

从诊断到改进：探讨视觉语言模型中的时空物理推理

Tiancheng Han^{1, 2}, Yunfei Gao³, Yong Li¹, Wuzhou Yu¹, Qiaosheng Zhang^{†2, 4}, Wenqi Shao^{†2, 4}

¹Tongji University, ²Shanghai Innovation Institute, ³East China University of Science and Technology, ⁴Shanghai AI Laboratory

Abstract

空间物理推理是理解真实物理世界的基础能力，是构建稳健世界模型的重要一步。尽管最近的视觉语言模型 (VLMs) 在多模态数学和纯空间理解等专门领域取得了显著进展，但它们在空间物理推理方面的能力基本上未被探索。本文对主流视觉语言模型进行了全面的诊断分析，揭示了当前模型在这一关键任务上的表现不佳。进一步的详细分析表明，这种欠佳的表现主要归因于由类人先验引起的偏差和缺乏深度推理。为了应对这些挑战，我们对 Qwen2.5-VL-7B 应用了监督微调，随后进行了基于规则的强化学习，结果在空间物理推理能力上取得了显著改善，并超过了领先的专有模型。尽管如此，尽管取得了这一成功，该模型对新物理场景的泛化能力仍然有限——这凸显了在空间物理推理中引入新方法的迫切需求。

1 介绍

直观地推理物理世界的能力——预测物体如何在物理定律下相互作用——是智能的基石。我们将这种能力称为空间物理推理，它包括对空间信息的感知以及基于感知对物理定律的推理。对于任何代理在现实世界中有效运作而言，这都是一个基本的前提条件。

近期在视觉语言模型 (VLMs) 方面的进展在各种任务中显示了显著的能力。然而，对其推理能力的现有研究主要局限于特定领域，如多模态数学 (Meng et al. 2025) 或纯空间理解 (Ouyang et al. 2025; Liao et al. 2025)，这仅涉及空间关系而不需要任何物理推理。当涉及到物理学时，它通常位于明确的教科书式情境 (Shen et al. 2025; Xiang et al. 2025; Dai et al. 2025; Zheng et al. 2025) 中，这留下了一个关键的漏洞，即理解 VLMs 在空间物理推理任务中的表现。

弥合这一差距对于实现世界模型等有雄心的 AI 目标至关重要，因为对物理结果的强健内部模型是通向 AGI (Assran et al. 2025) 的潜在途径。此外，强大的物理认知直接提高了在下游具身 AI 任务中的表现 (Chow et al. 2025)。在这项工作中，我们首先对主流 VLM 在时空物理推理中的表现进行了诊断分析，然后探讨微调是否可以弥补它们的不足并改善泛化能力。

我们的主要贡献是：(1) 我们对 VLM 在空间物理推理中的表现不佳进行了全面的诊断分析。超越简单的准确性指标，我们识别出其糟糕表现的根本原因：系统

性的人类偏见以及缺乏深入推理。重要的是，我们的分析揭示了推理质量，而不仅仅是推理过程的存在，是成功的决定性因素。(2) 我们系统地研究了微调的有效性和泛化能力。我们表明，监督微调 (SFT) 和强化学习 (RL) 的结合可以显著提升 VLM 在空间物理推理任务中的域内性能，且所得模型甚至超越了领先的专有模型。另一方面，其新的物理情境中促进泛化的能力仍然有限。这个发现突出了当前范式在超越模式匹配以灌输稳健、可推广的物理原则方面的挑战。

最近，为了增强 VLMs 的推理能力的努力集中在专门领域。在多模态数学推理中，像两阶段强化学习 (Peng et al. 2025) 和在线过滤 (Meng et al. 2025) 等方法取得了显著成功。同时，通过大规模视频或 3D 数据集和量身定制的训练方案，空间推理得到了推进，以改善空间布局的理解 (Ouyang et al. 2025; Daxberger et al. 2025)。

物理推理 对于物理推理的研究主要有两个方向：评估和增强。早期的评估基准使用物理模拟器来测试直觉的动力学和静态学 (例如，IntPhys (Riochet et al. 2020), ShapeStacks (Groth et al. 2018), Physion (Bear et al. 2021))，而最近的基准评估符号知识 (Zhang et al. 2025) 或现实世界的物理理解 (Chow et al. 2025)。相比之下，我们的工作针对一个更有结构性和更细化的物理任务：静态稳定性分析。增强方面的努力或者通过集成外部模块进行上下文场景建模 (Balazadeh et al. 2025)，或者追求对一般物理常识的端到端学习 (NVIDIA et al. 2025)，但这些可能并不优先考虑细粒度的物理规律。

2 VLM 的推理行为建模

我们分析了 VLM 在一个原型空间-物理推理任务中的行为：使用 ShapeStacks 基准 (Groth 等, 2018) 来判断堆叠物体的稳定性，该基准由 MuJoCo 引擎 (Todorov, Erez, and Tassa 2012) 生成。这个基准的特征是具有不同高度 (2-6) 和难度级别的堆叠物体，要求模型通过结合堆叠系统的空间信息与质心标准来进行推理。模型基于多个视角将场景分类为平衡 (‘True’) 或不平衡/将会倒塌 (‘False’)。我们对一个包含 888 个样本的测试子集进行分析，研究模型性能、推理策略、错误类型和固有偏差。关于数据集的更多细节，详见附录 A。

[†]Corresponding author: zhangqiaosheng@pjlab.org.cn, shaowenqi@pjlab.org.cn

Model	Accuracy	Difficulty Bias		Height Bias					Duplicated Height Bias		
		Easy	Hard	2	3	4	5	6	2	4	6
InternVL3-8B	0.542	0.794	0.385	0.938	0.805	0.472	0.221	0.249	0.996	0.700	0.252
InternVL3-14B	0.547	0.711	0.405	0.984	0.819	0.505	0.013	-0.162	0.988	0.622	-0.354
InternVL3-38B	0.547	0.997	0.967	0.999	0.993	0.978	0.942	0.868	1.000	0.939	0.804
InternVL3-78B	0.575	0.430	-0.311	0.764	0.264	-0.144	-0.317	-0.421	1.000	0.798	0.076
Qwen2.5VL-7B	0.522	0.218	-0.115	0.059	0.281	-0.107	0.118	-0.094	-0.032	0.038	-0.179
Qwen2.5VL-32B	0.546	0.927	0.335	0.938	0.692	0.692	0.573	0.317	0.258	0.514	-0.036
Qwen2.5VL-72B	0.547	0.967	0.596	0.724	0.781	0.933	0.928	0.715	0.018	1.0	0.722
Gemma3-12B-it	0.507	-0.698	-0.714	-0.674	-0.681	-0.713	-0.735	-0.736	-0.653	-0.695	-0.748
Gemma3-27B-it	0.550	-0.533	-0.663	-0.297	-0.582	-0.640	-0.705	-0.733	-0.145	-0.524	-0.736
GPT-4o	0.560	-	-	-	-	-	-	-	-	-	-
o4-mini	0.593	-	-	-	-	-	-	-	-	-	-
o3	0.641	-	-	-	-	-	-	-	-	-	-
Human Expert	0.779	-	-	-	-	-	-	-	-	-	-

Table 1: 准确性和行为建模参数。 T_{pref} 用于量化模型的偏差。正的 T_{pref} 表示模型倾向于回答为真，而负分表示偏向于假。

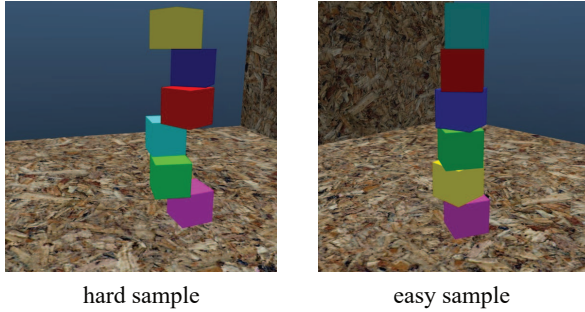


Figure 1: ShapeStacks 基准中的简单和困难样本。虽然展示的两个例子都处于平衡状态，困难样本在不同层之间表现出明显的错位，容易导致误判。

2.1 评估

我们评估了 9 个主流开源 VLMs ($\geq 7B$)，包括它们指令调整过的变体，但不包括 InternVL 系列¹，并以几个商业模型（例如，GPT-4o、o3、o4-mini）作为参考。对于开源模型，我们评估三次并报告平均值；对于商业模型，我们仅评估一次。超参数设置在附录 B.1 中提供。如表 1 所示，所有开源模型的准确率均未超过 0.6，即使是领先的商业模型 o3 也仅达到 0.641，远低于人类专家表现的 0.779。人类评估的详情在附录 B.2 中提供。

此外，我们观察到准确率与 VLMs 语言模型（不包括视觉编码器）的对数尺寸之间存在微弱但统计显著的正相关，如图 2 左侧所示。线性回归模型证实了这一趋势，得出了一个固定效应斜率为 0.0352 (95 % 置信区间 [0.0031, 0.0673])，其中 $p = 0.0360$ 。这表明，对于时空物理推理任务，扩展定律仍然成立，并且性能与语言模型的大小更密切相关，因为不同总体规模的 VLMs 可能共享相同大小的视觉编码器。完整的细节在附录 B.4 中提供。

为了理解这些表现不佳的根本原因，我们对他们的思

¹InternVL3 Instruct 模型未用混合偏好优化 (Zhu et al. 2025) 进行训练

维链 (CoT) 响应进行手动分析，以描述其逻辑步骤，识别常见错误来源，并检查高级认知行为的作用。

通过对模型响应的人工分析，我们在模型的 CoT 中识别出一种逐步的行为模式：(1) 任务识别：模型首先解释提示中呈现的任务定义。(2) 视觉分析：然后，模型分析图像中的视觉信息作为推理的基础。(3) 物理推理：基于任务目标和视觉输入，模型应用相关物理原则进行推理。(4) 结论：最终模型综合以上信息以得出最终决策。值得注意的是，模型常采用一种排除过程：检查是否有任何对象违反稳定性条件，并仅在未发现违反条件时得出稳定性结论。

主要三种错误原因 我们识别出三种主要的失败来源：(1) 视觉感知错误：尽管模型很少在场景中误判物体的数量、颜色或大致位置，但在识别细粒度空间关系（如层错位）时，经常出现错误。这表明模型仅能识别相对显著的视觉线索。(2) 物理推理错误：模型应用了不正确的物理原理。例如，在估计重心时，将水平放置的圆柱体视为竖直放置。(3) 因果推理错误：这些错误出现在模型从通常正确的中间推论中得出错误结论时。一种常见的失败模式是模型将视觉静止视为稳定的证据（即，“系统没有移动，所以它一定是稳定的”），尽管提示明确指出图像显示的是一个静态时刻。

在这些错误中，视觉感知错误是最关键的：如我们对 CoT 的早期分析所揭示的，错误的视觉输入不可避免地会导致后续错误的推理和预测。因果推理错误在很大程度上与模型的逻辑先验相关，而不是空间感知或物理先验。

有限实用性的丰富高级认知行为 受到之前研究 (Gandhi et al. 2025) 的启发，该研究指出高层次的认知行为是符号推理表现的关键，我们调查这些相同的行为是否也能提升视觉语言模型在空间物理推理中的表现。待考虑的四种认知行为是：验证：检查或比较生成的结果以确认其正确性。回溯：放弃失败的推理路径并明确尝试新的策略。子目标设定：将复杂问题分解成一系列较小和较易管理的子问题。逆向推理：从期望目标逆向推理以推断实现它所需的前提条件。

基于此，我们研究这些行为的出现是否与我们任务的

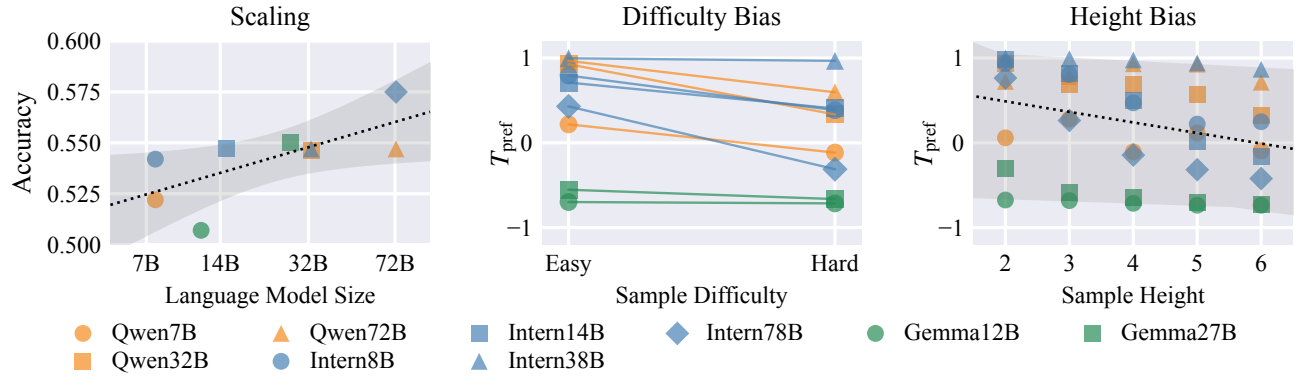


Figure 2: 左图：模型准确性与语言模型组件的对数大小呈正相关。中图：模型偏差 (T_{pref}) 在困难样本上趋向于预测“错误”。右图：随着栈高度的增加，模型越来越偏向于“错误”。在左图中，趋势线和 95 % CI 灰色阴影使用线性回归拟合。在右图中，每个模型的趋势线阴影和整体趋势的虚线使用线性混合效应模型拟合。

正确性相关。我们使用 o4-mini 模型对 InternVL 系列的响应进行标注，比较正确和错误响应中这些行为的分布。更多细节见附录 C。

如图 3 所示，对于所有 InternVL 模型，认知行为的分布在正确和错误的响应之间没有显著差异。这表明推理过程的质量和内容的存在比仅仅存在的认知行为模式更重要。该模型只是在执行表层推理，而不是真正的深入推理。

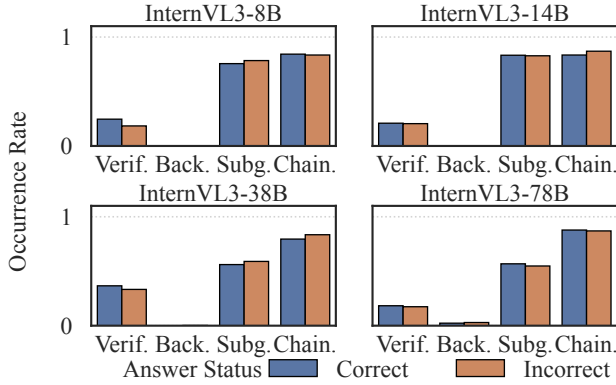


Figure 3: InternVL 模型系列中正确和错误响应中的高级推理行为比例没有显著差异。

2.2 具有类人偏好的偏见

接下来，我们研究是否在人类生成的数据上预训练的 VLMs 在空间物理推理中表现出类似人类的偏见。具体而言，我们考虑两个先验：困难先验，其中较大的物体间位移被视为不稳定，以及高度先验，其中较高的堆叠被认为较不稳定。

为了量化这一点，我们分析了模型预测的混淆矩阵并定义了偏好得分 T_{pref} ：

$$T_{\text{pref}} := \tanh(\psi), \quad \psi := \frac{\text{Recall} - \text{Specificity}}{\text{Specificity}} \quad (1)$$

，其中 $\text{Recall} := \frac{\text{TP}}{\text{TP} + \text{FN}}$ 和 $\text{Specificity} := \frac{\text{TN}}{\text{TN} + \text{FP}}$ 分别

表示正确预测的 True 和 False 样本的比例。指标 T_{pref} 反映了模型的倾向：正分数表示偏向“True”，而负分数则表示偏向“False”。

对于物体间位移的先验，我们分析了模型对简单和困难样本的偏好分数 T_{pref} 。如图 1 所示，困难样本在稳定和不稳定的情况下都涉及较大的不对齐，这可能会误导模型。我们假设如果模型具有类人偏差，它将为困难样本分配较低的 T_{pref} 分数，错误地将平衡的堆判断为不平衡。

我们的假设得到了先前评估中所用响应数据的支持。如表 1 的“困难偏差”列以及图 2 的中间图所示，与简单样本相比，所有模型在困难样本上得出的 T_{pref} 分数较低。这证实了困难样本中的视觉位移特征会对模型产生误导。一致的偏差模式表明，这些模型表现出类似于人类的先验：明显的物体间位移与不稳定性相关联。

此外，模型的“基于消除”的推理过程也可能导致这种偏差：在难处理样本中，较大的层间错位通常被识别为不稳定因素，从而直接导致模型得出结构不平衡的结论。

为了研究高度优先级，我们首先分析堆栈高度上的 T_{pref} 分数。如表 1 的高度偏差列和图 2 的右侧面板所示，模型随着高度的增加而倾向于呈现较低的 T_{pref} 分数。线性混合效应模型分析进一步确认了这一趋势，得出了固定效应斜率 -0.1244 (95% 置信区间: $[-0.2025, -0.0463]$, $p = 0.0018$)，表明统计上的显著性。

为了测试这种偏差的鲁棒性，我们创建了重复样本：基于 $h = 2$ 立方体样本，我们通过垂直堆叠相同立方体而不改变水平位置生成 $h = 4$ 和 $h = 6$ 版本 (图 4a)。虽然这些样本在高度上有所不同，但它们的机械结构基本相同。

在这些重复样本中，大多数模型——尤其是 InternVL 和 Gemma 系列——随着物体高度增加在 T_{pref} 上表现出下降趋势 (图 4b, 表 1)。这表明存在持续的高度相关偏差。线性混合效应模型得到一个固定效应斜率为 -0.1008 (95 % 置信区间: $[-0.1965, -0.0051]$, $p = 0.0389$)，表明存在统计学显著的负相关，因此存在普遍趋向于此偏差的趋势。然而，Qwen 系列的偏差在这些样本上消失，表明其先验是脆弱且不系统的。有关拟合的详细信息见附录 B.4。

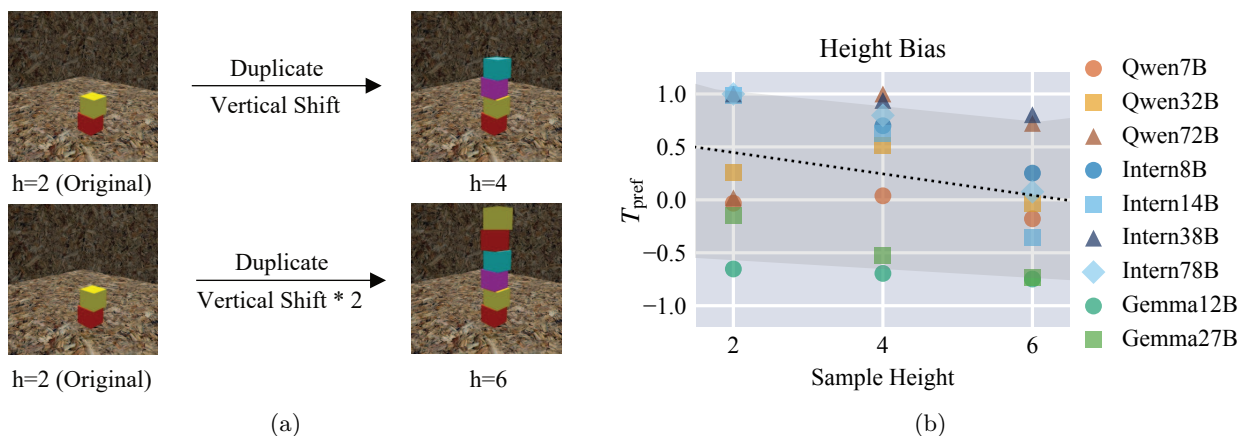


Figure 4: (a) 通过复制和平移生成具有高度相似机械结构的更多样本。(b) 在复制样本上对大多数模型的偏差仍然存在。用线性混合效应模型拟合显示出一个随着高度增加而普遍下降的趋势。尽管某些模型在更高高度下显示出预测“真实”的增加趋势，这些偏差并不影响整体趋势。灰色阴影是每个模型趋势线的包络线。

这两项实验证实，人类般的高度先验是一个常见的 VLM 特征，虽然其强度在不同的模型家族中有所不同。我们推测，这种高度偏差受到现实世界数据分布的影响，其中较高的物体通常比较低的物体更不稳定。在此类数据上训练的模型可能会继承这种偏差。此外，不同 VLM 中这种偏差的不同强度可能归因于它们不同的训练方法。

为研究先验知识转移，我们评估了两种专门化模型：具有强空间先验的 SpaceR (Ouyang et al. 2025) 和具有强物理先验的 MM-Eureka-Qwen-7B (Meng et al. 2025)。我们将它们的性能与其基础模型 Qwen2.5-VL-7B 在 ShapeStacks 上的表现进行比较。

如表 2 所示，尽管这些模型具有很强的先验，它们几乎没有表现出精度的提升，表明性能没有提高。这表明，仅在单一领域进行微调对于组合推理是不够的。

Model	Accuracy
Qwen2.5-VL-7B	0.522
MM-Eureka-Qwen-7B	0.521
SpaceR	0.522

Table 2: 先验知识迁移：仅仅拥有空间或物理推理能力很难提高在时空物理推理任务中的表现。

在我们的诊断分析的基础上，本节探讨微调是否可以弥补已识别的视觉语言模型的限制。我们展示了一个两阶段训练流程——首先进行监督微调 (SFT)，然后进行强化学习 (RL)——在 Qwen2.5-VL-7B 上显著提高其在 ShapeStacks 上的表现，并与领先的专有模型相比实现了最先进的性能。此外，我们测试改进后的模型对跨维度、动态和高度的新物理场景的泛化能力。我们发现该模型表现出一定程度的泛化，但当数据分布偏离训练域时，其性能会下降。

2.3 领域内性能提升

实验设置 我们的微调数据是 ShapeStacks 训练集 (13,618 个样本)。我们使用 o4-mini 模型对 3,000 个随机选择的样本 (种子 = 0) 进行 CoT 样式推理响应的

蒸馏，构建 SFT 数据集。剩余的样本则被格式化用于具有结果监督的强化学习。

我们使用 TRL 框架 (von Werra et al. 2020)，结合 GRPO 算法对 Qwen2.5-VL-7B 进行 SFT 和 veRL 的微调。在 SFT 阶段，我们使用学习速率为 $1e-4$ 的 LoRA 进行训练，总共 40 个周期，并采用余弦调度器进行 5% 的预热。在 RL 阶段，我们进行 8 个周期的全参数微调。学习速率设定为 $2e-6$ (使用余弦调度器和 10% 的预热)，批量大小为 1024。对于每个提示，我们展开 6 个响应，总奖励是格式得分和正确性得分的加权和。我们要求模型在不同的特殊标记之间输出思维链 (CoT) 和最终答案 (在 ‘<think>’ 和 ‘</think>’ 之间放置 CoT，并在 ‘<answer>’ 和 ‘</answer>’ 之间放置最终答案)，这被用作二元格式奖励 (0 或 1)。模型最终答案的正确性也用作二元答案奖励。在总奖励中，格式奖励占 10%，答案奖励占 90%。最终得到的模型被命名为 Qwen-SFT-RL。详细信息见附录 D.1。

我们在上一节中使用的 ShapeStacks 测试集上评估了 Qwen-SFT-RL 模型。结果如表 3 所示。如表所示，SFT+RL 显著提升了模型的性能，相比 Instruct 模型取得了 47.7% 的提升，并且性能超越了领先的专有模型 o3。这表明，SFT+RL 微调范式可以有效提升 VLMs 在叠放系统稳定性任务上的性能。

我们进行了消融研究以理解 SFT 的作用。我们通过在 SFT 阶段的情况下直接将 RL 应用于基础模型来创建 Qwen-RL 模型。我们还评估了仅使用 SFT 进行微调的 Qwen-SFT 模型。超参数与用于训练 Qwen-SFT-RL 的超参数相同。

如表 3 所示，Qwen-RL 在准确性上有显著的 42.7% 的提升，但仍低于 Qwen-SFT-RL 的 47.7%，这表明仅靠 RL 是不理想的。虽然 Qwen-SFT 显示出很小的改善 (0.4%)，但训练动态分析 (图 5) 揭示了 SFT 初始化的好处。Qwen-RL 表现出过拟合 (验证奖励在步骤 20 左右急剧下降)，未能保持响应长度，并保留了基础模型的因果推理错误率，而 Qwen-SFT-RL 则无错误。

这表明仅依靠结果监督的强化学习不足以指导推理。SFT 扮演了两个关键角色：(1) 它为强化学习提供了一

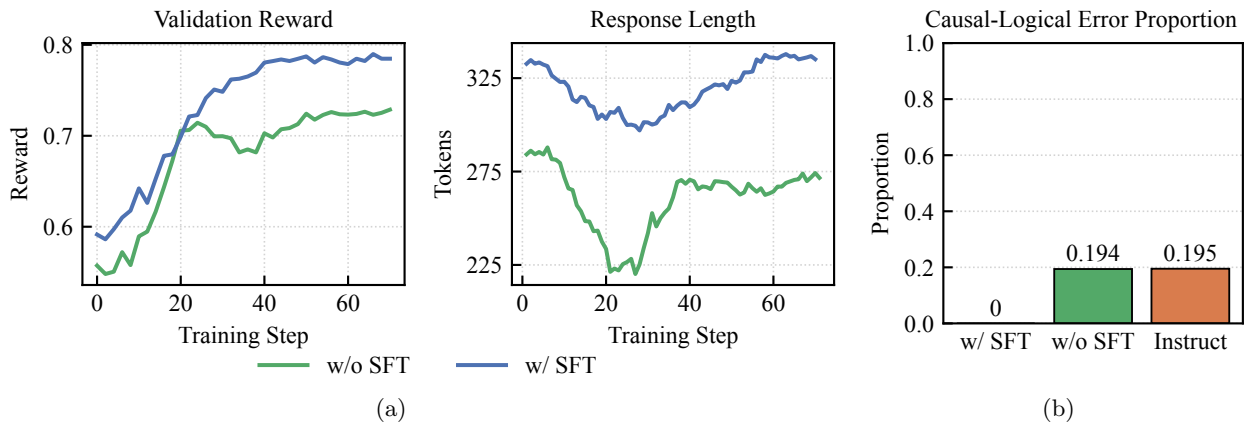


Figure 5: SFT 作为 RL 初始化的好处。我们将使用 SFT 初始化的 RL 微调 (w/ SFT) 与来自 Instruct 模型的 RL 微调 (w/o SFT) 进行比较。(a) SFT 初始化在训练过程中导致更稳定的验证奖励并保持更长的响应。(b) SFT 还消除了没有它时存在的因果逻辑错误。

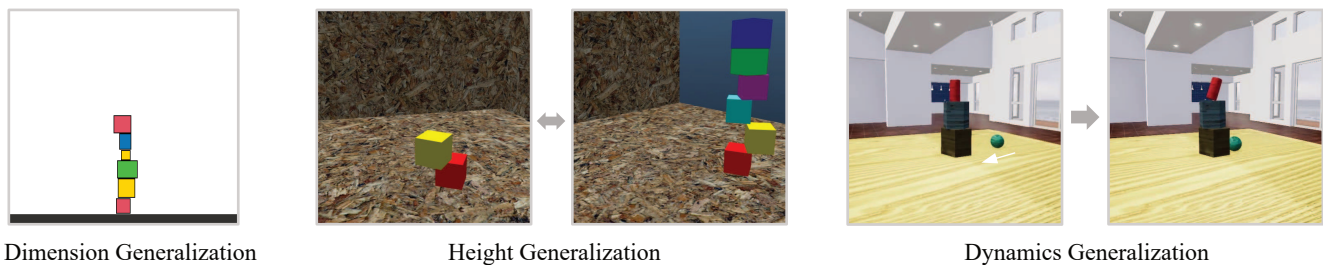


Figure 6: 三个基于物理的泛化测试的示例。维度泛化测试在 2D 和 3D 数据之间进行转移。高度泛化测试在低堆和高堆之间进行转移。动力学泛化测试从静态稳定性转移到涉及外部力的动态场景 (Physion)。

个稳健的初始化，提高了训练稳定性和最终表现；(2) 其 CoT 数据约束模型学习正确的推理逻辑。

2.4 泛化能力测试

我们研究模型是否可以通过后训练来推广特定的时空物理推理模式。具体来说，我们探索了基于不同特征模式的三种类型的推广。

- 维度泛化：我们评估模型在不同维度上下文间转换推理的能力。具体来说，我们测试一个在三维数据 (ShapeStacks) 上训练的模型是否能泛化到一个定制的二维数据集，反之亦然。
- 动态泛化：我们评估 Qwen-SFT-RL (在 ShapeStacks 上训练的) 是否能够将堆叠系统的时空物理知识泛化到动态物理环境中。具体来说，我们使用 Physion (Bear et al. 2021) 数据集的支持场景作为评估基准。
- 高度泛化：我们研究模型是否能够学习并泛化与高度相关的抽象原则，而不仅仅是记住特定的高度模式。我们通过分别对两个独立的模型进行微调，使其只接触低堆和高堆数据，然后评估它们在未见过的高度类别上的表现来实现这一点。

三种类型的推广在图 6 中进行了说明。在所有的推广实验中，微调的模型基于 Qwen2.5-VL-7B。除了与训练数据相关的差异，如微调集的内容，最终检查点的步骤和

时代数量，所有其他超参数与用于 Qwen-SFT-RL 的相同。详细信息在附录 D.1 中提供。

尽管三维分析涉及额外的空间维度，但分析质心投影的核心物理原理仍然相同。我们研究一个在一维空间数据上训练得到的模型是否能够将其推理能力推广到不同维度。

对于 2D 堆叠系统数据，我们使用 Box2D 物理引擎 (Catto. 2009) 生成样本，塔的高度范围从 3 到 6 (因为 2D 高度为 2 的塔过于简单)，并将此数据集称为 2d-towers。完整的数据集被分为训练集和测试集。有关数据生成的更多详细信息，我们建议读者参阅附录 D.2。

我们首先在 2d-towers 测试集上评估 Qwen-SFT-RL，以分析其从 3D 到 2D 的泛化能力。评估结果 (表 3) 显示了 29.0 % 的增益，并揭示了泛化能力的显著提升。

为了测试从 2D 到 3D 的逆向转换能力，我们对 2d-towers 训练集进行模型微调。得到的模型被标记为 Qwen-2d。我们在 ShapeStacks 和 2d-towers 的测试集上评估 Qwen-2d。如表 3 所示，Qwen-2d 在域内准确性上提高了 80 %，在推广到 ShapeStacks 时取得了适度的 17.4 % 增益，这低于 Qwen-SFT-RL 在 2d-towers 上实现的 29.0 % 改进。

这表明虽然维度泛化是可行的，但从 3D 到 2D 的泛化更为有效。我们假设这是因为 3D 样本提供了更丰富的空间信息，使得一个在 3D 环境中熟练的模型更容易处理更简单的 2D 情况。我们接下来探索从静态到动态

Models	Test Samples				
	ShapeStacks	ShapeStacks high	ShapeStacks low	2d-towers	Physion Support
Qwen2.5VL-7B	0.522	0.580	0.531	0.510	0.504
Qwen-SFT-RL (ShapeStacks)	0.771 / 47.7 %	0.781 / 34.7 %	0.799 / 50.5 %	0.658 / 29.0 %	0.509 / 1.0 %
Qwen-high (ShapeStacks high)	0.606 / 16.1 %	0.678 / 16.9 %	0.573 / 7.9 %	0.670 / 31.4 %	0.493 / -2.2 %
Qwen-low (ShapeStacks low)	0.677 / 29.7 %	0.683 / 17.8 %	0.751 / 41.4 %	0.661 / 29.6 %	0.507 / 0.6 %
Qwen-2d (2d-towers)	0.613 / 17.4 %	0.674 / 16.2 %	0.609 / 14.7 %	0.918 / 80.0 %	0.522 / 3.6 %
Qwen-RL (ShapeStacks)	0.745 / 42.7 %	0.766 / 32.1 %	0.792 / 49.2 %	0.520 / 2.0 %	0.5 / -0.8 %
Qwen-SFT (ShapeStacks)	0.524 / 0.4 %	0.532 / -8.3 %	0.521 / -1.9 %	0.590 / 15.7 %	0.493 / -2.2 %

Table 3: 泛化能力分析中的指标。在“模型”列中，微调所用的数据集在模型名称后用括号表示。微调模型的评估结果以以下格式呈现：“准确率/相对于微调前指令模型的改进”。加粗的数字表示域内评估结果。

时空物理推理的泛化。具体来说，我们在 Physion 基准的支持场景中评估 Qwen-SFT-RL（在 ShapeStacks 数据集上训练的），其中 Physion 是一个包含各种物理场景的动态经典力学基准。在支持场景中，一个堆叠的几何系统受到外力作用，要求模型预测结构是否保持稳定或坍塌，如果坍塌则预测坍塌的过程。这个实验旨在评估一个已经学习了稳健的静态物理稳定性推理的模型是否可以将这种理解迁移到动态场景中。

Qwen-SFT-RL 的评估结果如表 3 所示。显然，精通静态物理推理的模型在推广到动态样本时仍然存在困难，仅显示出 1 % 的提升。这表明模型在将其对堆叠系统的理解转移到动态领域时遇到了困难，尽管动态场景中包含潜在的静态平衡原理。一个可能的解释是静态和动态样本之间的巨大视觉差异，这可能阻碍了模型将其学习的推理模式与新的视觉上下文相结合的能力。

我们研究高度泛化，其中核心推理任务——估计质心——在堆叠高度变化的情况下仍保持不变。我们将原始训练集分为 ShapeStacks-low（高度为 2-3）和 ShapeStacks-high（高度为 4-6）。由此得出的模型，在高低数据上进行微调，分别被称为 Qwen-high 和 Qwen-low。为了进行评估，我们将相同的基于高度的分类应用于原始测试集，并从低和高子集中各随机抽取 1,000 个例子。

如表 3 所示，Qwen-low 在领域内表现更强，并且比 Qwen-high 具有更好的高度泛化能力。从高到低的泛化仅提升了 7.9 %，而从低到高的泛化则显示了 17.8 % 的提高，甚至超过了 Qwen-high 在域内的表现（提高了 16.9 %）。在混合高度的 ShapeStacks 集合上，Qwen-low 实现了 29.7 % 的性能提升，明显高于 Qwen-high 的 16.1 %。

这些发现表明，VLMs 在两个方向上都具备高度泛化能力，但从较低样本到较高样本的效果更为明显。我们将这种不对称性归因于高叠样本的视觉和物理复杂性增加，这可能导致训练过程中产生具有挑战性且信息不足的梯度。相比之下，较简单的低叠样本有助于学习正确的推理策略，这导致更强的性能和更好的泛化能力。

域转移导致的性能下降 我们的最终分析评估了模型在所有基准场景下的表现。结果如表 3 所示，其中域内样本上的性能提升以粗体显示。我们观察到一个明显的趋势：经过微调的 VLM 能够对密切相关的领域进行有意义的泛化，但这种能力随着测试样本和训练样本之间的领域差距扩大而减弱。

具体来说，我们观察到：(1) Qwen-SFT-RL 在 ShapeS-

tacks 高度变化上的显著提升在 2D 任务上有所减弱，在 Physion 任务上几乎消失。(2) Qwen-low 在域内的提升在域外的 ShapeStacks 高度和 2D 样本上明显下降，在 Physion 上几乎没有改进，尽管它仍然超越了基础模型。(3) Qwen-high 在 2D 样本上明显提升，但在其他 ShapeStacks 高度变化上的提升减少，在 Physion 的支持上完全消失。(4) Qwen-2d 尽管在域内 2D 提升强劲，但在任何更高维度的 ShapeStacks 样本上改善有限，在 Physion 上仅有边际提升。

此外，比较结果表明，Qwen-SFT-RL 在所有泛化测试集上始终优于 Qwen-RL，这表明 SFT 也有助于提高微调模型的泛化能力。

我们的综合分析揭示了当前开源 VLMs 在表面能力与实际物理理解之间存在的断层。我们首先识别了它们的基本局限性，包括性能较弱、人类般的认知偏见，以及缺乏深度推理，这表明质量比形式更重要。然后我们证明，虽然主流的微调方法可以在特定领域内显著提升性能，这种改进在很大程度上是表面的。这些模型在新颖物理场景中的泛化能力有限，尤其是在领域差异增加时，表明它们依赖于统计捷径而非强健的物理原理。这暴露了当前范式的一个根本限制：它们在模式匹配上表现出色，但未能灌输建立世界模型所需的可泛化理解。我们的研究强烈表明，未来的发展道路不在于仅仅扩大现有方法的规模，而在于开发新的方法。未来的工作应优先考虑如利用具有物理基础的模拟数据和设计新颖的训练范式等策略，明确鼓励因果性和可泛化物理法则的学习。

References

- Assran, M.; Bardes, A.; Fan, D.; Garrido, Q.; Howes, R.; Mojtaba; Komeili; Muckley, M.; Rizvi, A.; Roberts, C.; Sinha, K.; Zholus, A.; Arnaud, S.; Gejji, A.; Martin, A.; Hogan, F. R.; Dugas, D.; Bojanowski, P.; Khalidov, V.; Labatut, P.; Massa, F.; Szafraniec, M.; Krishnakumar, K.; Li, Y.; Ma, X.; Chandar, S.; Meier, F.; LeCun, Y.; Rabbat, M.; and Ballas, N. 2025. V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning. arXiv:2506.09985.
- Balazadeh, V.; Ataei, M.; Cheong, H.; Khasahmadi, A. H.; and Krishnan, R. G. 2025. Physics Context Builders: A Modular Framework for Physical Reasoning in Vision-Language Models. arXiv:2412.08619.
- Bear, D.; Wang, E.; Mrowca, D.; Binder, F.; Tung, H.-

- Y.; RT, P.; Holdaway, C.; Tao, S.; Smith, K.; Sun, F.-Y.; Li, F.-F.; Kanwisher, N.; Tenenbaum, J.; Yamins, D.; and Fan, J. 2021. Physion: Evaluating Physical Prediction from Vision in Humans and Machines. In Vanschoren, J.; and Yeung, S., eds., *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.
- Catto., E. 2009. Box2D: Box2D is a 2D physics engine for games. <https://github.com/erincatto/box2d>. Accessed: 2025-5-30.
- Chow, W.; Mao, J.; Li, B.; Seita, D.; Guizilini, V. C.; and Wang, Y. 2025. PhysBench: Benchmarking and Enhancing Vision-Language Models for Physical World Understanding. In *The Thirteenth International Conference on Learning Representations*.
- Dai, S.; Yan, Y.; Su, J.; Zihao, D.; Gao, Y.; Hei, Y.; Li, J.; Zhang, J.; Tao, S.; Gao, Z.; and Hu, X. 2025. PhysicsArena: The First Multimodal Physics Reasoning Benchmark Exploring Variable, Process, and Solution Dimensions. arXiv:2505.15472.
- Daxberger, E.; Wenzel, N.; Griffiths, D.; Gang, H.; Lazarow, J.; Kohavi, G.; Kang, K.; Eichner, M.; Yang, Y.; Dehghan, A.; and Grasch, P. 2025. MM-Spatial: Exploring 3D Spatial Understanding in Multimodal LLMs. arXiv:2503.13111.
- Gandhi, K.; Chakravarthy, A.; Singh, A.; Lile, N.; and Goodman, N. D. 2025. Cognitive Behaviors that Enable Self-Improving Reasoners, or, Four Habits of Highly Effective STaRs. arXiv:2503.01307.
- Groth, O.; Fuchs, F. B.; Posner, I.; and Vedaldi, A. 2018. ShapeStacks: Learning Vision-Based Physical Intuition for Generalised Object Stacking. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part I*, 724–739. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-030-01245-8.
- Liao, Z.; Xie, Q.; Zhang, Y.; Kong, Z.; Lu, H.; Yang, Z.; and Deng, Z. 2025. Improved Visual-Spatial Reasoning via R1-Zero-Like Training. arXiv:2504.00883.
- Meng, F.; Du, L.; Liu, Z.; Zhou, Z.; Lu, Q.; Fu, D.; Han, T.; Shi, B.; Wang, W.; He, J.; Zhang, K.; Luo, P.; Qiao, Y.; Zhang, Q.; and Shao, W. 2025. MM-Eureka: Exploring the Frontiers of Multimodal Reasoning with Rule-based Reinforcement Learning. arXiv:2503.07365.
- NVIDIA; ; Azzolini, A.; Bai, J.; Brandon, H.; Cao, J.; Chattopadhyay, P.; Chen, H.; Chu, J.; Cui, Y.; Diamond, J.; Ding, Y.; Feng, L.; Ferroni, F.; Govindaraju, R.; Gu, J.; Gururani, S.; Hanafi, I. E.; Hao, Z.; Huffman, J.; Jin, J.; Johnson, B.; Khan, R.; Kurian, G.; Lantz, E.; Lee, N.; Li, Z.; Li, X.; Liao, M.; Lin, T.-Y.; Lin, Y.-C.; Liu, M.-Y.; Lu, X.; Luo, A.; Mathau, A.; Ni, Y.; Pavao, L.; Ping, W.; Romero, D. W.; Smelyanskiy, M.; Song, S.; Tchapmi, L.; Wang, A. Z.; Wang, B.; Wang, H.; Wei, F.; Xu, J.; Xu, Y.; Yang, D.; Yang, X.; Yang, Z.; Zhang, J.; Zeng, X.; and Zhang, Z. 2025. Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning. arXiv:2503.15558.
- Ouyang, K.; Liu, Y.; Wu, H.; Liu, Y.; Zhou, H.; Zhou, J.; Meng, F.; and Sun, X. 2025. SpaceR: Reinforcing MLLMs in Video Spatial Reasoning. arXiv:2504.01805.
- Peng, Y.; Zhang, G.; Zhang, M.; You, Z.; Liu, J.; Zhu, Q.; Yang, K.; Xu, X.; Geng, X.; and Yang, X. 2025. LMM-R1: Empowering 3B LMMs with Strong Reasoning Abilities Through Two-Stage Rule-Based RL. arXiv:2503.07536.
- Riochet, R.; Castro, M. Y.; Bernard, M.; Lerer, A.; Ferguson, R.; Izard, V.; and Dupoux, E. 2020. IntPhys: A Framework and Benchmark for Visual Intuitive Physics Reasoning. arXiv:1803.07616.
- Shen, H.; Wu, T.; Han, Q.; Hsieh, Y.; Wang, J.; Zhang, Y.; Cheng, Y.; Hao, Z.; Ni, Y.; Wang, X.; Wan, Z.; Zhang, K.; Xu, W.; Xiong, J.; Luo, P.; Chen, W.; Tao, C.; Mao, Z.; and Wong, N. 2025. PhyX: Does Your Model Have the “Wits” for Physical Reasoning? arXiv:2505.15929.
- Todorov, E.; Erez, T.; and Tassa, Y. 2012. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, 5026–5033. IEEE.
- von Werra, L.; Belkada, Y.; Tunstall, L.; Beeching, E.; Thrush, T.; Lambert, N.; Huang, S.; Rasul, K.; and Gallouédec, Q. 2020. TRL: Transformer Reinforcement Learning. <https://github.com/huggingface/trl>.
- Xiang, K.; Li, H.; Zhang, T. J.; Huang, Y.; Liu, Z.; Qu, P.; He, J.; Chen, J.; Yuan, Y.-J.; Han, J.; Xu, H.; Li, H.; Sachan, M.; and Liang, X. 2025. SeePhys: Does Seeing Help Thinking? – Benchmarking Vision-Based Physics Reasoning. arXiv:2505.19099.
- Zhang, X.; Dong, Y.; Wu, Y.; Huang, J.; Jia, C.; Fernando, B.; Shou, M. Z.; Zhang, L.; and Liu, J. 2025. PhysReason: A Comprehensive Benchmark towards Physics-Based Reasoning. arXiv:2502.12054.
- Zheng, S.; Cheng, Q.; Yao, J.; Wu, M.; He, H.; Ding, N.; Cheng, Y.; Hu, S.; Bai, L.; Zhou, D.; Cui, G.; and Ye, P. 2025. Scaling Physical Reasoning with the PHYSICS Dataset. arXiv:2506.00022.
- Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; Gao, Z.; Cui, E.; Wang, X.; Cao, Y.; Liu, Y.; Wei, X.; Zhang, H.; Wang, H.; Xu, W.; Li, H.; Wang, J.; Deng, N.; Li, S.; He, Y.; Jiang, T.; Luo, J.; Wang, Y.; He, C.; Shi, B.; Zhang, X.; Shao, W.; He, J.; Xiong, Y.; Qu, W.; Sun, P.; Jiao, P.; Lv, H.; Wu, L.; Zhang, K.; Deng, H.; Ge, J.; Chen, K.; Wang, L.; Dou, M.; Lu, L.; Zhu, X.; Lu, T.; Lin, D.; Qiao, Y.; Dai, J.; and Wang, W. 2025. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. arXiv:2504.10479.