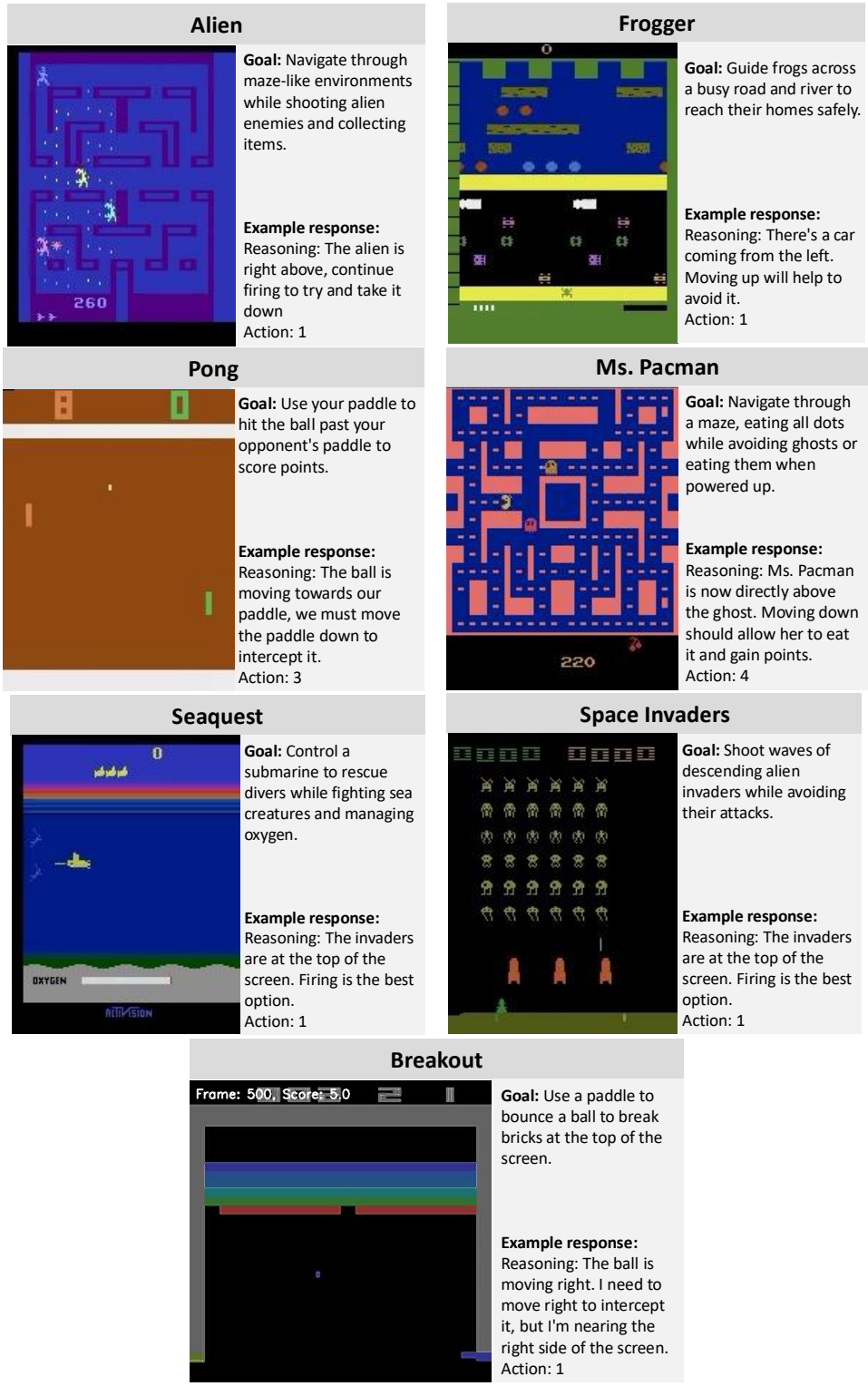


Figure 6 | 用于评估的 Atari 游戏模型的目标和示例响应。我们实现了 Atari-GPT 中的 [53] 中的 7 种 Atari 游戏。

www.xueshuxiangzi.com

为了评估分布外泛化能力，我们在 Atari-GPT [53] 上测试了 ViGaL ，如图 6 所示，这是一个用于评估 MLLMs 在 Atari 视频游戏中作为决策代理的基准。该基准包含七款不同的 Atari 游戏：Alien、Frogger、Pong、Ms. Pacman、Seaquest、Space Invaders 和 Breakout。这些游戏提供了与 Snake 游戏和 Rotation 游戏不同的多样视觉环境，并需要不同的策略来达成目标，从而使其成为评估 ViGaL 分布外泛化能力的理想测试平台。

为了评估，我们将游戏帧作为像素观察输入到我们的模型中，遵循 Atari-GPT 中建立的协议。具体来说，每个游戏帧从 $210 \times 160 \times 3$ 调整到 $512 \times 512 \times 3$ ，然后连同游戏特定的动作信息一起提供给我们的模型。我们维护一个上下文缓冲区，其中包含前两个帧和响应以及当前帧，以实现时间推理。按照 Atari-GPT 的要求，我们实现了 8 帧跳帧，这将 ALE 中的标准 4 帧跳帧扩展以降低计算强度，同时保持游戏连续性。

我们通过四次独立的每次 1000 步的展开过程评估我们的方法，并报告平均累计奖励，结果如表 **??** 所示。

B.2. 在旋转游戏上的消融

Table 5 | 消融研究。类似于在表 4 中的评估，我们分析了在旋转游戏中我们训练后策略的不同方面如何影响下游泛化基准。基础模型是 Qwen2.5-VL-7B，结果显示在 灰色 中。表 1 和表 2 中的默认设置在 蓝色 中被突出显示。我们观察到表 4 中报告的每个策略同样的改进趋势。

(a) 提示设计。

prompt	Avg.	Math	CLEVR+	Geo.
base model	49.1	47.7	54.9	44.8
w/o Reasoning Instruction	61.4	48.9	80.4	54.8
w/ Reasoning Instruction	62.6	49.3	80.7	57.9

(b) SFT 与 RL。

post-training	Avg.	Math	CLEVR+	Geo.
base model	49.1	47.7	54.9	44.8
SFT	55.6	44.0	75.4	47.5
RL	62.6	49.3	80.7	57.9

(c) 难度控制。

difficulty control	Avg.	Math	CLEVR+	Geo.
base model	49.1	47.7	54.9	44.8
w/o difficulty control	61.0	48.0	80.2	54.8
w/ difficulty control	62.6	49.3	80.7	57.9

如表 5 所示，我们进行了一项类似于表 4 的消融研究，但将蛇游戏环境替换为旋转游戏。我们的结果证明了每种策略在下游泛化基准上的一致改进趋势。

具体而言，我们通过改变两张图像之间的旋转角度来控制任务难度。在不受控制的难度设置下，图像之间的旋转角度可以是顺时针 90ř、逆时针 90ř 或 180ř。然而，我们发现显式要求模型区分顺时针和逆时针旋转会导致训练困难。因此，我们去掉了这一要求，仅保留顺时针 90ř 和 180ř 旋转的选项。

与贪吃蛇游戏不同，我们无法执行表格 4e 中显示的消融实验，因为旋转游戏本质上依赖于视觉且需要视觉输入。同样地，我们也无法执行表格 4b 中的消融实验，因为旋转游戏只提供二元答案选项，无法同时有意义地指定" 最佳" 和" 最差" 答案。

www.xueshuxiangzi.com

B.3. 与数学数据的协同整合

Model	Math Avg.	MathVista	MathVerse	MathVision
base model	47.7	68.0	49.0	26.0
MM-Eureka-Qwen-7B	50.1	73.0	50.3	26.9
ViGaL (w/o Math Data)	50.6	71.9	52.4	27.5
ViGaL (w/ Math Data)	51.8	72.3	54.5	27.7

Table 6 | 在数学数据上的消融研究。我们进行了一项实验，在数学数据 MMK12 上进一步训练 ViGaL 。基础模型是 Qwen2.5-VL-7B，其结果在 灰色 中展示。平均准确率最高的设置在 蓝色 中被突出显示。

虽然我们的工作主要展示了在没有数学数据训练的情况下数学性能的提升，但我们进行了额外的实验，以探索在我们的训练流程中整合数学数据的协同效益。在我们的实验设置中，我们实施了一个两阶段的训练过程。在阶段 1 中，我们遵循原来的方法，仅在 Snake 和 Rotation 游戏上训练模型。对于阶段 2，我们在 MMK12 [38] 上训练了我们的模型，这是一个包含大约 12k 个例子的多模态数学推理数据集。我们保持与 MM-Eureka-Qwen-7B [38] 相同的数据和训练设置。唯一的区别是我们的模型在视觉游戏上的额外阶段 1 训练。

如表 6 所示，与仅使用阶段 1 训练相比，在阶段 2 中整合数学数据在三个数学基准测试中平均提高了 0.9 %。这表明了我们的视觉游戏学习方法与数学数据微调之间的协同关系。此外，尽管两个模型使用相同的数学数据，ViGaL （带数学数据）在数学基准测试中平均表现显著优于 MM-Eureka-Qwen-7B 1.4 %。这些结果表明，视觉游戏学习可以作为一个有效的基础训练阶段，使用领域特定的数据进一步增强以提升目标任务的性能。根据第 2.1 节中的讨论，我们的分析显示每个游戏在模型中发展了不同的推理能力。为研究潜在的组合效益，我们进行了实验，模型同时接受 Snake 和 Rotation 游戏的训练。图 **??** 表明联合训练有效地结合了每个独立游戏的优势，提升了在每个游戏展示特别效果的数学领域中的表现，在 Mathverse 上获得了更大的整体提升。这些结果表明，策略性地结合具有互补优势的游戏提供了一种简单但有效的方法来增强模型的泛化能力。

我们通过评估通过 @k 指标来探索使用不同 RL 方法训练的模型的推理能力边界。此指标衡量的是至少有一个 k 个独立模型样本解决给定问题的概率，这表明了模型推理能力的真实范围或边界——本质上是指模型在给予足够多的采样尝试后可能解决的问题。

我们评估了三个模型的 @k 表现：未经过 RL 训练的基础模型、MM-Eureka-Qwen-7B-Instruct 和我们的 ViGaL 。如图 **??** 所示，随着 k 增加，我们的 ViGaL 在 Mathverse 上的 @k 得分一致地呈现上升趋势。这一发现表明，当允许进行多次推理尝试时，我们的方法可以有效解决复杂问题，揭示在单次样本评估中无法显现的能力。

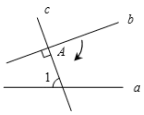
此外，与其他通过强化学习训练的模型 MM-Eureka-Qwen-7B-Instruct 相比，我们的模型在随着 k 增加而通过 @k 的改进更加显著。这表明 ViGaL 具有更广泛的推理边界和更强的推理能力，使其在有足够的机会探索不同的解决途径时，能够解决更广泛的问题。

最后，我们的结果表明，随着 k 增加，没有经过强化学习训练的基础模型最终会优于经过强化学习训练的模型。这与 [58] 中的发现一致，该研究指出了可验证奖励的强化学习（RLVR）的一个基本限制：虽然强化学习训练可以显著提高在小 k 值（例如，通过 @1 ）的性能，但基础模型对可解问题有更广泛的覆盖率。这表明存在一种权衡，即强化学习优化专注于解决高概率问题，但以牺牲更广泛的解决方案覆盖率为代价。未来的工作应探索 RLVR 算法，以提高在所有 k 值上的通过 @k 性能，从而有效地将推理边界扩展到基础模型之外。

www.xueshuxiangzi.com

C. 案例研究

Rotation

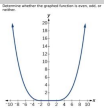


Question:
As shown in the figure, the straight line a and the straight line b are intercepted by the straight line c. Line b is perpendicular to c, and the foot of the perpendicular is point A. ∠1 = 70°. If line b is parallel to line a, how many degrees clockwise can line b be rotated around point A?
Choices: A: 70° B: 50° C: 30° D: 20°

Qwen2.5-VL-7B-Instruct	Ours
1. **Identify the given information:** Line b is perpendicular to line c. The foot of the perpendicular is point A. Angle 1 is 70°. Line b is parallel to line a.	1. **Identify the given information:** Line b is perpendicular to line c . The foot of the perpendicular is point A . ∠ 1 = 70°. Line b is parallel to line a.
2. **Understand the geometric relationships:** Since line b is perpendicular to line c, angle 1 is the angle between line a and line c. Since line b is parallel to line a, the angle between line b and line c is the same as the angle between line a and line c, which is 70°.	2. **Analyze the geometric relationships:** Since line b is perpendicular to line c , ∠BAC = 90°. Line b is parallel to line a, so the corresponding angles are equal. The angle ∠ 1 is given as 70°.
3. **Determine the angle of rotation:** The angle of rotation is the difference between the initial angle 70°. and the final angle 0°, since line b is perpendicular to line c.	**3. Determine the rotation angle:** - The angle ∠ 1 represents the angle between line b and the vertical line c . Since line b is perpendicular to line c , the angle between line b and line a (which is parallel to line b) is 90° - 70° = 20° .
Therefore, the angle of rotation is 70°.Thus, the line b can be rotated 70°clockwise around the point A. **The correct answer is: A**	**Therefore,** the line b can be rotated 20° clockwise around the point A .**The correct answer is D.**

(a) 来自 Mathverse 的一个案例研究。基础模型误解了几何配置和旋转方向，而我们的模型正确识别出垂直关系并计算出正确的角度。

Coordinate



Question:
Determine whether the graphed function is even, odd, or neither.

Qwen2.5-VL-7B-Instruct	Ours
Let's analyze the graph:	
1.Symmetry with respect to the y-axis: If the graph were symmetric with respect to the y-axis, then for every point (x, y) on the graph, the point (-x, y) would also be on the graph. **However, the graph in the image does not appear to be symmetric with respect to the y-axis. For example, the point (3, y) does not have a corresponding point (-3, y) on the graph.**	**Let's examine the graph of the function.** The graph shows a parabola that opens upwards and is symmetric about the y-axis. This symmetry indicates that the function is even.
2. Symmetry with respect to the origin: If the graph were symmetric with respect to the origin, then for every point (x, y) on the graph, the point (-x, -y) would also be on the graph. However, the graph in the image does not appear to be symmetric with respect to the origin. For example, the point(3, y) does not have a corresponding point (-3, -y) on the graph.	**To confirm, we can check a few points:** For x = 2 , the function value is f(2). For x = -2 , the function value is f(-2) .Since the graph is symmetric about the y-axis, f(-2) = f(2) . This confirms that f(-x) = f(x) , which is the definition of an even function.
Since the graph does not exhibit symmetry with respect to either the y-axis or the origin, the function is neither even nor odd.**The final answer is: neither.**	Therefore, the function is even.**The final answer is: even.**

(b) 来自 Mathverse 的案例研究。基础模型误解了图形中的关键视觉信息，如对称性和坐标，而我们的模型显示出对数学元素的准确视觉感知。

Figure 7 | 基于规则的强化学习训练后，基础模型与我们的模型比较，显示在几何和坐标问题上的视觉-数学推理能力有所提升。

我们在下面提供了定量比较示例，以展示在 RL 训练后数学问题推理上的改进。在图 7a 中，解决几何角问题时，基础模型未能正确地解释垂直线与对应角之间的关键关系。它对角度测量做出了矛盾的假设，导致所需旋转的计算错误。相比之下，我们的 ViGaL 准确地跟踪了几何约束，并正确计算了初始位置和目标位置之间的角度差异。在图 7b 中，分析图中的函数特性时，基础模型错误地声称函数没有对称性，尽管有明显的视觉证据。它未能识别图像中抛物线的基本 y 轴对称性。我们的模型立即识别出这一关键对称模式，并正确应用偶函数的适当数学定义，展示了增强的对数学结构的视觉感知能力。

www.xueshuxiangzi.com