DIVA-GRPO：通过难度自适应变体优势提升多模态推理能力

Haowen Gao[1], [2†]、Zhenyu Zhang[3†]、Liang Pang[1*]、Fangda Guo[1],Hongjian Dou[3],Guannan Lv[3], Shaoguo Liu[3*]、Tingting Gao[3],Huawei Shen[1],Xueqi Cheng[1]

[1] 人工智能安全国家重点实验室，计算技术研究所，中国科学院，北京，中国[2] 中国科学院大学，北京，中国[3] 快手科技，北京，中国

gaohaowen23s@ict.ac.cn, zhangzhenyu06@kuaishou.com
pangliang@ict.ac.cn

arXiv:2603.01106v1 [cs.AI] 1 Mar 2026

摘要

强化学习（RL）结合组相对策略优化（GRPO）已成为增强多模态大语言模型（MLLMs）推理能力的广泛采用方法。尽管 GRPO 无需传统评论家模型即可实现长链推理，但它常常受困于稀疏奖励（源于困难问题上正向反馈的稀缺）以及优势消失（当组级奖励对过于简单或过于困难的问题表现出高度一致性时发生）。现有解决方案可分为三类：样本增强与扩展，可能因难度分布控制不佳而加剧优势消失；选择性样本利用，未能充分利用全部数据的价值；以及间接奖励设计，可能因推理与最终结果之间的错位而引入有偏的优化方向。然而，这些方法忽略了一个根本问题：对于给定问题，如何确保回答的组内奖励分布具有足够的方差，从而为每个回答产生清晰的优化信号？ 为解决这些问题，我们提出 DIVA-GRPO，一种难度自适应变体增强优势方法，从全局视角动态调整每个问题的变体难度分布。本方法动态评估问题难度，采样具有适当难度等级的变体，并采用难度加权归一化缩放，在局部与全局（一个问题及其变体）组内计算优势。该设计缓解了奖励稀疏性与优势消失，最大限度减少数据浪费，并提升训练稳定性。在六个推理基准上的大量实验表明，DIVA-GRPO 在训练效率与推理性能方面均优于现有方法。代码可于 https://github.com/Siaaaaaa1/DIVA-GRPO 获取。

1 引言

多模态大语言模型（MLLMs）（Chen 等，2024b；Hurst 等，2024；Laurençon 等，2024；Liu 等，2023；Yin 等，2024；Wu 等，2024b；Alayrac 等，2022；Zhu 等，2023；Li 等，2023）展现出卓越的能力，能够整合文本与视觉信息（Pang 等，2024），以完成复杂推理任务，例如视觉问答（Antol 等，2015；Xiao 等，

[†]贡献相同[*]通讯作者

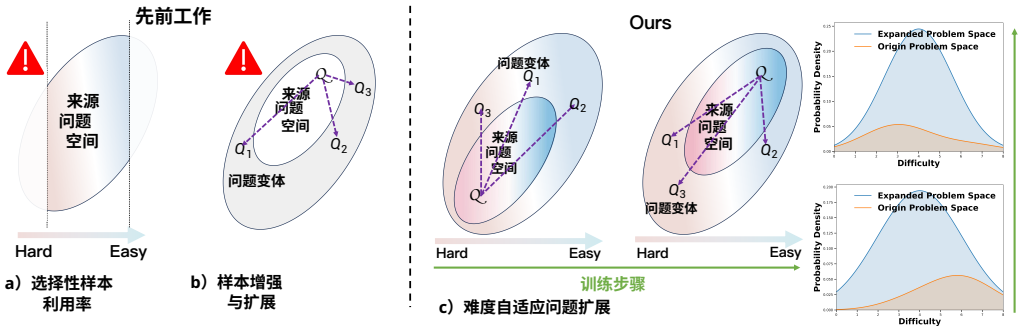


图 1：（a）选择性样本利用仅依赖部分数据，导致数据利用不足。（b）样本增强在缺乏难度感知的情况下扩展数据，导致优势稀疏性进一步加剧。（c）我们的方法根据问题难度自适应地扩展样本空间，确保难度分布稳定。

2024）和多模态逻辑推理（Huang & Chang，2022；Guo 等，2025年）。然而，文本和视觉模态的异质性使长链推理颇具挑战，需要仔细观察并逐步解决问题（Xu 等，2024年）。为应对这些挑战，近期研究探索了多模态思维链方法（Zhou 等，2024年；Chen 等，2024a；Wang 等，2025年b），包括 Video-of-Thought（Fei 等，2024年）、Det-CoT（Wu 等，2024a）、CoI（Meng 等，2023）、Grounded-RL（Sarch 等，2025年），这些方法旨在通过分解问题或利用结构化策略来提升推理能力。与这些进展相辅相成的是，强化学习（RL）已成为进一步增强多模态大语言模型的强大框架：近端策略优化（PPO）（Schulman 等，2017）和直接偏好优化（DPO）（Rafailov 等，2023）广泛用于对齐，而 GRPO（Shao 等，2024）则通过将奖励与组相对优势估计相结合，推动长链推理的发展，显著提升了对齐效果和推理效率。

然而，将 GRPO 直接应用于多模态大语言模型面临着一个关键挑战——优势消失（Huang 等，2025年；Park 等，2025年；Yao 等，2025年；Wang 等，2025年a；Meng 等，2025年；Zhang 等，2025年；Zhang 和 Zuo，2025年）。GRPO 通过比较针对给定问题采样得到的回答来计算相对优势，然后对所有样本的优势进行标准化。当某个问题对当前模型来说过于简单（或过于困难）时，正奖励（或负奖励）过少可能导致优势值过大，从而造成较高的梯度噪声和不稳定的训练。在极端情况下，所有回答都可能完全正确或完全错误，导致相对优势为零。这会减缓学习进程，并削弱模型在不同难度下的推理性能。现有方法尝试从多个方面解决这一问题。一种方法是样本增强与扩展，即通过为困难实例添加提示词，或生成多样化的文本和图像变体来扩大问题空间，不过由于难以控制难度分布，这可能会加剧优势消失（Huang 等，2025年；Park 等，2025年；Yao 等，2025年），并引入隐藏偏差（Gao 等，2025年）。另一种方法是选择性样本利用，即优先使用效果显著的样本，或舍弃学习贡献较低的样本，以提高效率；然而，这可能会限制模型接触复杂问题的机会，或降低数据多样性（Wang 等，2025年a；Meng 等，2025年）。最后，间接奖励设计提供更细粒度的反馈信号，以缓解奖励稀疏性，但这些奖励可能无法与最终任务目标完全一致，进而可能引导模型走向次优方向（Zhang 等，2025年；Zhang 和 Zuo，2025年）。

尽管上述方法能够缓解某些优势消失情形，但它们都忽略了一个关键问题：如图***1(c)***所示，随着训练推进，问题难度持续下降，优势消失不断加剧，模型训练效率稳步下降。 为解决这一问题，我们提出 DIVA-GRPO，这是一种自适应方法，可在保持语义一致性的同时动态调整问题难度和变体分布，并同时整合局部（单个问题）与全局（由原始问题派生的多个变体）奖励计算。在训练过程中，根据模型回答迭代评估问题难度：简单问题通过复杂文本和图像扰动进行增强，中等难度问题通过多样化文本变体进行增强，困难问题则通过“思维步骤”推理变体进行增强，从而确保有效的样本利用并缓解优势消失。为平衡不同难度等级之间的反馈，DIVA-GRPO 采用批次 Z 分数归一化和难度加权缩放，并进一步引入基于奖励范围的重新缩放，以防止微小奖励差异导致优势被夸大。这

实现稳定的策略优化，鼓励探索，并增强长链推理，同时广泛适用于基于 GRPO 的方法，以提升优化效率。

大量实验验证了 DIVA-GRPO 的有效性和效率。我们的模型在六项具有挑战性的基准测试中表现出色，在 7B规模下达到最先进的性能，准确率接近规模大得多的模型和专有商业系统。此外，DIVA-GRPO 显著加速了训练。在使用相同原始数据集合的情况下，它以少于原先一半的步数达到最佳性能，并将端到端墙钟时间加速 1.76× 倍以上，同时在各难度等级下保持稳定且有信息量的优势信号。这些结果凸显了我们的模型在性能和训练效率方面带来的显著提升。

总的来说，DIVA-GRPO 通过结合自适应难度评估、变体生成和稳健的优势估计，克服了传统 GRPO 的局限性。该框架缓解了奖励稀疏性和优势消失问题，同时提升了从视觉问答到复杂逻辑推理等广泛多模态任务中的推理稳定性和性能。我们的核心贡献如下：

- 提出 DIVA-GRPO，基于动态难度评估，利用难度自适应变体生成与优势共享来缓解奖励稀疏性和优势消失。
- 引入局部与全局优势的联合估计，并结合批次 z 分数归一化与难度加权缩放，以提升训练稳定性。
- 提出一种基于奖励范围的优势重新缩放方法，有效防止不合理的优势膨胀并加速收敛。
- 在六个主流多模态推理基准上开展实验，证明了 DIVA-GRPO 的有效性与优越性。

2 挑战：奖励稀疏性与优势消失

我们首先回顾组相对策略优化（GRPO）的关键概念，它们构成了我们难度自适应变体策略的基础。

组相对策略优化（GRPO）。令 $\mathcal{Q} = \{q_1, q_2, \ldots, q_N\}$ 表示训练问题的集合。对于给定的问题 q，模型 π_θ 通过 rollout 生成 k 个候选回答 $\mathcal{Y}_q = \{y_1, y_2, \ldots, y_k\}$。每个回答 y_i 都会获得一个标量奖励 $r(y_i) \in \mathbb{R}$，该奖励由奖励函数确定。奖励函数可以是基于规则的评估器（例如答案正确性），也可以是训练得到的奖励模型。GRPO 通过对奖励进行标准化，计算每个回答在其所属组内的相对优势：

$$A(y_i) = \frac{r(y_i) - \mu_r}{\sigma_r + \epsilon}, \quad \mu_r = \frac{1}{k}\sum_{j=1}^{k} r(y_j), \quad \sigma_r = \sqrt{\frac{1}{k}\sum_{j=1}^{k}(r(y_j) - \mu_r)^2},$$

其中，ϵ 是一个用于保证数值稳定性的小常数。随后，策略梯度估计为

$$\nabla_\theta \mathcal{L}(\theta) = \mathbb{E}_{q\sim\mathcal{Q},\, y\sim\pi_\theta}\big[A(y)\,\nabla_\theta \log \pi_\theta(y|q)\big].$$

奖励稀疏性与优势消失。GRPO 及其扩展方法常常面临**奖励稀疏性**：当基础多模态模型的能力有限或问题较难时，只有少数推理路径能获得正奖励，这种情况在训练早期尤为明显。它们还会遭遇**优势消失**：由于相对优势是在响应组内计算的，过难或过易的问题会导致输出全部答对或全部答错，从而使优势为零。随着训练的推进，这一问题会愈发严重，降低优化效率和样本利用率。

动机。现有解决方案通过以下方式应对这些问题：(i) 样本增强与扩充，在不调整内在难度的情况下扩大问题空间；(ii) 样本选择与利用，提高效率，但可能丢弃有价值的困难样例并降低多样性；以及 (iii)间接奖励设计，丰富监督，但可能使优化产生偏差。尽管这些方法有一定成效，却无法保证每个组内奖励方差的稳定性。

这引出了我们的核心问题：如何不受问题难度影响，保持奖励方差稳定且信息丰富？我们认为，关键在于动态评估问题

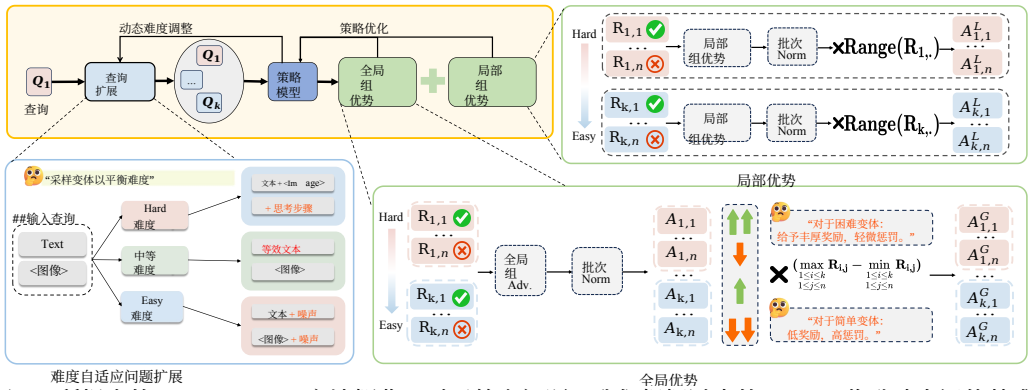


图 2：所提出的 **DIVA-GRPO** 方法概览。对于给定问题，我们根据过去的 rollout奖励动态评估其难度，并自适应地采样不同难度等级的变体。如图所示，当原始问题较难时，会采样较简单的变体，以确保奖励多样性。随后，我们计算局部（问题本身）和全局（问题及其变体）优势，并通过难度感知重加权和奖励范围重缩放得到最终优势，用于更新策略模型。

难度，并自适应地调整变体采样，确保每个问题都能产生正向和负向反馈。

3 基于 GRPO 的难度自适应变体优势

为应对这一挑战，我们设计了一种难度自适应方法 **DIVA-GRPO**。其核心思想是动态评估问题难度，自适应地采样不同难度的语义一致变体，并提升问题及其变体空间内奖励的多样性。通过计算难度加权的局部（针对原始问题）和全局（针对问题及其变体）组优势，该框架缓解了奖励稀疏和优势消失问题。图 2 展示了整体流程。

3.1 样本难度评估

实施自适应策略的前提是为每个问题分配适当的难度等级。值得注意的是，问题难度既非固有属性，也非静态不变；相反，应根据模型能力的演变动态评估。为此，DIVA-GRPO 引入了一种基于历史 rollout 结果的难度评估机制：如果某个问题的大多数 rollout 结果正确，则认为该问题对当前模型相对简单；否则，则认为相对困难。每个训练轮次都会重新校准此评估，确保感知到的难度随模型能力提升而调整。随后，这些难度估计将用于指导变体生成和采样策略，使模型能够在持续提供有意义的优势信号的同时，平衡探索与利用。

我们将问题难度定义在范围 $[D_{\min}, D_{\max}]$ 内，其中 $D_{\min}$ 表示最简单的等级，$D_{\max}$ 表示最困难的等级。在训练开始时，所有问题都初始化为该范围的中点：$D_{\text{mid}} = \frac{D_{\min}+D_{\max}}{2}$。设每个原始问题包含 m 个变体，且每个变体进行 k 次 rollout。我们将这些 rollout 的经验准确率记为 $\alpha = \frac{1}{mk}\sum_{i=1}^{m}\sum_{j=1}^{k}\mathbb{I}[y_{i,j}\text{ is correct}],$，其中 $\mathbb{I}[\cdot]$ 是指示函数。随后按照以下规则更新难度：

$$D^{\text{new}} = \text{clip}\left(D^{\text{old}} + \eta \cdot (0.5 - \alpha),\, D_{\min},\, D_{\max}\right),$$

其中，$\eta > 0$ 是控制调整幅度的学习率，而 clip(·，$D_{\min}$，$D_{\max}$) 可确保难度保持在 $[D_{\min}, D_{\max}]$ 范围内。该更新规则保证：

- 如果大多数 rollout 都答对（$\alpha \to 1$），难度就会降低；
- 如果大多数 rollout 都失败（$\alpha \to 0$），难度就会提高；
- 如果准确率在 50% 左右，难度就保持稳定。

这样一来，难度等级会随着模型性能平滑变化，相比简单的“全对/全错”规则，能够进行更细致的调整，并确保在整个训练过程中持续校准采样策略。关于超参数 η，其消融实验及选择建议见附录 D.2。

3.2 难度自适应变体生成

我们的目标是为每个问题提供适当的优势。为此，我们引入了一种基于上述每个问题的难度评估来生成自适应变体的方法。将每个原始问题表示为 $q = (I_q, T_q)$，其中 I_q 是关联图像，T_q 是文本描述和问题。在评估每个问题的难度 $D_q \in [D_{\min}, D_{\max}]$ 后，我们生成一组难度特定一致变体 $\mathcal{V}_q = \{q^{(1)}, q^{(2)}, \dots\}$ 用于训练，确保每个变体保留原始答案 y_q^*，同时有选择地调整其难度，以优化模型的学习。通过根据 D_q 动态调整每个变体的特征，我们旨在确保每个问题的奖励分布能够提供有意义的优势信号，同时保持更加平衡的特性。采样方法详见附录 D.2，附录 H 中还展示了若干变体案例研究。

- **较简单问题**（$D_q < D_{\text{中}}$）：通过扰动文本和图像来生成变体：$q^{(i)} = (I^{(i)}_q, T^{(i)}_q)$。文本变体 $T^{(i)}_q$ 通过改写、重组原问题，或对其进行轻微的语言扰动来创建，同时确保答案保持不变。这些修改能增强模型对细微差异的敏感度，从而提高鲁棒性。图像变体 $I^{(i)}_q$ 通过扰动函数生成，例如旋转、添加高斯噪声、椒盐噪声、散斑噪声以及模糊。更强或多重扰动会在保持正确性的同时增加难度。或者，也可以将文本信息 T_q 嵌入变体集 $\mathcal{V}_q$ 中的图像，并以图像形式提供，以传达问题信息。

- **中等难度问题**（$D_q \approx D_{\text{中}}$）：通过创建语义等价的文本版本来生成变体：$\mathcal{V}_q = \{(I_q, T^{(i)}_q) \mid T^{(i)}_q \text{是} T_q\}$的释义。这些变体在保持原有难度的同时，使表达更加多样化。这样，模型便能接触到同一问题的多种表述，从而提升对未见表达的泛化能力。

- **困难问题**（$D_q > D_{\text{中}}$）：变体纳入部分推理引导：$q^{(i)} = (I_q, T_q \oplus R^{(i)}_q)$，其中 $R^{(i)}_q$ 是由闭源模型生成并验证的中间推理步骤序列。对于更困难的问题，会提供额外的推理步骤作为提示。这确保即使是具有挑战性的问题也能产生有意义的优势信号，从而缓解梯度消失，并促进对复杂推理的逐步掌握。

3.3 难度加权与归一化优势平衡

在介绍我们的平衡策略之前，我们首先回顾基于GRPO的训练中所使用的局部与全局优势概念。虽然GRPO在单个问题内计算优势，但近期工作Yao等人（2025）引入了语义一致变体，以实现更广泛的比较。设原始问题记为 $q = (I_q, T_q)$，其关联图像为 I_q，文本描述为T_q。对于每个 q，构建一个变体集合 $\mathcal{V}_q = \{q^{(1)}, \dots, q^{(N)}\}$，仅修改文本或图像，同时保留标准答案。定义了两类优势：

- **局部优势：** $A_{\text{局部}}(y_{(i)\,i})$，针对每个问题使用标准GRPO公式计算。
- **全局优势：** $A_{\text{全局}}(y_{(j)\;i})$，在变体集的所有回答上计算：$A_{\text{全局}}(y_{(j)\,i}) = {}^{r}{}_{(y_{(j)}}{}^{i}{}_{\sigma_q + \epsilon}{}^{-\mu_q})$，其中 μ_q 和 σ_q 是 $\mathcal{V}_q$中所有回答奖励的均值和标准差。

为确保局部与全局优势都能有效发挥作用，同时考虑问题难度的差异，我们提出了一种两步平衡策略：(1) 归一化，使局部与全局信号具有可比性；(2) 难度加权缩放，根据每个问题的动态难度系数自适应地重新缩放优势。这一设计通过防止全局优势占据主导，并使奖励信号与问题难度相匹配，从而稳定优化过程。

形式上，对每个问题采样多个变体后，我们会计算局部与全局优势。然而，这一过程中会出现两个问题。

（1）局部–全局不平衡。局部与全局优势的量级并不相等：局部优势由 k 个 rollout 计算得出，而全局优势基于 $m \times k$ 个样本。因此，全局优势往往更大并主导优化，而局部信号的权重不足。对原始局部与全局优势分布的详细分析见附录 D.6。为解决该问题，我们分别应用批次级 z-score 归一化：

$$\tilde{A}_{\text{local}}(y) = \frac{A_{\text{local}}(y) - \mu_{\text{local}}}{\sigma_{\text{local}} + \epsilon}, \quad \tilde{A}_{\text{global}}(y) = \frac{A_{\text{global}}(y) - \mu_{\text{global}}}{\sigma_{\text{global}} + \epsilon},$$

其中 $\mu_{\text{local}}, \sigma_{\text{local}}$ 与 μ_{global},σ_{global} 分别是一个批次内局部优势与全局优势的均值和标准差。这确保两种信号保持可比，并充分参与训练。

（2）难度加权缩放。现有方法在计算优势时对简单问题与困难问题一视同仁，而未考虑其不同的难度等级。为鼓励模型攻克更难的问题，我们在归一化之后引入难度加权缩放。

令 $\{D_q^{(i)}\}_{i=1}^N$ 表示问题组 $\mathcal{V}_q$中 N 个变体的难度系数，并令$\bar{D}_q = \frac{1}{N}\sum_{i=1}^N D_q^{(i)}$为组内平均难度。对于与变体$q^{(i)}$相关联的每个回答 y_i，重新缩放的优势计算如下

$$\hat{A}(y_i \mid q^{(i)}) = \exp\Big(k \cdot (D_q^{(i)} - \bar{D}_q) \cdot \text{sgn}(\tilde{A}(y_i))\Big) \cdot \tilde{A}(y_i),$$

其中， $\tilde{A}(y_i)$ 表示归一化优势，sgn(·) 表示符号函数， $k > 0$ 表示敏感度。附录 D.5 通过消融实验分析并验证了采用相对难度加权的原因，而敏感度参数 k 的分析见附录 D.1。

直观而言，当某个变体的难度高于组平均水平（$D_q^{(i)} > \bar{D}_q$）时，正确答案（$\tilde{A} > 0$）会得到强化，而错误答案则会被弱化。相反，当某个变体的难度低于平均水平时，正确答案的权重会降低，错误答案则会受到更重的惩罚。

通过这种方式，训练过程实现了难度自适应优化，有效平衡了问题组内的相对难度和优势幅度各自的贡献。附录 B 严谨地证明了这一平衡策略的理论有效性：定理 B.1 表明，降低梯度方差能够加快收敛；推论 B.2 则表明，我们的归一化与加权策略在保持无偏梯度估计的同时，显著降低了方差。此外，附录 C 从数学上证实，当正确样本与错误样本的比例约为 1:1 时，优化信号最强，为我们的动态难度调整机制奠定了坚实的理论基础。

3.4 基于奖励范围的优势重缩放（RRB 重缩放）

在基于 GRPO 的强化学习中，我们观察到，当采样组内的奖励范围较小时，优势估计可能变得不可靠。如果奖励高度集中，标准 z-分数归一化可能会夸大细微差异，从而产生误导性的优化信号。因此，模型可能会对微不足道的提升给予过高奖励，却忽视显著的提升。

考虑这样一种奖励方案：格式正确的回答获得 0.1，完全正确的回答获得 0.9。假设有两个样本，各自进行 5 次 rollout：样本 A 的奖励为$[0, 0, 0, 0, 0.1]$，样本 B 的奖励为 $[0, 0, 0, 0, 1]$。应用标准 z- 分数归一化后，两个组都会为零奖励的 rollout 分配相同的优势（-0.45），并为最后一次 rollout 分配相同的高优势（1.79）。这会高估样本 A 中格式正确性略有提升所带来的收益，同时低估样本 B 中更大的成就。

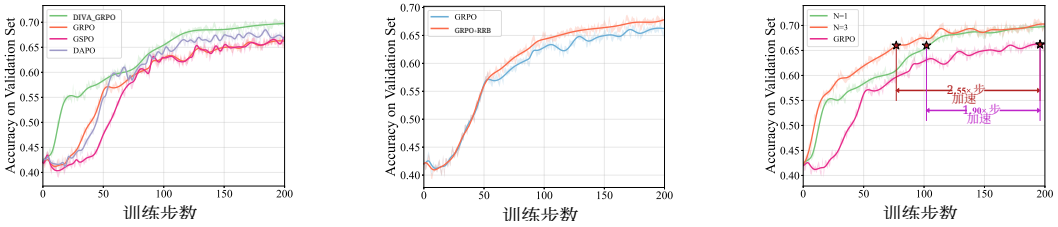
为解决这一问题，我们提出基于奖励范围的优势重缩放。令 $\mathcal{R}_q = \{r_1, r_2, \ldots, r_k\}$ 表示问题 q 的所有 rollout 的奖励，最大可能的奖励范围为 $R_{\max}$。我们将奖励范围和重新缩放的优势定义为

$$\Delta r_q = (\max(\mathcal{R}_q) - \min(\mathcal{R}_q))/R_{\max}, \quad \hat{A}_{\text{range}}(y_i) = \Delta r_q \cdot \tilde{A}(y_i)$$

其中， $\tilde{A}(y_i)$ 是通过标准 z 分数获得的归一化优势。直观而言，该方法确保优势的幅度反映奖励的实际变化程度，从而避免

表1：多模态数学基准上的性能比较。粗体表示 7B 模型中的最佳性能，下划线表示整体最佳性能。评估使用 VLMEvalKit（Duan 等人，2024年）进行，其他模型的结果则取自 Meng 等人（2025年）和 Yao 等人（ 2025 年）。对于每个条目，“/” 前的分数是我们使用官方发布的检查点重新评估所得，“/” 后的分数则是原论文报告的分数。

模型	MathVista	MathVerse	MathVision	OlympiadBench	WeMath	MMK12test	Avg.
Claude3.7-Sonnet	66.8	52.0	41.3	*48.9*	72.6	55.3	56.15
GPT-4o	63.8	50.2	30.4	*35.0*	68.8	49.9	49.68
o1	73.9	57.0	<u>60.3</u>	<u>68.0</u>	<u>98.7</u>	<u>73.9</u>	<u>72.30</u>
Gemini2-flash	70.4	<u>59.3</u>	41.3	51.0	71.4	65.2	59.77
InternVL2.5-VL-8B	64.4	39.5	19.7	12.3	53.5	45.6	39.17
Qwen-2.5-VL-7B	68.2	47.9	25.4	20.2	62.1	53.6	46.23
InternVL2.5-VL-38B	71.9	49.4	31.8	32.0	67.5	58.0	51.77
Qwen-2.5-VL-32B	71.7/74.7	49.9	40.1	30.0	69.1	66.8	54.6
InternVL2.5-VL-78B	72.3	51.7	32.2	31.1	66.3	61.6	52.53
Qwen-2.5-VL-72B	<u>74.8</u>	57.6	38.1	40.4	72.4	70.5	59.0
InternVL2.5-8B-MPO	68.9	35.5	21.5	7.8	53.5	34.5	36.95
InternVL2.5-38B-MPO	73.8	46.5	32.3	25.6	66.2	48.3	48.78
QVQ-72B-Preview	71.4	48.2	35.9	33.2	65.4	61.5	52.60
Adora-7B	73.5	50.1	23.0	20.1	64.2	58.1	48.17
R1-Onevision-7B	64.1	47.1	23.5/29.9	17.3	61.8	39.8	42.27
OpenVLThinker-7B	70.2	47.9	25.3	20.1	64.3	60.6	48.07
MM-Eureka-7B	71.7/73.0	50.3	26.9	20.1	66.1	64.5	49.93
R1-ShareVL-7B	73.5/75.4	52.8	29.5	21.3	67.9	68.8	52.30
SFT-7B	66.4	49.6	21.3	16.2	57.6	54.3	44.23
DIVA-GRPO-7B（我们的方法）	**74.2**	**57.6**	**32.1**	**23.1**	**69.3**	**70.2**	**54.58**



（a）RQ1：DIVA-GRPO 的有效性

（b）RQ3：RRB 对通用 GRPO 方法的有效性

（c）RQ4：DIVA-GRPO 对效率和速度的影响

图3：训练步数对验证集上模型性能的影响。

避免细微差异被夸大，并提供更可靠的优化信号。这种重新缩放方法可以独立于难度加权缩放应用，并且兼容任何基于GRPO的框架，从而提高策略优化的稳定性和效率。

4 实验

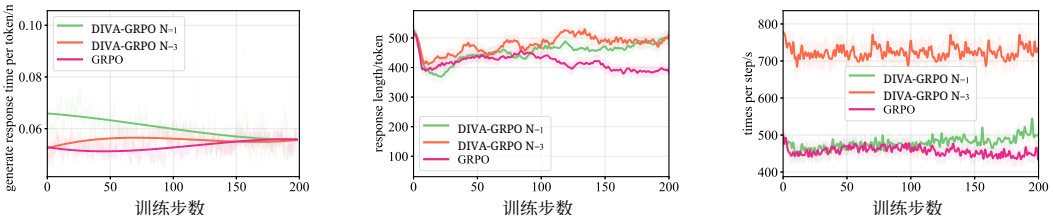
本节通过大量实验评估 DIVA-GRPO 的有效性，并辅以消融研究来评估其关键组件的作用。我们的评估围绕以下研究问题展开：

- **RQ1（有效性）**：DIVA-GRPO 与近期先进系统相比表现如何?
- **RQ2（完整性）**：移除关键组件会对模型性能产生什么影响?
- **RQ3（泛化）**：RRB 重缩放对标准GRPO的泛化能力如何?
- **RQ4（效率）**：DIVA-GRPO 会对训练效率产生什么影响?

4.1 实验设置

基准测试。 我们的评估涵盖六个多样化的基准测试：*MathVista*（Lu等人，2023）、*Math-Verse*（Zhang 等，2024年）、*MathVision*（Wang 等，2024年）、*OlympiadBench*（He等人，2024年）、*We-Math*（Qiao等人，2024年）以及*MMK12-test*（Meng 等，2025年）。

基线。 我们的比较涵盖三类模型：*(i)* 闭源专有多模态大语言模型；*(ii)* 开源基础多模态大语言模型；*(iii)* 开源微调多模态大语言模型。



(a) 每个词元的生成时间　(b) 回答的平均长度　(c) 每步耗时

图4：训练和测试期间的耗时与训练速度统计，表明我们的方法仅使每步训练时间略有增加，主要是由于输出长度增加所致。

实现。 我们使用 AdamW 对 **Qwen2.5-VL-7B-Instruct** （Bai 等人，2025年）进行训练，学习率为 10^{-6}。我们采用最新的 EasyR1 框架进行训练。具体而言，对于不生成变体的方法，我们设置 $k = 10$；对于生成变体的方法，我们设置 $k = 5$，并采用 $N = 1$ 和 $N = 3$ 两种设置，以确保实验的公平性。此设置确保了对所有基线模型进行严格且公平的比较。难度分数 D_q 初始化为 5（$D_{\min} = 1$、 $D_{\max} = 9$），并使用 $\eta = 4$。详细的参数分析见附录 D。文本推理提示预先离线生成，而图像扰动则在线应用。所有文本变体和推理序列均使用 GPT-o3 离线生成。

4.2 主要结果（**RQ1：有效性**）

我们报告了 7B 规模模型上的结果，主要实验在 R1-ShareVL-52K 数据集上进行。与基础模型和其他多模态训练方法相比，DIVA-GRPO 展现出更优的性能。如表1所示，我们针对五个广泛使用且具有挑战性的基准测试，与最先进系统进行了全面比较。

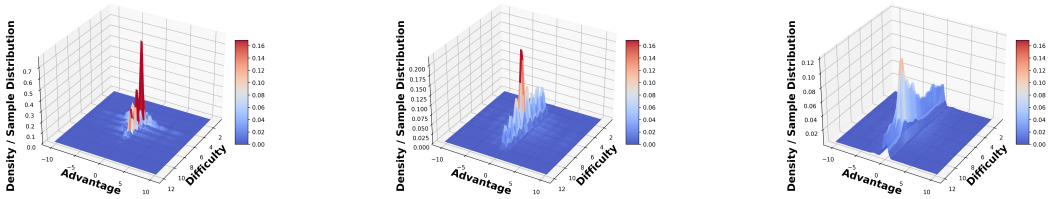
我们的方法在 7B规模下的所有数据集上均达到了最先进的性能，在中文和英文基准测试上都取得了稳定且出色的结果，平均得分为 54.58。值得注意的是，在 MathVista、MathVerse 和 WeMath 上，其性能已经达到更大规模的 Qwen2.5-VL-72B 的水平，同时显著优于多个开源基础模型（例如 Qwen2.5-VL-32B、InternVL2.5-VL-78B）和专有系统（例如 GPT-4o、Claude 3.7-Sonnet），展现出显著的效率和成本效益。然而，在更具挑战性的竞赛级数学任务上，我们的方法仍逊于超大规模开源模型和前沿专有系统，这可能是由于模型容量和训练数据覆盖范围存在固有限制。我们进一步探索了使用 GPT-o3 生成的推理轨迹，直接对骨干模型进行监督微调，结果发现，这种直接训练不仅未能提升性能，甚至低于基础模型，进一步凸显了我们方法的有效性。与其骨干模型 Qwen2.5-VL-7B 相比，我们的方法在所有基准测试上都持续取得性能提升，平均准确率提高了 8.23 点，突显了其相对于现有多模态训练框架的明显优势。关于不同外部模型的讨论见附录 D.4。

图 5 进一步展示了训练动态：在早期阶段，样本尚未形成清晰的难度分布，大多数集中在中等难度附近，且仅表现出较小的优势。随着训练的推进，分布逐渐向更简单和更困难的问题两端扩展，优势信号在不同难度等级上变得更加分明。在后期阶段，模型即使在高难度样本上仍能保持信息丰富且平衡的优势信号，而不会坍缩为平凡的“全对”或“全错”状态。这些动态凸显了 DIVA-GRPO 在维持稳定且高效优化方面的有效性。

4.3 消融研究

RQ2（完整性）：DIVA-GRPO 的组件消融

我们在 7B 规模下，使用具有代表性的基准测试（*MathVista*、*MathVerse* 和*MMK12test*），对 DIVA-GRPO 的三个核心组件进行了消融研究。为确保比较及时且公平，实验使用从*MMK12* 数据集中随机抽取的 5,000 个样本进行。每 20 步为一个训练周期，总计训练 10 轮。我们评估了移除



(a) 训练阶段 1　　(b) 训练阶段 2　　(c) 训练阶段 3

图 5：不同训练阶段中问题难度与全局优势联合分布的三维核密度估计（KDE）曲面。曲面高度反映样本密度，展示了模型学习动态如何随难度和优势而演变。

自适应变体生成（包括全局-局部优势和平衡）、难度加权、基于奖励范围的优势重缩放（RRB 重缩放）以及全局-局部平衡（G-LBalance）的影响，并将每个变体与完整模型进行比较。

如表 2 所示，移除任一组件都会导致性能持续下降，而完整的 DIVA-GRPO 模型在所有基准测试中均取得最高准确率。这些结果表明，所有组件都能带来互补的增益，缺少任何一个都会造成性能下降，凸显了完整模型设计的必要性。

此外，如图 3a 所示，我们在整个训练过程中将 DIVA-GRPO 与 GRPO、GSPO（Zheng 等人，2025年）和 DAPO（Yu 等人，2025年）进行比较。实验结果表明，在严格控制时间和回答总长度的情况下，我们的方法在训练速度和稳定性方面均优于 GRPO、GSPO 和 DAPO。此外，我们还评估了我们的方法在 Qwen3-VL-4B 和 8B 上的性能，详细结果见附录 D.3。

表 2：DIVA-GRPO 的消融研究，表明每个组件都能带来提升，且完整模型表现最佳（准确率，%）。

方法	MathVista	MathVerse	MMK12test	Avg.
不含变体生成	70.0	53.7	61.1	61.6
不使用难度加权	69.9	55.7	66.5	64.0
不使用 RRB 重缩放	71.5	55.2	64.7	63.8
不使用 G-L 平衡	70.8	55.4	66.0	64.1
完整的 DIVA-GRPO	**73.2**	**56.3**	**68.8**	**66.1**

RQ3（泛化）：RRB 重缩放的泛化能力

为验证基于奖励范围的优势重缩放（RRB-Rescaling）更广泛的适用性，我们将其纳入标准GRPO，且不做任何其他修改。表 3显示，加入 RRB 重缩放后，平均准确率从60.23提升至62.17，并且在全部三个基准测试中均有所改善。图3b进一步展示了这一点。这表明，在通用GRPO设置下，RRB 重缩放能够同时提升稳定性和性能，证实该组件并不局限于DIVA-GRPO，也能使其他GRPO风格训练框架受益。

表 3：在标准GRPO上评估RRB 重缩放。

方法	MathVista	MathVerse	MMK12test	Avg.
GRPO（基线）	69.3	52.8	58.6	60.23
+ RRB	**70.0**	**53.6**	**62.9**	**62.17**

RQ4（效率）：训练效率与稳定性

我们在Qwen2.5-VL-7B上比较DIVA-GRPO与GRPO的训练性能。如图3c所示，我们测量DIVA-GRPO在GRPO训练的10个训练周期内达到测试集最优值所需的步数。当GRPO使用 $n = 10$且DIVA-GRPO使用 $n = 5$时，我们确保在 $N = 1$ 时采样句子数量保持不变，并在 $N = 3$时输入数据量保持一致。图4a展示了生成过程中每个词元所消耗的时间，图4b说明了平均响应长度，图4c描绘了每步所需的时间。接下来，我们从步数和时间两个方面计算DIVA-GRPO的加速比。

• **对于** $N = 1$，我们确保回答数量一致，也就是说，样本总数保持不变，同时 DIVA-GRPO 将步骤加速了 1.90× 倍。根据统计数据，尽管每个词元的耗时仅增加 13%，每个步骤的耗时仅增加 4%，但从总耗时来看，我们的方法平均每步耗时 7.98 分钟，而 GRPO 方法每步耗时 7.61 分钟。生成变体和思考步骤需要

每 5000 个样本大约需要 32 分钟，而视觉变体的生成几乎是即时完成的，不会产生额外的时间成本，因为这部分时间已计入训练时间。即使将生成变体所需的时间也计算在内，我们的方法在耗时方面仍实现了 1.76× 倍加速。

• **对于** $N = 3$，我们确保数据需求保持不变，同时 DIVA-GRPO 将步骤加速了 2.55× 倍。在这种情况下，每步平均耗时为 12.03 分钟。即使将生成变体所需的时间计算在内，我们的方法在时间上仍实现了 1.56× 倍加速。这清楚地展现了我们方法的效率。

5 相关工作

随着 GRPO 在多模态大语言模型中越来越多地用于增强推理能力，仍有若干挑战尚未解决，尤其是稀疏奖励和优势消失。现有解决方案可分为三种主要途径：样本增强与问题空间扩展、样本选择与利用，以及奖励设计。

样本增强与扩展 方法旨在通过难度感知机制或生成多样化样本来扩大问题空间，从而提升探索能力与泛化能力。Noisy Rollout（Liu 等，2025a）采用图像加噪来促进强化学习，引入视觉多样性以防止过拟合并提升鲁棒性。Hint-GRPO（Huang 等，2025）根据任务难度调整提示，从而提高困难样本的数据效率。DeepVideo-R1（Park 等，2025）调整难度，帮助变体重新获得优势，但并未改变单个样本的难度分布，因此在极端情况下仍会出现稀疏奖励。R1-shareVL（Yao 等，2025）通过多个变体扩展问题空间，并采用全局优势来调整难度，但缺少缓解稀疏性的完整机制。

样本选择与利用 方法将计算资源集中于有效数据，同时忽略无益样本，从而提升效率，但也有过早放弃困难样例的风险。VL-Rethinker（Wang 等，2025a）采用选择性重放来缓解奖励稀疏性，并通过重放过去表现出色的样本来保持优势的多样性。MM-Eureka（Meng 等，2025）选择部分正确的提示词，过滤掉优势为零的情况，以稳定训练。尽管这些方法有所助益，却未能最大限度地利用整体样本，也未提供全局优化视角。

奖励设计 方法旨在提供更密集、连续的信号，以减少稀疏性并稳定推理。MM-Eureka 引入超参数 λ，用于调整格式奖励的权重。R1-VL（Zhang 等，2025）提出 StepGRPO，将推理准确率奖励和逐步有效性奖励纳入其中，以提供更密集的反馈。因此，引入的间接奖励可能偏离主要目标，并可能误导模型。DeepVideo-R1 提出 Reg-GRPO，将 GRPO 转换为回归损失；而 GRPO-LEAD（Zhang 和 Zuo，2025）则通过动态难度感知，在优势计算中对困难问题加权。Dr. GRPO（Liu 等，2025b）通过移除优势估计中的方差项，力求降低方差并稳定学习信号。这与我们的 RRB 方法动机相同，因为两者都旨在提供更稳定的学习信号。尽管这些方法缓解了收敛和稳定性问题，却未能解决奖励稀疏和优势递减这一根本挑战。

6 结论

在本研究中，我们发现并解决了基于 GRPO 的多模态大语言模型强化学习的一项根本局限：奖励稀疏性和优势消失，这些问题阻碍了有效的长链推理。为此，我们提出了 DIVA-GRPO，这是一种难度自适应的变体优势框架，可动态评估问题难度、生成定制变体，并通过难度感知归一化和基于奖励范围的重新缩放计算局部与全局优势。该设计确保了面对不同难度的问题时，优化信号稳定且信息丰富，同时缓解奖励稀疏性并提升训练稳定性。在六个多模态推理基准上的大量实验表明，DIVA-GRPO 在推理准确率、收敛速度和训练效率方面始终优于现有的基于 GRPO 的方法。总的来说，我们的方法为改进多模态大语言模型中的强化学习提供了一个具有泛化能力的框架，并为复杂多模态推理任务中的探索、样本利用和优势估计提供了一种有原则的方法。

致谢

本研究得到中国科学院战略性先导科技专项（项目编号：XDB0680302）、国家自然科学基金委员会（NSFC，项目编号：62276248）、新疆维吾尔自治区重点研发计划（项目编号：2024B03026）、北京市科技新星计划（项目编号：20250484765）以及中国科学院青年创新促进会（项目编号：2023111）的资助。

伦理声明

在本研究中，所有数据收集与分析均遵循伦理标准进行，且未涉及人类受试者或动物实验。本研究未使用任何敏感个人数据或隐私信息，所有实验设计与数据处理方法均遵守相关法律与伦理准则。在数据集发布方面，我们严格遵循适用的学术伦理规范，确保所有数据与代码共享均符合法律与伦理要求。本研究不存在利益冲突，所有资金来源与支持均已明确披露。本研究确保透明、诚信与公平，并严格遵守相关的研究伦理与合规标准。

可复现性声明

为确保我们实验的完全可复现性，我们在代码仓库中提供了复现结果所需的全部代码、配置与说明：https://github.com/Siaaaaaa1/DIVA-GRPO。该代码仓库包含训练与评估脚本、预处理流程，以及详细的 README 文档，其中说明了依赖项以及复现所报告结果的步骤。

参考文献

Jean-Baptiste Alayrac、Jeff Donahue、Pauline Luc、Antoine Miech、Iain Barr、Yana Hasson、Karel Lenc、Arthur Mensch、Katherine Millican、Malcolm Reynolds 等。Flamingo：一种用于少样本学习的视觉语言模型。神经信息处理系统进展，35：23716–23736，2022年。

Stanislaw Antol、Aishwarya Agrawal、Jiasen Lu、Margaret Mitchell、Dhruv Batra、C Lawrence Zitnick 和 Devi Parikh。VQA：视觉问答。载于电气电子工程师学会国际计算机视觉大会论文集，第2425–2433页，2015年。

Shuai Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wenbin Ge、Sibo Song、Kai Dang、Peng Wang、Shijie Wang、Jun Tang 等。Qwen2.5-VL 技术报告。*arXiv* 预印本 *arXiv:2502.13923*，2025年。

Qiguang Chen、Libo Qin、Jin Zhang、Zhi Chen、Xiao Xu 和 Wanxiang Che。M^3cot：一种用于多领域、多步多模态思维链的新型基准。*arXiv* 预印本 *arXiv:2405.16473*，2024a。

Zhe Chen、Weiyun Wang、Hao Tian、Shenglong Ye、Zhangwei Gao、Erfei Cui、Wenwen Tong、Kongzhi Hu、Jiapeng Luo、Zheng Ma 等。我们距离 GPT-4V 还有多远？通过开源套件缩小与商业多模态模型之间的差距。中国科学：信息科学，67(12):220101，2024b。

Haodong Duan、Junming Yang、Yuxuan Qiao、Xinyu Fang、Lin Chen、Yuan Liu、Xiaoyi Dong、Yuhang Zang、Pan Zhang、Jiaqi Wang 等。VLMEvalKit：用于评估大型多模态模型的开源工具包。载于第*32*届 *ACM*国际多媒体会议论文集，第11198–11201页，2024年。

Hao Fei、Shengqiong Wu、Wei Ji、Hanwang Zhang、Meishan Zhang、Mong-Li Lee 和 Wynne Hsu。Video-of-Thought：从感知到认知的逐步视频推理。*arXiv* 预印本 *arXiv:2501.03230*，2024年。

Haowen Gao、Liang Pang、Shicheng Xu、Leigang Qu、Tat-Seng Chua、Huawei Shen 和 Xueqi Cheng。生成式幽灵：探究人工智能生成的视频中隐藏的排序偏差。载于第*33*届 *ACM* 国际多媒体会议论文集，第6403–6411页。计算机协会，2025年。doi：10.1145/3746027.3754815。

Shasha Guo、Liang Pang、Xi Wang、Yanling Wang、Huawei Shen 和 Jing Zhang。GeoVLMATH：通过用于辅助线构造的跨模态奖励增强视觉语言模型的几何推理。*arXiv* 预印本 *arXiv:2510.11020*，2025年。

Chaoqun He、Renjie Luo、Yuzhuo Bai、Shengding Hu、Zhen Leng Thai、Junhao Shen、Jinyi Hu、Xu Han、Yujie Huang、Yuxiang Zhang 等。OlympiadBench：一项旨在通过奥林匹克级双语多模态科学问题推动通用人工智能发展的高难度基准测试。*arXiv* 预印本 *arXiv:2402.14008*，2024年。

Jie Huang 和 Kevin Chen-Chuan Chang。迈向大语言模型中的推理：一项综述。*arXiv* 预印本 *arXiv:2212.10403*，2022年。

Qihan Huang、Weilong Dai、Jinlong Liu、Wanggui He、Hao Jiang、Mingli Song、Jingyuan Chen、Chang Yao 和 Jie Song。通过文本去偏 HINT-GRPO 提升多模态大语言模型的推理能力。*arXiv* 预印本 *arXiv:2503.23905*，2025年。

Aaron Hurst、AdamLerer、AdamP Goucher、Adam Perelman、Aditya Ramesh、Aidan Clark、AJ Ostrow、Akila Welihinda、Alan Hayes、Alec Radford 等。GPT-4o 系统卡片。*arXiv* 预印本 *arXiv:2410.21276*，2024年。

Hugo Laurençon、Andrés Marafioti、Victor Sanh 和 Léo Tronchon。构建并深入理解视觉语言模型：见解与未来方向。*arXiv* 预印本 *arXiv:2408.12637*，2024年。

Junnan Li、Dongxu Li、Silvio Savarese 和 Steven Hoi。BLIP-2：利用冻结的图像编码器和大语言模型引导语言-图像预训练。载于国际机器学习会议，第 19730–19742 页。PMLR，2023。

Haotian Liu、Chunyuan Li、Qingyang Wu 和 Yong Jae Lee。视觉指令微调。《神经信息处理系统进展》经信息处理系统，36:34892–34916，2023。

Xiangyan Liu、Jinjie Ni、Zijian Wu、Chao Du、Longxu Dou、Haonan Wang、Tianyu Pang 和 Michael Qizhe Shieh。Noisyrollout：通过数据增强强化视觉推理。*arXiv* 预印本 *arXiv:2504.13055*，2025a。

Zichen Liu、Changyu Chen、Wenjun Li、Penghui Qi、Tianyu Pang、Chao Du、Wee Sun Lee 和 Min Lin。理解类 r1-zero 训练：批判性视角。*arXiv* 预印本 *arXiv:2503.20783*，2025b。

Pan Lu、Hritik Bansal、Tony Xia、Jiacheng Liu、Chunyuan Li、Hannaneh Hajishirzi、Hao Cheng、Kai-Wei Chang、Michel Galley 和 Jianfeng Gao。MathVista：评估基础模型在视觉上下文中的数学推理能力。*arXiv* 预印本 *arXiv:2310.02255*，2023。

Fanqing Meng、Lingxiao Du、Zongkai Liu、Zhixiang Zhou、Quanfeng Lu、Daocheng Fu、Botian Shi、Wenhai Wang、Junjun He、Kaipeng Zhang 等。Mm-eureka：利用基于规则的大规模强化学习探索视觉顿悟时刻。*CoRR*，2025年。

Fanxu Meng、Haotong Yang、Yiding Wang 和 Muhan Zhang。用于直观推理的图像链。*arXiv* 预印本 *arXiv:2311.09241*，2023。

Jinhui Pang、Xinyun Yang、Xiaoyao Qiu、Zixuan Wang 和 Taisheng Huang。Mmaf：用于减少命名实体识别中视觉特征偏差的掩码多模态注意力融合。*DataIntelligence*，6(4):1114–1133，2024年。doi：10.3724/2096-7004.di.2024.0049。

Jinyoung Park、Jeehye Na、Jinyoung Kim 和 Hyunwoo J Kim。Deepvideo-r1：通过难度感知回归式 GRPO 进行视频强化微调。*arXiv* 预印本 *arXiv:2506.07464*，2025年。

Runqi Qiao、Qiuna Tan、Guanting Dong、Minhui Wu、Chong Sun、Xiaoshuai Song、Zhuoma GongQue、Shanglin Lei、Zhe Wei、Miaoxuan Zhang 等。We-math：你的大型多模态模型能否实现类人的数学推理？*arXiv* 预印本 *arXiv:2407.01284*，2024年。

Rafael Rafailov、Archit Sharma、Eric Mitchell、Christopher D Manning、Stefano Ermon和Chelsea Finn。直接偏好优化：你的语言模型实际上是一个奖励模型。神经信息处理系统进展，36:53728–53741，2023。

Gabriel Sarch、Snigdha Saha、Naitik Khandelwal、Ayush Jain、MichaelJ Tarr、Aviral Kumar和Katerina Fragkiadaki。面向视觉推理的基于视觉落地的强化学习。*arXiv*预印本*arXiv:2505.23678*，2025年。

John Schulman、Filip Wolski、Prafulla Dhariwal、Alec Radford和Oleg Klimov。近端策略优化算法。*arXiv*预印本*arXiv:1707.06347*，2017。

Zhihong Shao、Peiyi Wang、Qihao Zhu、Runxin Xu、Junxiao Song、Xiao Bi、Haowei Zhang、Mingchuan Zhang、YK Li、Yang Wu等。DeepSeekMath：推动开放语言模型中数学推理的极限。*arXiv*预印本*arXiv:2402.03300*，2024年。

Haozhe Wang、Chao Qu、Zuming Huang、Wei Chu、Fangzhen Lin 和 Wenhu Chen。VL-Rethinker：通过强化学习激励视觉语言模型进行自我反思。*arXiv* 预印本 *arXiv:2504.08837*，2025a。

Ke Wang、Junting Pan、Weikang Shi、Zimu Lu、Houxing Ren、Aojun Zhou、Mingjie Zhan 和 Hongsheng Li。使用 Math-Vision 数据集评测多模态数学推理。神经信息处理系统进展，37：95095–95169，2024年。

Yaoting Wang、Shengqiong Wu、Yuecheng Zhang、Shuicheng Yan、Ziwei Liu、Jiebo Luo 和 Hao Fei。多模态思维链推理：综合综述。*arXiv* 预印本 *arXiv:2503.12605*，2025b。

Yixuan Wu、Yizhou Wang、Shixiang Tang、Wenhao Wu、Tong He、Wanli Ouyang、Philip Torr 和 Jian Wu。DetToolChain：一种释放多模态大语言模型检测能力的新提示范式。载于欧洲计算机视觉会议，第164–182页。施普林格，2024a。

Zhiyu Wu、Xiaokang Chen、Zizheng Pan、Xingchao Liu、Wen Liu、Damai Dai、Huazuo Gao、Yiyang Ma、Chengyue Wu、Bingxuan Wang 等。Deepseek-vl2：用于高级多模态理解的混合专家视觉语言模型。*arXiv* 预印本 *arXiv:2412.10302*，2024年b。

Yijia Xiao、Edward Sun、Tianyu Liu 和 Wei Wang。Logicvista：视觉上下文中的多模态大语言模型逻辑推理基准测试。*arXiv* 预印本 *arXiv:2407.04973*，2024年。

Shicheng Xu、Liang Pang、Huawei Shen、Xueqi Cheng 和 Tat-Seng Chua。Search-in-the-chain：利用搜索交互式增强大语言模型，以处理知识密集型任务。载于*2024*年 *ACM Web Conference* 论文集，第 1362–1373 页。计算机协会，2024年。doi：10.1145/3589334.3645363。

Huanjin Yao、Qixiang Yin、Jingyi Zhang、Min Yang、Yibo Wang、Wenhao Wu、Fei Su、Li Shen、Minghui Qiu、Dacheng Tao 等。R1-sharevl：通过 share-GRPO 激励多模态大语言模型的推理能力。*arXiv* 预印本 *arXiv:2505.16673*，2025年。

Shukang Yin、Chaoyou Fu、Sirui Zhao、Ke Li、Xing Sun、Tong Xu 和 Enhong Chen。多模态大语言模型综述。*National ScienceReview*，11(12):nwae403，2024年。

Qiying Yu、Zheng Zhang、Ruofei Zhu、Yufeng Yuan、Xiaochen Zuo、Yu Yue、Weinan Dai、Tiantian Fan、Gaohong Liu、Lingjun Liu 等。DAPO：一个大规模开源大语言模型强化学习系统。*arXiv* 预印本 *arXiv:2503.14476*，2025年。

Jingyi Zhang、Jiaxing Huang、Huanjin Yao、Shunyu Liu、Xikun Zhang、Shijian Lu 和 Dacheng Tao。R1-vl：通过逐步组相对策略优化，让多模态大语言模型学会推理。*arXiv* 预印本 *arXiv:2503.12937*，2025年。

Jixiao Zhang 和 Chunsheng Zuo。GRPO-LEAD：一种用于语言模型简洁数学推理的难度感知强化学习方法。*arXiv* 预印本 *arXiv:2504.09696*，2025年。

Renrui Zhang、Dongzhi Jiang、Yichi Zhang、Haokun Lin、Ziyu Guo、Pengshuo Qiu、Aojun Zhou、Pan Lu、Kai-Wei Chang、Yu Qiao 等。MathVerse：你的多模态大语言模型真的能看懂视觉数学问题中的图表吗？发表于欧洲计算机视觉会议，第 169–186 页。施普林格，2024年。

Chujie Zheng、Shixuan Liu、Mingze Li、Xiong-Hui Chen、Bowen Yu、Chang Gao、Kai Dang、Yuqiong Liu、Rui Men、An Yang 等。组序列策略优化。*arXiv* 预印本 *arXiv:2507.18071*，2025年。

Xiongtao Zhou、Jie He、Lanyu Chen、Jingyu Li、Haojing Chen、Víctor Gutiérrez-Basulto、Jeff Z Pan 和 Hanjie Chen。Miceval：通过图像描述和推理步骤揭示多模态思维链的质量。*arXiv* 预印本 *arXiv:2410.14668*，2024年。

Deyao Zhu、Jun Chen、Xiaoqian Shen、Xiang Li 和 Mohamed Elhoseiny。MiniGPT-4：利用先进的大语言模型增强视觉语言理解。*arXiv* 预印本 *arXiv:2304.10592*，2023。

A 详细基准测试与基线模型

A.1 基准测试

我们评估了六项多模态数学推理基准测试：

• **MathVista**： 一项综合性基准测试，汇集了涵盖多种数学与视觉任务的挑战。它包含 *6,141* 个样例，源自 *28* 个现有的数学多模态数据集，以及 *3* 个新建数据集（即 *IQTest*、*FunctionQA* 和 *PaperQA*）。

• **MathVerse**： 一项全面的视觉数学基准测试，旨在公平、深入地评估多模态大语言模型（*MLLMs*）。它包含 *2,612* 道带有图表的高质量多学科数学问题，每道题均转换为六种不同版本，共计 *15K* 个测试样本。

• **MathVision**： 精心整理的 *3,040* 道高质量数学问题，均带有源自真实数学竞赛的视觉上下文。内容涵盖 *16* 个不同的数学学科，并按难度划分为 *5* 个等级。

• **OlypamidBench**： 一个包含 *8,476* 道数学和物理问题的数据集，题目来源于国际奥林匹克竞赛、中国奥林匹克竞赛和中国高考（高考）。该数据集提供专家级的逐步推理标注。

• **WeMath**： *WeMath*以类人数学推理为灵感，是首个专门用于探索端到端性能之外的问题求解原则的基准测试。它精心收集并分类了 *6.5K* 道视觉数学问题，涵盖 *67* 个层级化知识概念和 *5* 层知识粒度。

• **MMK12_测试**： 一个聚焦于物理学、化学、生物学和数学等学科专属任务的数据集。该数据集包含答案经过验证的问题，旨在评估学科专属的推理能力。

A.2 基线

按照表1采用的分类体系，我们将基线系统分为三类：

- **闭源专有多模态大语言模型**：Claude3.7-Sonnet、GPT-4o、o1、Gemini2-flash。
- **开源基础多模态大语言模型**：InternVL2.5-VL-8B、Qwen-2.5-VL-7B、InternVL2.5-VL-38B、InternVL2.5-VL-78B。
- **开源微调多模态大语言模型**：InternVL2.5-8B-MPO、InternVL2.5-38B-MPO、QVQ- 72B-Preview、Adora-7B、R1-Onevision-7B、OpenVLThinker-7B、MM-Eureka-7B、R1- ShareVL-7B。

为进行公平比较，我们提出的系统 **DIVA-GRPO-7B**采用与 Qwen2.5-VL-7B 相同的骨干网络。

B 优势归一化与难度加权平衡的理论分析

定理 B.1 （梯度方差控制）。令 g（θ）为随机策略梯度估计量

$$g(\theta) = \sum_{t=1}^{T} A_t \, \nabla_\theta \log \pi_\theta(a_t|s_t),$$

其中， A_t 表示优势函数。假设 $\mathbb{E}[g(\theta)] = \nabla J(\theta)$ 是无偏的，且 $\mathit{Var}[g(\theta)] < \infty$。那么对于步长 $\eta > 0$，期望平方误差满足

$$\mathbb{E}\big[\|\theta_{t+1} - \theta^*\|^2\big] \le \mathbb{E}\big[\|\theta_t - \theta^*\|^2\big] - \eta\|\nabla J(\theta_t)\|^2 + \eta^2 \,\mathrm{Var}[\tilde{g}(\theta_t)]. \tag{1}$$

该不等式表明，收敛速度在很大程度上取决于梯度估计量的方差。降低梯度方差可以提高优化稳定性并加快收敛。

推论 B.2 （优势归一化与难度加权平衡）。考虑调整后的梯度估计量

$$\tilde{g}(\theta) = \sum_{i=1}^{k} \hat{A}(y_i)\, \nabla_\theta \log \pi_\theta(y_i),$$

其中，调整后的优势定义为

$$\hat{A}(y_i) = \exp\Big(k \cdot (D_i - \bar{D}) \cdot \mathrm{sgn}(\tilde{A}(y_i))\Big)\, \tilde{A}(y_i).$$

~此处 $A(y_i)$ 表示批次级归一化优势（局部或全局）， D_i 表示变体的难度系数 $q^{(i)}$, $\bar{D}$ 表示组内的平均难度， $k > 0$ 表示敏感度参数。

因此：

1. 无偏性： 归一化保证缩放后的均值为零，因此 $\mathbb{E}[\tilde{g}(\theta)] = \nabla J(\theta)$。

2. 方差降低： 难度加权缩放会自适应地重新加权样本。对于困难问题，正确信号会被放大，错误答案的影响则会减弱；而对于简单问题，正确信号会被降权，错误答案则会受到更强惩罚。这种重新平衡可避免异常值占据主导地位，从而确保

$$\mathit{Var}[\tilde{g}(\theta)] \;\le\; \mathit{Var}[g(\theta)].$$

代入定理 *B.1* 可得

$$\mathbb{E}\big[\|\theta_{t+1} - \theta^*\|^2\big] \le \mathbb{E}\big[\|\theta_t - \theta^*\|^2\big] - \eta \|\nabla J(\theta_t)\|^2 + \eta^2\, \mathrm{Var}[\tilde{g}(\theta_t)]. \tag{2}$$

因此，所提出的平衡策略在降低更新方差的同时保留了无偏梯度估计，从而实现更稳定、高效的收敛。

讨论。 定理 B.1 证明，方差直接决定策略梯度方法的收敛行为。推论 B.2 表明，我们提出的难度加权与归一化优势平衡在保持无偏性的同时降低了方差。直观而言，这种方法平衡了不同问题难度下的贡献，避免全局或局部优势主导优化过程，从而稳定训练并提升效率。

C z 分数归一化下最优奖励平衡的理论证明

为进一步论证我们的设计选择，我们给出理论证明。我们的方法会自适应地评估每个问题的难度，并生成变体，使当前模型面对的整体难度保持在中等水平，旨在维持正确样本与错误样本约为 50/50 的比例（而非依赖单个正确或错误样本来产生优势）。为简化起见，我们考虑二元奖励设定：正确答案获得最高奖励，错误答案获得零奖励。证明表明，在这种平衡设定下，优化效率最高：当正确样本与错误样本数量相等时，估计梯度与最优更新方向最为一致，从而更有效地改进模型。

设置与符号。 考虑 GRPO 类流程中的一个批次（或小批次），其中包含 n 个 rollout。奖励仅取两个值 $\{0, R_{\max}\}$。令 k 为奖励为 $R_{\max}$ 的 rollout 数量，并设

$$\mu := \frac{k}{n} \in [0, 1].$$

我们通过对批次上的奖励进行 z 分数归一化来计算优势：

$$A_i = \frac{r_i - \bar{r}}{\sigma_r}, \qquad \bar{r} = \mu R_{\max}, \qquad \sigma_r = \sqrt{\mathrm{Var}(r)} = R_{\max}\sqrt{\mu(1-\mu)}.$$

令 $g_i := \nabla_\theta \log \pi(a_i|s_i) \in \mathbb{R}^d$ 表示每个样本的得分梯度向量。我们假设参数更新方向（在相差一个正常数的意义下）遵循常见的策略梯度形式

$$\Delta\theta \propto \frac{1}{n}\sum_{i=1}^{n} A_i\, g_i.$$

固定一个表示期望"最优"方向的参考单位向量 $v \in \mathbb{R}^d$。定义

$$s_+ := v^\top \mathbb{E}[g \mid r = R_{\max}], \qquad s_- := v^\top \mathbb{E}[g \mid r = 0],$$

类别梯度在 v 上的平均投影。

引理 C.1（二元奖励的优势）。 采用上述记号，两个优势值为

$$A_+ := \frac{R_{\max} - \bar{r}}{\sigma_r} = \sqrt{\frac{1-\mu}{\mu}}, \qquad A_- := \frac{0 - \bar{r}}{\sigma_r} = -\sqrt{\frac{\mu}{1-\mu}}.$$

特别地， $R_{\max}$ 会抵消，因此不会影响对 μ 的依赖关系。

证明。 直接代入可得

$$A_+ = \frac{R_{\max} - \mu R_{\max}}{R_{\max}\sqrt{\mu(1-\mu)}} = \frac{1-\mu}{\sqrt{\mu(1-\mu)}} = \sqrt{\frac{1-\mu}{\mu}},$$

$$A_- = \frac{0 - \mu R_{\max}}{R_{\max}\sqrt{\mu(1-\mu)}} = -\frac{\mu}{\sqrt{\mu(1-\mu)}} = -\sqrt{\frac{\mu}{1-\mu}}.$$

□

引理 C.2（批次更新投影到 v）。 在类别均值近似下（将每个奖励类别的样本聚合为其类别均值），更新在 v 上的投影满足（至多相差一个正常数）

$$v^\top \Delta\theta \;\propto\; \mu A_+ s_+ + (1-\mu) A_- s_- = \sqrt{\mu(1-\mu)}\,(s_+ - s_-).$$

因此，投影绝对幅度为

$$\left|v^\top \Delta\theta\right| \;\propto\; \sqrt{\mu(1-\mu)}\,|s_+ - s_-|.$$

证明。 按奖励类别对求和进行聚合：

$$\Delta\theta \propto \mu A_+ \,\mathbb{E}[g \mid r = R_{\max}] + (1-\mu) A_- \,\mathbb{E}[g \mid r = 0].$$

投影到 v 并代入 $A_\pm$ 的表达式，得到

$$v^\top \Delta\theta \propto \mu A_+ s_+ + (1-\mu) A_- s_- = \sqrt{\mu(1-\mu)}\, s_+ - \sqrt{\mu(1-\mu)}\, s_- = \sqrt{\mu(1-\mu)}\,(s_+ - s_-).$$

取绝对值即得到所述表达式。 □

定理 C.3（ $\mu = \frac{1}{2}$的最优性）。 对于任意固定的 s_+、 $s_- \in \mathbb{R}$，绝对投影更新幅度

$$F(\mu) := \sqrt{\mu(1-\mu)}\,|s_+ - s_-|$$

在 $\mu \in [0,1]$ 上于 $\mu = \frac{1}{2}$ 处达到最大值。因此，在所述假设下，一半样本的奖励为 $R_{\max}$、另一半样本的奖励为 0 的批次，能使方向 v 上的期望更新幅度最大化。

证明。 由于 $|s_+ - s_-|$ 与 μ 无关，只需在 $[0,1]$ 上最大化 $g(\mu) := \sqrt{\mu(1-\mu)}$。注意

$$g(\mu)^2 = \mu(1-\mu) = -\mu^2 + \mu,$$

一个顶点位于 $\mu = \frac{1}{2}$ 的凹二次函数，其最大值为 $\frac{1}{4}$。因此， $g(\mu)$ 在 $\mu = \frac{1}{2}$ 处取得最大值， $F(\mu)$ 也是如此。 □

两个具有代表性的几何情形。这里我们分别讨论两种极端但具有代表性的情形。

推论 C.4 (情形 A：相反类别梯度)。 假设平均类别梯度沿参考方向恰好相反，即

$$s_- = -s_+.$$

则投影更新满足（相差一个正常数）

$$v^\top \Delta\theta \;\propto\; \sqrt{\mu(1-\mu)}\,(s_+ - s_-) = 2\sqrt{\mu(1-\mu)}\,s_+,$$

因此，投影绝对幅度与以下量成正比

$$\left|v^\top \Delta\theta\right| \;\propto\; 2|s_+|\sqrt{\mu(1-\mu)}.$$

因此，该幅度在 $\mu = \frac{1}{2}$ 处达到最大值。

证明。将 $s_- = -s_+$ 代入一般表达式 $v^\top \Delta\theta \propto \sqrt{\mu(1-\mu)}(s_+ - s_-)$。这将得到所示表达式，它是 $\sqrt{\mu(1-\mu)}$ 的标量倍数。因子 $\sqrt{\mu(1-\mu)}$ 在 $\mu = \frac{1}{2}$ 处达到最大值，因此结论成立。 □

推论 C.5 (情形 B：正交类别梯度)。 假设负类的类别均值梯度与参考方向正交，即

$$s_- = 0,$$

而 $s_+ \neq 0$。则

$$v^\top \Delta\theta \;\propto\; \sqrt{\mu(1-\mu)}\,s_+,$$

因此，投影绝对幅度为

$$\left|v^\top \Delta\theta\right| \;\propto\; |s_+|\sqrt{\mu(1-\mu)},$$

其最大值同样在 $\mu = \frac{1}{2}$ 处取得。

此外，考虑一种更强的几何情形：正类的均值梯度沿着 v，范数为 $\|g_+\| = \alpha$；负类的均值梯度与其正交，范数为$\|g_-\| = \beta$。如果两个类别的归一化优势加权贡献幅度相等（即 $\alpha\sqrt{\mu(1-\mu)} = \beta\sqrt{\mu(1-\mu)}$，等价地，$\alpha = \beta$），那么所得更新向量就是两个大小相等的正交向量之和，因此更新向量与 v之间的角度 ϕ 满足余弦 $\phi = 1/\sqrt{2}$（与 μ无关），而在 v 上的投影绝对幅度仍按 $\sqrt{\mu(1-\mu)}$ 变化，并在$\mu = \frac{1}{2}$ 处达到最大值。

证明。第一条结论可通过将 $s_- = 0$ 代入一般投影公式立即得出。对于几何解释，若两个（正交的）类别均值向量的加权幅值相等，则它们的向量和长度为单个幅值的 $\sqrt{2}$ 倍，而在 v 上的投影等于该幅值；除以总范数后得到 $\cos\phi = 1/\sqrt{2}$。对 μ 的依赖仅通过公共幅值因子 $\sqrt{\mu(1-\mu)}$体现，该因子在 $\mu = \frac{1}{2}$处达到最大值。 □

备注。

1. 结论不依赖于 $R_{\max}$：z 分数标准化会抵消奖励尺度。2. 如果 $s_+ > s_-$，那么 $v^\top \Delta\theta > 0$，更新会朝着 v 移动；如果 $s_+ < s_-$，更新则会朝着与 v 相反的方向移动。无论哪种情况，投影的绝对值都在 $\mu = \frac{1}{2}$ 时达到最大。3. 证明采用了简化的类别均值近似。实践中，每个类别都存在内部方差；尽管如此，期望投影对 μ 的主导阶依赖仍由 $\sqrt{\mu}$（$1-\mu$）决定。在类内方差满足温和条件时，定性结论（在 $\mu = \frac{1}{2}$ 时达到最大）在期望意义下仍然成立。4. 对于非二元奖励，代数形式会发生变化：必须针对完整的奖励分布分析 $\frac{r_i - \bar{r}}{\sigma_r}$。二元情形提供了清晰的解析基线，并展示了 z 分数归一化在使有效信号尺度与 $\sqrt{\mu}$（$1-\mu$）成正比方面的关键作用。

变体	描述
1	高斯（强度 = 0.45）和旋转（1°的倍数） + 变体文本 / 向图像添加文本
2	高斯（强度 = 0.45）或旋转（30°的倍数） +变体文本 / 向图像添加文本
3	高斯（强度 = 0.3）或旋转（45°的倍数） + 变体文本 / 向图像添加文本
4	高斯（强度 = 0.3）或旋转（90° 的倍数）
5	变体文本
6	原始查询 / 变体文本
7	1 个思考步骤
8	2 个思考步骤
9	3 个思考步骤

表 C.1：由难到易的变体。

表 C.2：不同基础模型的性能比较。

模型	数学	物理学	化学	生物学	Avg.
Qwen3-VL-4B					
Qwen3-VL-4B-Instruct	65.5	55.5	58.5	56.1	58.9
+GRPO	75.5	68.5	71.0	68.6	70.9
+DIVA-GRPO（我们的方法）	78.3	70.7	72.7	71.9	73.4
改进。	+2.8	+2.2	+1.7	+3.3	+2.5
Qwen3-VL-8B					
Qwen3-VL-8B-Instruct	71.5	64.5	66.5	65.5	67.0
+GRPO	77.5	72.5	74.0	73.2	74.3
+DIVA-GRPO（我们的方法）	82.2	74.3	78.2	79.3	78.5
提升。	+4.7	+1.8	+4.2	+6.1	+4.2

结论。这项理论分析严格论证了我们方法的核心动机。当批次优势通过 z 分数归一化计算，且奖励为二元奖励 $\{0, R_{\max}\}$ 时，更新在任意固定参考方向 v 上的期望投影幅度与 $\sqrt{\mu(1-\mu)}$ 成正比，并在 $\mu = \frac{1}{2}$ 时达到最大。换言之，正确样本与错误样本数量相等的平衡批次，能够为优化提供最强的期望方向信号。这一结果支持了我们的设计选择：自适应生成问题变体，使正确样本与错误样本的比例保持在约 50/50，从而确保梯度更新最高效地与最优更新方向对齐，进而最大限度地促进模型改进。

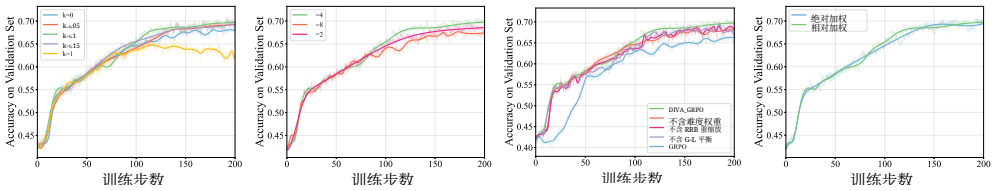
D 补充实验

D.1 参数 k

在与消融研究相同的实验设置下，我们评估了所提方法在参数 k 取不同值时的性能。如图 D.1a 所示，当 $k \approx 0.1$ 时，模型的表现最佳，准确率差异峰值仅为 0.5%。当 $k = 0$（即禁用难度加权）或 $k = 1$（即难度加权过高）时，由于样本难度未能区分，算法在训练早期阶段的表现相近。然而，随着训练的推进，样本难度的差异愈发明显，过于极端的加权会导致性能下降或训练不稳定。当 $k = 0.1$ 时，施加的最大权重约为原始优势的 1.5 倍，而最大降幅约为原始值的 0.67 倍。根据加权公式，我们推导得出 $k = \ln(1.5)$ 最大值$\left(D_q^{(i)} - \bar{D}_q\right)$。

D.2 参数 η 与样本采样

在与消融研究相同的实验设置下，我们评估了所提出方法在参数 η取不同值时的表现。参数 η 表示答对或答错问题时难度评估的变化幅度。具体而言，当 $\eta = 2$时，样本需要连续答对（或答错）四次，才能将其难度等级更新为最简单（或最困难）等级。当 $\eta = 4$时，更新需要答对（或答错）两次；当 $\eta = 8$时，答对（或答错）一次即可更新。



(a) 在 k 的不同取值下 (b) 在 η的不同取值下 (c) 组件消融研究 (d) 组件消融研究

图 D.1：不同超参数对验证集上模型性能的影响。

表 D.3：Qwen-2.5-VL-7B 上不同外部模型的性能比较。

模型	数学	物理学	化学	生物学	Avg.
Qwen-2.5-VL-7B					
+ Qwen-Plus	78.1	61.5	69.1	70.9	69.8
+ GPT-o3	78.3	62.2	69.6	70.7	70.2
改进。	+0.2	+0.7	+0.5	-0.2	+0.3

我们观察到，难度评估变化过快时，模型固有的随机性可能导致难度评估不准确。相反，难度评估变化过慢时，模型性能的提升可能无法反映在难度评估中。如图 D.1b 所示，当 $\eta = 4$时，模型取得最佳表现。根据难度更新公式，我们建议将 η 设为 $\frac{D_{\max}-D_{\min}}{2}$。

此外，在样本采样方面，我们设置了共九种难度级别变体，详见表 C.1。采样时，我们使用正态分布，从与目标难度相邻的难度等级中进行采样。

D.3 不同规模的模型

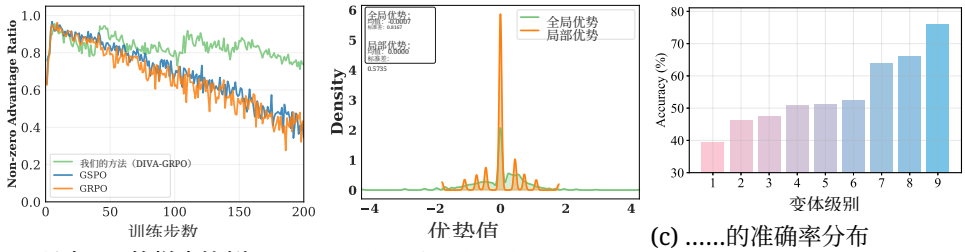
如表 C.2 所示，我们在不同规模和架构的模型上评估了 GRPO 和我们的方法 DIVA-GRPO 的性能，实验设置与前文所述一致。在 Qwen3-VL-4B-Instruct 和 Qwen3-VL-8B-Instruct 上的实验结果表明，与基础模型相比，我们的方法分别提升了 14.5 和 11.5。此外，与 GRPO 相比，我们的方法平均提升了 2.5 和 4.2。这些结果表明，我们的方法适用于各种规模和架构的模型。

D.4 使 用不同外部模型生成 DELS

如表 D.3 所示，我们使用 Qwen-plus 和 GPT-o3 生成变体和思考步骤，并在相同的训练条件下测试其表现。结果表明，使用较强或较弱的外部模型对最终模型训练性能的影响微乎其微，GPT-o3 的准确率仅比 Qwen-plus 高 0.3。这可能归因于我们采用了结果监督：我们只选取能够正确给出最终答案的样本进行训练，而非使用所有样本进行训练。此外，我们使用 Qwen-plus 验证生成的问题变体与原始问题的一致性。我们为实验抽取了 1,000 个问题变体，发现其中 98.8% 与原始问题一致。这进一步证明了我们方法的可靠性。实验中使用的提示见第 G 节。

D.5 绝对难度加权与相对难度加权

如第 3.3 节所述，我们的方法采用相对难度加权，而非绝对难度加权。这是因为问题及其变体的难度在两个尺寸上存在差异。例如，难度为 1 的问题加上难度为 8 的变体，无法直接与难度为 9 的问题进行比较。难度增加的变体仅仅是添加了一些视觉噪声，而具有固有难度的问题则因其复杂的逻辑而具有挑战性。因此，直观地应用绝对难度加权并不合理。然而，对于同一原始问题的变体，其难度排序是明确的：图像噪声变体 $>$ 原始问题 $>$ 变体



(a) 具有……的样本比例非零优势

(b) 来源全局与局部优势-分布

(c) ……的准确率分布不同难度的样本等级。

图 D.2：DIVA-GRPO 的有效性分析。

其中添加了思考步骤。因此，我们采用相对难度加权，即可以比较同一问题内部的难度。此外，如图 D.1d 所示，我们测试了绝对难度加权和相对难度加权，发现二者对模型最终性能的影响都很小。然而，从逻辑一致性的角度来看，我们最终选择了相对难度加权。

D.6 方法有效性分析

非零优势样本比例。 我们的方法动态估计问题难度，并采样适当难度的变体，从而有效缓解奖励稀疏性并提高样本效率。如图 D.2a 所示，GRPO 和 GSPO 中非零优势样本的比例在训练初期迅速上升，表明训练准确率快速提高。然而，随着模型变强，大多数问题变得简单，导致非零优势样本比例急剧下降，并限制了后期优化。相比之下，我们的方法自适应地调整问题难度，使非零优势样本比例保持稳定，确保模型持续遇到具有挑战性的变体，维持平衡的奖励分布，并在整个过程中保持有效训练。

来源局部与全局优势分析。 我们首先对局部优势与全局优势在每个组内应用 z 分数归一化。由于与局部组相比，全局组具有更高的稀疏性和方差，即使经过标准归一化后，优势范围 max(adv) – min(adv) 在不同组之间仍可能存在显著差异，导致范围较大的组占据主导（见图 D.2b）。为缓解这一影响，我们进一步分别对局部优势与全局优势应用批归一化，然后在每个批次内进行标准归一化。这一额外的归一化有效减小了两种分布之间的差异。

变体难度分析。 为验证我们的方法生成的变体难度是否符合预期，我们从 MMK12_test 数据集中随机抽取 500 个样本。针对每个样本，我们在每个难度等级下各生成一个变体，具体的变体设置详见表 C.1。每个难度等级下共计 500 个样本。对于每个样本，我们使用 Qwen-2.5-VL-7B 进行 10 次 rollout，并计算最终准确率。最终准确率的分布如图 D.2c 所示。尽管准确率的提升并非完全呈线性，但总体趋势表明，我们生成的样本变体呈现出预期的规律：较简单的变体准确率更高，而较困难的变体准确率更低。这证明了我们方法的有效性。

表 F.4：MMK12 不同学科的性能比较。粗体表示 7B 模型中的最佳性能，下划线标示整体最佳性能。

模型	数学	物理学	化学	生物学	Avg.
闭源模型					
Claude3.7-Sonnet	57.4	53.4	55.4	55.0	55.3
GPT-4o	55.8	41.2	47.0	55.4	49.9
o1	<u>81.6</u>	<u>68.8</u>	<u>71.4</u>	<u>74.0</u>	<u>73.9</u>
Gemini2-flash	76.8	53.6	64.6	66.0	65.2
开源通用模型					
InternVL2.5-VL-8B	46.8	35.0	50.0	50.8	45.6
Qwen-2.5-VL-7B	58.4	45.4	56.4	54.0	53.6
InternVL2.5-VL-38B	61.6	49.8	60.4	60.0	58.0
Qwen-2.5-VL-32B	71.6	59.4	69.6	66.6	66.8
InternVL2.5-VL-78B	59.8	53.2	68.0	65.2	61.6
Qwen-2.5-VL-72B	75.6	64.8	69.6	72.0	70.5
开源推理模型					
InternVL2.5-8B-MPO	26.6	25.0	42.4	44.0	34.5
InternVL2.5-38B-MPO	41.4	42.8	55.8	53.2	48.3
QVQ-72B-Preview	61.4	57.4	62.6	64.4	61.5
Adora	63.6	50.6	59.0	59.0	58.1
R1-Onevision	44.8	33.8	39.8	40.8	39.8
OpenVLThinker	63.0	53.8	60.6	65.0	60.6
MM-Eureka-7B	71.2	56.2	65.2	65.2	64.5
DIVA-GRPO-7B（我们的方法）	**78.3**	**62.2**	**69.6**	**70.7**	**70.2**

E 问题难度对 GRPO 优化的影响

为进一步验证向模型提供中等难度的问题更有利于通过 GRPO 进行优化，我们开展了一项新的实验。从 MMK12 训练数据中抽取两组各含 5,000 个样本的数据：第一组由随机抽取的实例组成，第二组则由中等难度的实例组成。我们使用这两组数据，在 rollout 参数为 $k = 5$ 的设置下，通过标准 GRPO训练模型。如图 E.3 所示，使用中等难度数据进行训练，能够使模型优化得更快、效果更好。

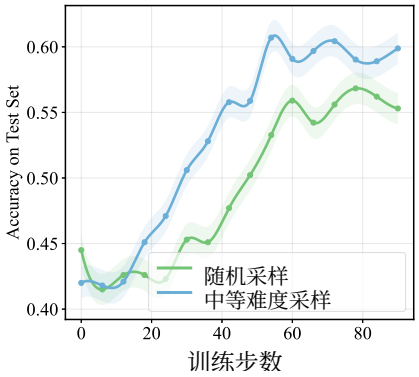


图 E.3：随机抽样问题与中等难度问题上的 GRPO 优化对比。

F MMK12 上的跨学科表现

在 MMK12 数据集上，我们观察到一个有趣的现象：尽管我们的模型主要针对数学题进行训练，但它在物理学、化学和生物学方面的推理能力也有所提升。实验结果汇总于表 F.4。我们的模型在全部四个学科中均取得了 7B 模型中的最佳表现，甚至超过了许多闭源商业模型和开源通用大型模型。其他模型的对比结果见 Meng 等人（2025）。

G 实验中使用的提示

本节详细介绍我们流程中使用的提示词，包括变体生成、推理和验证。

G.1 第一部分：变体生成

目的：生成 5 个不同的问题变体，同时保持正确答案完全一致。变体必须包含词元 <图像> ，并使用 <variant*>...</variant*> 标签包裹。

提示词 1：变体生成模板

```
针对以下问题生成 5 个不同的变体，要求：1. 保持与原始问题完全相同的正确答案。2. 使用显
```

预期输出（示例）：

```
<变体1> [.....文本.....  <图像>  .....] </变体1><变体2> [.....文本
```

G.2 第二部分：推理与验证

提示词 2：逐步推理（P R O M P T_F O R M A T_T H I N K_S T E P）

Pur姿势：生成一个逐步推理链（3–5 步），详细描述图示观察，并以方框框出最终答案，g wit并以方框框出最终答案。

```
你是一位数学家、统计学家和几何学家。下面我将给出一道数学问题及其配图。请仔细观察图像
```

预期输出（示例）：

```
<步骤1>  ... 观察 ...  </step1><步骤2>  ... 推理 ...
```

提示词 3：答案比较

目的：比较两个数值答案（answer1、answer2），忽略格式差异。如果两者相等，则返回<answer>True</answer>，否则返回 False。

以忽略任何格式差异的方式比较以下两个答案：答案 1：{result['answer1']} 答案 2：{result['

提示4：反思与修订 (PROMPT_FORMAT_THINK_STEP_2)

目的： 在收到正确性反馈后完善推理步骤，但不要直接在步骤中填入正确答案，也不要从中反向推导答案。

你是一位数学家、统计学家和几何学家。请仔细观察图像中的细节。你已经在上文给出了推理步

H 训练数据集案例研究

H.1 案例1

问题： 下方的图 <图像> 展示了同一笛卡尔坐标系中函数 $y = x^2 - 1, y = x^2 + 6x + 8,$ $y = x^2 - 6x + 8, y = x^2 - 12x + 35$ 的图像。最有可能表示 $y = x^2 - 6x + 8$ 的图像是 ___。

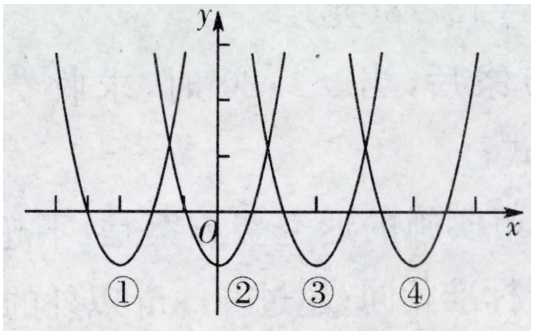


图 H.4：原始问题

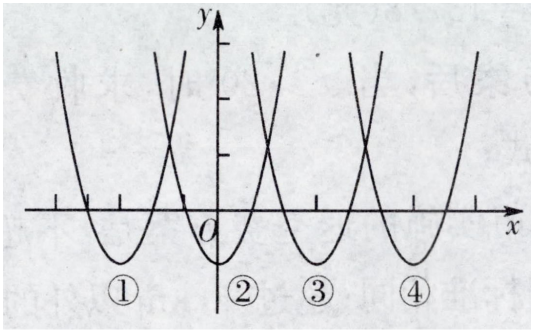
困难变体

文本与图像变体： 将原始问题描述和问题直接嵌入图像中，并将提示词文本替换为以下固定指令：

如 <图像>所示，请根据图回答问题。

图像变体：

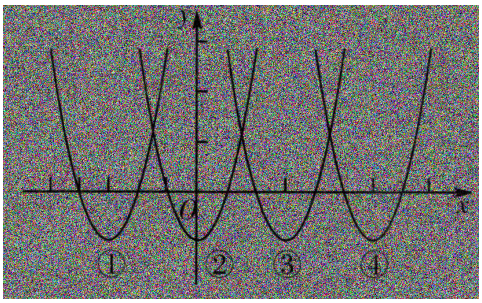
- **椒盐噪声：** 图像受到椒盐噪声干扰，以增加物体识别的复杂度。
- **模糊：** 图像经过模糊处理，使人更难聚焦于特定细节。
- **高斯噪声：** 图像受到高斯噪声的干扰。



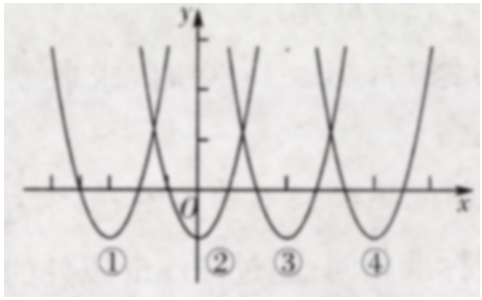
The figure <image> below shows the graphs of the functions: $$y=x^{2}-1$$, $$y=x^{2}+6x+8$$, $$y=x^{2}-6x+8$$, $$y=x^{2}-12x+35$$ in the same Cartesian coordinate system. The most likely graph for $$y=x^{2}-6x+8$$ is ___.

图 H.5：视觉主导问题

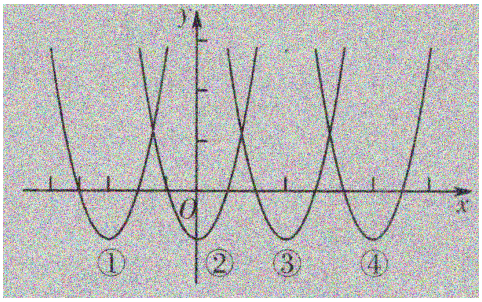
• **旋转：**图像经过旋转，增加了空间推理的难度。



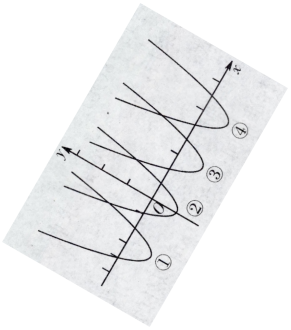
(a) 椒盐噪声



(b) 模糊



(c) 高斯噪声



(d) 旋转

图 H.6：图像变体：椒盐噪声、模糊、高斯噪声和旋转。

文本变体：对原始问题进行语义等价的改写，同时保留其含义和正确答案。

问题变体 1：

在图 <图像>所示的笛卡尔平面中，绘有四条不同的抛物线，每条分别对应以下方程之一： $y = x^2 - 1$、 $y = x^2 + 6x + 8$、 $y = x^2 - 6x + 8$和 $y = x^2 - 12x + 35$。你的挑战是辨别并指出由方程 $y = x^2 - 6x + 8$生成的那条特定曲线。

问题变体 2：

在图 <图像>所提供的图表中，请识别哪一个对应于二次函数 $y = x^2 - 6x + 8$。与其他函数——$y = x^2 - 1$、 $y = x^2 + 6x + 8$和 $y = x^2 - 12x + 35$——相比，该函数呈现出独特的顶点和开口方向。你的任务是选出准确表示 $y = x^2 - 6x + 8$的图表。

较简单变体

思考步骤：

- **步骤 1：** 图像中展示了四个抛物线图像，分别标记为 ①、 ②、 ③ 和 ④。每个图像都表示一个形如 $y = ax^2 + bx + c$ 的二次函数。这些抛物线的总体特征包括它们的开口方向（均向上，表明首项系数为正）以及各自的顶点。
- **步骤 2：** 要确定哪个图像对应函数 $y = x^2 - 6x + 8$，我们首先将该函数改写为顶点式。配方法：

$$y = x^2 - 6x + 8 = (x^2 - 6x + 9) - 9 + 8 = (x - 3)^2 - 1$$

这表明抛物线的顶点位于点 $(3, -1)$。

- **步骤 3：** 接下来，我们分析每个图表中顶点的位置：
  - 图表 ①的顶点位于 y 轴左侧。- 图表 ② 的顶点位于原点或非常接近原点。- 根据顶点式，图表 ③ 的顶点位于 $(3, -1)$。- 图表 ④ 的顶点位于 y 轴右侧，但不在 $(3, -1)$。
- **步骤 4：** 因此，顶点位于 $(3, -1)$ 的图表是图表 ③。根据分析，函数 $y = x^2 - 6x + 8$ 对应标记为 ③ 的图表，因为只有这个图表的顶点坐标正确。

H.2 案例 2

问题：

<图像> 减去所有棕色立方体。减去所有蓝色圆柱体。还剩下多少个立方体?

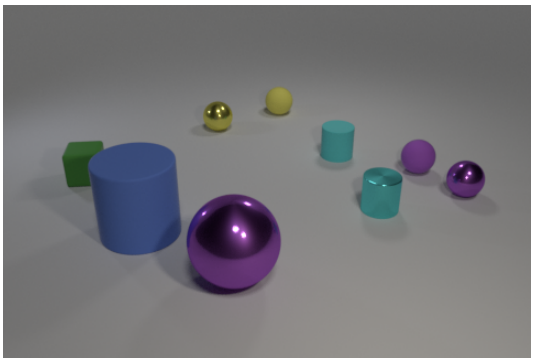


图 H.7：原始问题

困难变体

Text 和图像变体： 将原始问题描述和问题直接嵌入 into 图像中，并将提示词文本替换为以下固定指令：

如 <图像>所示，请根据图回答问题。

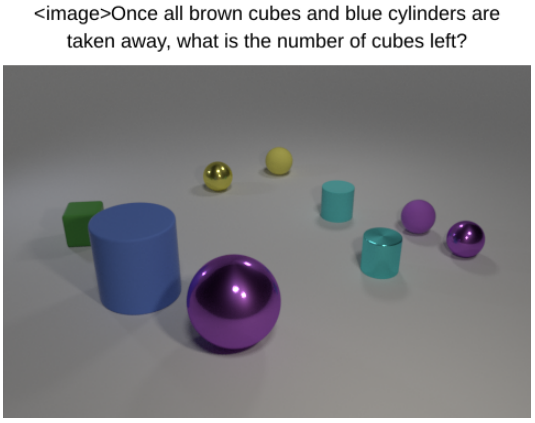
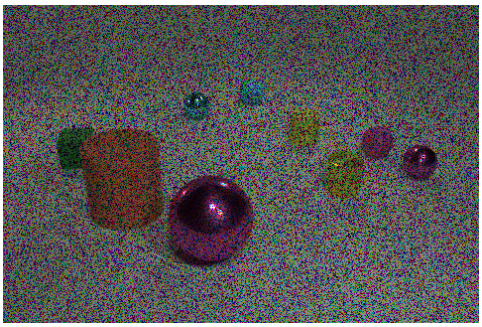


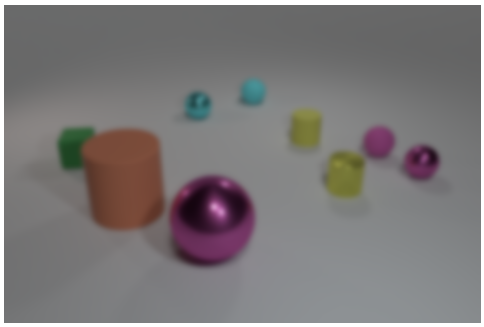
图 H.8：视觉主导问题

图像变体：

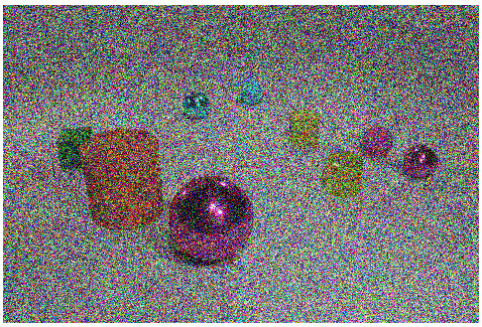
- **椒盐噪声：**图像中添加了椒盐噪声，以增加识别物体的难度。
- **模糊：**图像经过模糊处理，更难聚焦于具体细节。
- **高斯噪声：**图像受到高斯噪声扰动。
- **旋转：** 图像经过旋转，增加了空间推理的难度。



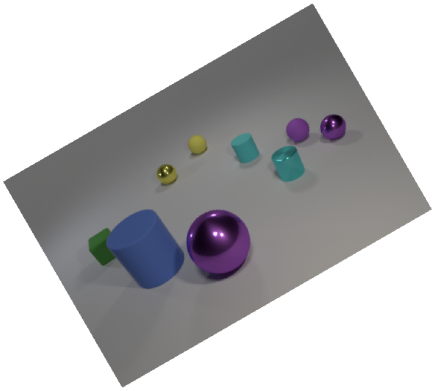
(a) 椒盐噪声



(b) 模糊



(c) 高斯噪声



(d) 旋转

图 H.9：图像变体：椒盐噪声、模糊、高斯噪声和旋转。

文本变体： 与原始问题语义等价的改写，同时保留其含义和正确答案。

问题变体 1：

<图像> 移走所有棕色立方体和蓝色圆柱体后，还剩多少个立方体?

问题变体 2：

<图像> 移除所有棕色立方体和所有蓝色圆柱体。还剩多少个立方体?

较简单变体

思考步骤：

• **步骤 1：** 最左侧物体：一个单独的立方体，明显为绿色。– 紧挨该立方体右侧：一个大型深蓝色圆柱体。– 更靠右：两个较小的圆柱体，呈较浅的青蓝色。– 出现若干球体（紫色、金色、黄色），但它们不是立方体或圆柱体。– 场景中没有任何物体呈现棕色。

因此，仅存在一个立方体（绿色），并且有三个蓝色系圆柱体（一个深色，两个青色）。

• **步骤 2：** 结合观察到的清单解读问题说明。– “减去所有棕色立方体。”由于不存在任何棕色立方体，此操作不影响立方体数量。– “减去所有蓝色圆柱体。”

• **步骤 3：** 尽管有三个圆柱体满足颜色条件，但移除圆柱体只会改变圆柱体清单；立方体保持不变。执行规定的减法。– 在减去零个棕色立方体后，立方体清单与步骤 1 中记录的相同。

• **步骤 4：** 移除三个蓝色圆柱体仅影响圆柱体类别；立方体类别与之前完全相同。总结两项操作的效果。– 指定的减法均未改变原始立方体数量，因此最终立方体计数等于步骤 1 中记录的初始立方体计数。

I 测试数据集案例研究

I.1 案例 1

问题。 如图所示， $\triangle ABC \cong \triangle ADE$。根据图可知， $\angle B = 70°$、$\angle C = 30°$ 和 $\angle DAC = 35°$。求 $\angle EAC$。

选项： A: $40°$ B: $45°$ C: $35°$ D: $25°$

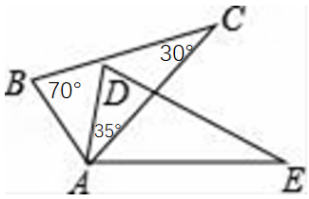


图 I.10：几何题：MathVerse 7。

我们的模型 — 最终答案：B: $45°$ $\boxed{}$

完整推导（外部，逐步）。

1. 在 $\triangle ABC$ 中，三角形内角和给出 $\angle BAC = 180° - \angle B - \angle C = 180° - 70° - 30° = 80°$。2. 由全等 $\triangle ABC \cong \triangle ADE$，对应角相等；因此 $\angle DAE = \angle BAC$ =80°。3. 在顶点 A 处，角的分解得到 $\angle DAE = \angle DAC + \angle CAE$。给定 $\angle DAC = 35°$，我们有 $\angle CAE = 80° - 35° = 45°$。4. 因此 $\angle EAC = \angle CAE = 45°$，对应于选项 B。

其他模型（基线模型）— 报告的答案：A: 40° $\boxed{}$

报告的推导（根据基线模型输出重建）。

1. 基线模型断言，根据全等关系 $\angle DAE = \angle BAC$（这一断言在几何形式上是正确的）。2. 随后，基线模型（错误地）将 $\angle BAC$ 读作或视为 $35°$，并将 $\angle CAD$（或 $\angle DAC$）读作 $30°$（这是对图中标签／给定数据的误读）。3. 基线模型利用（错误的）等式 $\angle BAC = \angle CAD + \angle EAC$，计算得出 $\angle EAC$ =$35° - 30° = 5°$。4. 随后，基线模型又提出另一种算术尝试：将某种组合视为“外角”或角度之和，计算得出 $35° + 30° = 65°$，然后（在没有有效几何论证的情况下）进行 $65° - 70°$，并报告 $40°$ 为最终选项。

明确的错误分析。

• **标记误读错误**。 基线模型将 $\angle BAC = 35°$ 设为该值的步骤与已知数值 $\angle B = 70°$ 和 $\angle C = 30°$不一致，因为这两个数值推出的结果应为 $\angle BAC = 80°$。错误识别哪个标记角对应于 $\angle BAC$ 是根本原因。

• **算术错误及缺乏论证的运算**。 基线模型混淆了不同的角和运算（内角与外角），却没有几何论证；此外，它执行的算术运算（例如报告文本中的 $65° - 70° \to 40°$）既不符合代数规则，在此语境下也缺乏几何意义。

• **逻辑链断裂**。 由于基线模型的前提有误，后续推导（例如对这些错误角度进行相减或求和）并不能从有效的几何事实（三角形内角和、全等对应或正确的角的分解）中得出。

总结——为什么在这种情况下我们的模型更可取。

• **准确的视觉解读：**我们的模型正确识别了标签，并从图中读出 $\angle B$、$\angle C$ 和 $\angle DAC$，避免了基线模型出现的标签混淆。
• **正确且可验证的推理：**每一步都遵循基本的几何规则（三角形内角和、全等 $\Rightarrow$ 对应角相等，以及顶点处角的分解），算术计算也简单明了、便于核验。
• **透明的呈现：**上述推导已明确呈现，可纳入附录，以便审稿人追溯推理过程并确认其正确性。

I.2 案例 2

问题。用区间表示法写出数轴上所表示的数的集合。



图 I.11：区间表示法问题：MathVerse 474。

我们的模型——最终答案：$(-2, \boxed{-1]}$

完整推导（外部，逐步进行）。

1. 查看数轴上标出的端点：一个端点位于 -2，另一个位于 -1。2. 根据标记样式判断是否包含：

- -2 标记为一个空心 圆圈 $\Rightarrow -2$，不 包含。
- -1 用实心（填充）圆圈标记 $\Rightarrow -1$ 被包含。

3. 将包含/排除转换为区间表示法：
• 排除 -2 时，左侧使用左圆括号“(”。
• 包含 -1 时，右侧使用方括号“]”。4. 因此，该区间为 $(-2,\ -1]$。5. 用不等式形式表示为 $-2 < x \leq -1$。

其他模型（基线模型）——报告的答案：$(-2,\ -1]$

报告的推导（根据基线模型的输出重构）。

1. 基线模型将端点识别为 -2 和 -1。2. 它（错误地）将两个端点标记都解释为空心圆圈，并得出两个端点都不包含在内的结论。3. 基于这一（错误的）解释，它在区间两侧都使用圆括号，并报告为 $(-2,\ -1)$。

明确的错误分析。

• **视觉误读错误**。 基线模型的核心错误是把 -1 处的标记看成空心，而实际上它是实心的（闭端点）。这一处视觉识别／映射错误，就会把右端点处的“$\leq$”改成“$<$”，从而改变括号类型。

• **记号上的后果**。 区间表示法对端点包含情况很敏感：将闭端点误认为开端点，会把方括号“]”变成圆括号“)”，从而得到不同的集合。这里的错误会导致集合是否包含点 -1 发生变化。

• **后续影响**。 由于基线模型误读了标记，因此它给出的任何不等式形式陈述或集合归属陈述（例如 $-2 < x < -1$），对于等于 -1 的点都是错误的。

总结——为何在此情形下我们的模型更可取。

• **正确的视觉到符号映射：** 我们的模型能正确地将空心与实心端点标记分别对应到不包含与包含，因此会使用恰当的圆括号／方括号组合。
• **步骤清晰且可验证：** 每一步都是简单、可检查的规则：确定端点、读取标记样式、转换为区间表示法，并可选择给出等价不等式。这种透明性让审阅者能够轻松验证。
• **精确的最终表达式：** 答案 $(-2,\ -1]$ 清晰明确地表示了大于 -2 且小于或等于 -1的实数集合，与图中 -1 为实心标记相符。

I.3 案例 3

问题。 根据下图，求直线的斜率。

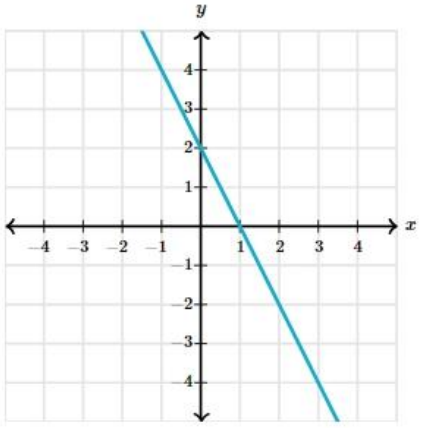


图 I.12：直线斜率问题：MathVerse 591。

我们的模型 — 最终答案： -1 $\boxed{}$

完整推导（外部，分步进行）。

1. 经过两点（x_1， y_1）和（x_2， y_2）的直线，其斜率由下式给出：

$$\text{slope} = \frac{y_2 - y_1}{x_2 - x_1}.$$

2. 根据图表，找出直线上的两个点：（-1， 2）和（1，0）。
3. 将这些点代入斜率公式：

$$\text{slope} = \frac{0 - 2}{1 - (-1)} = \frac{-2}{2} = -1.$$

4. 因此，该直线的斜率为 -1。

其他模型（基线模型）——报告的答案： -2 $\boxed{}$

报告的推导（根据基线模型输出重建）。

1. 基线模型将直线上的两个点识别为 (1,2) 和 (0,4)。2. 使用斜率公式：

$$\text{slope} = \frac{2 - 4}{1 - 0} = \frac{-2}{1} = -2.$$

3. 基线模型报告最终斜率为 -2。

明确的错误分析。

• **错误的点的选择：** 基线模型误读了直线上点的坐标。使用 (1,2) 和 (0,4) 与图中的实际直线不符。

• **由此导致的错误斜率：** 由于点的选择错误，计算出的斜率 -2 与直线的真实斜率 -1不匹配。

• **逻辑链断裂：** 后续所有推理在数学上都与所选点一致，但前提（点的选择）存在错误，导致最终答案不正确。

总结——为什么我们的模型在这种情况下更具优势。

• **准确的视觉解读：** 我们的模型正确识别出图中直线上的实际点。

• **正确的推理：** 逐步应用斜率公式可直接得到正确答案 -1。

• **透明的推导：** 推理过程得到完整呈现，使审阅者能够轻松验证其正确性以及与图的一致性。

J 无法产生优势的案例

J.1 始终回答正确的案例

J.1.1 案例 1

问题。 如图所示，点 P 是 $\triangle ABC$的内角 $\angle ABC$ 与外角 $\angle ACD$ 的角平分线的交点。点 P 到直线 AC的距离为6cm。点 P 到直线 AB的距离是多少?

答案： $\boxed{6}$

J.1.2 案例 2

问题。 如图所示，在菱形 $ABCD$ 中，对角线 AC 和 BD 相交于点 O， $OE \parallel DC$ 与 BC 相交于点 E。若 $AD = 8$ 厘米，则 OE 的长度为 厘米。 ____

答案： 4 $\boxed{}$

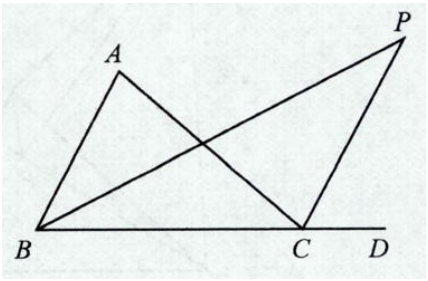


图 J.13：涉及角平分线的几何题。

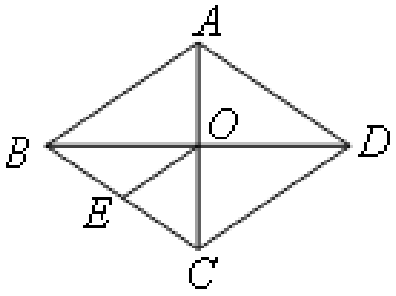


图 J.14：涉及菱形的几何题。

J.1.3 案例 3

问题。如图所示，在 $\triangle ABC$ 中，点 D 和 E 分别位于 AB 和 AC 上，且 $DE \parallel BC$。如果 $AD = 1$ 和 $AB = 4$，那么 $\frac{DE}{BC} =$ 。

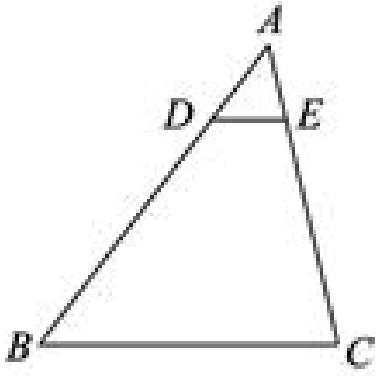


图 J.15：涉及相似三角形的几何题。

答案： $1\ \boxed{\frac{4}{-}}$

J.1.4 案例 4

问题。对于任意非零实数 a、b，如果 $a \otimes b$ 的运算规则由图中所示的流程图表示，那么 $3 \otimes 2 =$ 。

答案： $2\ \boxed{\ }$

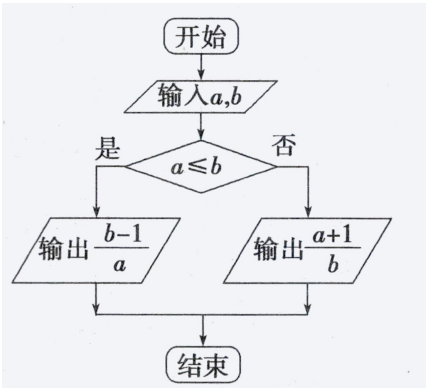


图 J.16：流程图运算问题。

J.2 一贯回答错误的案例

J.2.1 案例 1

问题。图 A 是一张直方图，显示了某学校教职员工月收入的频数分布。已知从左到右第一个组的频数为 80。在样本中，月收入（单位：元）落在区间 $[1000, 1500)$,$[1500, 2000)$, $[2000, 2500)$, $[2500, 3000)$, $[3000, 3500)$, $[3500, 4000]$ 内的频数分别记为 $A_1, A_2, \ldots, A_6$。图 B 是用于估算该校教职员工平均月收入的算法流程图。 S的输出是什么？（请用数字作答。）

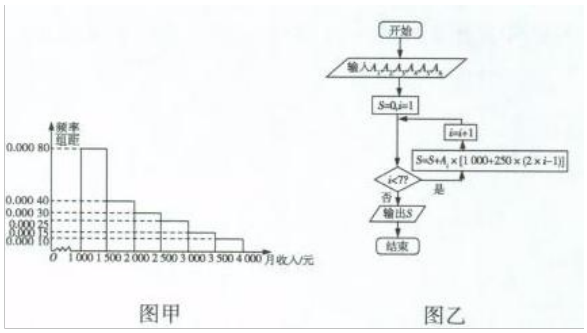


图 J.17：复杂统计与流程图问题。

答案： 1962.5

J.2.2 案例 2

问题。 要加工一个形状如图所示的零件，请根据标注的尺寸（单位：毫米）计算倾斜角 α 的度数。（使用计算器，精确到 1 秒）。

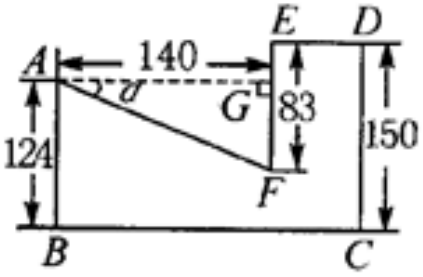


图 J.18：工业几何计算。

答案： $22^\circ\, 9'\, 12''$

J.2.3 案例 3

问题。 黑白六边形地砖按图示图案排列；那么第 n 种图案中的白色地砖数量为 。

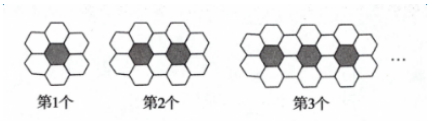


图 J.19：模式识别数列问题。

答案： $\boxed{a_n = 6n - 2(n-1) = 4n + 2}$

J.2.4 案例 4

Pro题。 如图所示，有一个由半球和正方形组成的几何立体 e pyr位于中间，如三视图所示。该立体的体积为。

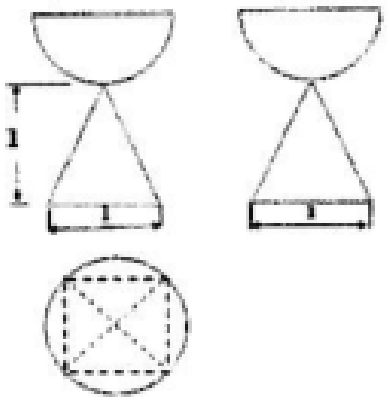


图 J.20：根据三视图绘制的立体几何图。

答案： $\boxed{3 + \frac{\sqrt{2}}{6}\pi}$

J.3 案例分析

分析持续得出正确或错误回答的样本，可以发现明显的规律。

持续正确的情形：我们发现，对于图像清晰、逻辑简单且结果定义明确、易于表达和验证的问题，我们的方法并未带来显著优势。这主要是因为视觉信息毫无歧义，而且对于基础视觉语言模型而言，解题路径足够直接。此外，最终答案通常是简单的整数或分数，因此验证起来更容易。

持续答错的案例：相反，对于涉及高度复杂图像内容，或结果难以描述或验证的问题，即使提供了推理步骤，仍然很难得出正确答案。这种难度源于仅凭推理推断正确答案本身就具有内在复杂性，而中间结果和最终结果难以验证这一点又进一步加剧了这种难度，从而使优化任务变得更加复杂。

K 生成式人工智能使用披露

在本研究中，我们使用大语言模型辅助完善和润色文稿。作者确认，本文的数据分析、方法开发和理论贡献均未借助生成式人工智能工具生成内容或提供研究协助。包括代码和数据在内的研究各个方面均由作者独立完成。

L 局限性

尽管 DIVA-GRPO 有效缓解了奖励稀疏性，并提升了在多模态推理任务上的表现，但仍存在一些局限。首先，对于需要中间推理步骤的困难问题，其变体生成过程依赖外部模型，可能引入潜在偏差。其次，难度加权缩放涉及需要谨慎调优的超参数，降低了开箱即用的便利性。最后，DIVA-GRPO 主要在多模态推理任务上进行评估，其向非多模态任务或奖励结构根本不同的任务的泛化能力仍有待探索。