arXiv:2603.09478v1 [cs.MM] 10 Mar 2026

MORE-R1：通过结合强化学习的逐步推理，引导大型视觉语言模型进行多模态对象-实体关系抽取

Xiang Yuan[1],Xu Chu[1],Xinrong Chen[1],Haochen Li[2],Zonghong Dai[2], Hongcheng Fan[2],Xiaoyue Yuan[3],Weiping Li[1],和 Tong Mo[1](✉)

[1] 北京大学软件与微电子学院，中国北京，xiangyuan@stu.pku.edu.cn, motong@ss.pku.edu.cn[2] AlignBase，中国北京[3] 上海市市场监督管理局信息应用研究中心，中国上海

摘要。 多模态对象-实体关系抽取（MORE）是信息抽取研究中的一项具有挑战性的任务，旨在识别视觉对象与文本实体之间的关系，需要复杂的多模态理解和跨模态推理能力。现有方法主要基于分类或基于生成，缺乏推理能力，难以应对 MORE 任务中的复杂抽取场景，且可扩展性有限、中间推理过程不透明。为应对这些挑战，我们提出了 MORE-R1，这是一种新型模型，通过引入结合强化学习（RL）的显式逐步推理，使大型视觉语言模型（大型视觉语言模型）能够有效完成 MORE 任务。MORE-R1采用两阶段训练流程，包括先进行监督微调（SFT）的冷启动训练阶段，随后进行用于优化推理能力的强化学习阶段。在初始阶段，我们设计了一种高效方法，自动构建高质量的监督微调数据集，其中包含针对 MORE 任务定制的细粒度逐步推理，从而使模型能够学习有效的推理范式。在后续阶段，我们采用组相对策略优化（GRPO）强化学习算法和渐进式样本混合策略，以稳定训练，并进一步增强模型处理困难样本的推理能力。在 MORE基准上进行的全面实验表明，MORE-R1取得了最先进的性能，显著优于基线方法。

关键词： 多模态关系抽取 · 大视觉语言模型 · 强化学习 · 大型推理模型。

1 引言

关系抽取（RE）是信息抽取（IE）研究中的一项基础任务，旨在从非结构化输入中抽取实体之间的关系。在各种与关系抽取相关的任务中，多模态对象-实体关系

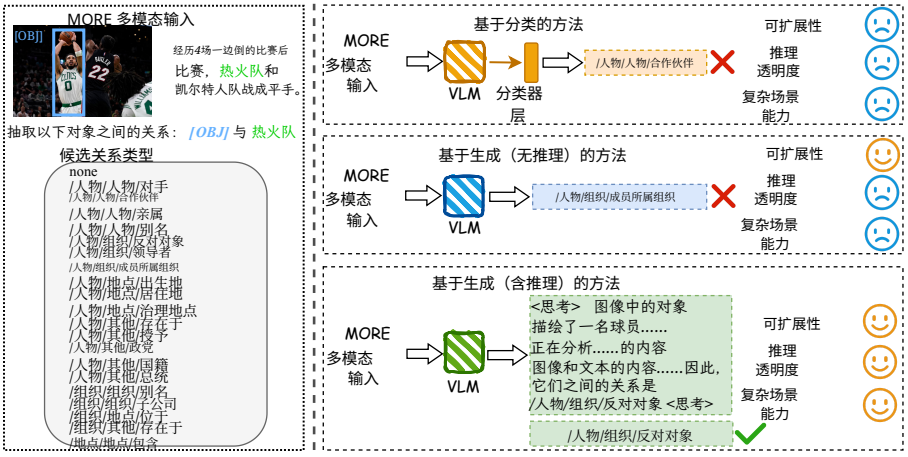


图1. 左侧：MORE任务示例：从候选集合中识别对象与实体之间的关系类型。右侧：MORE任务的不同框架。我们的模型属于基于生成（含推理）的方法。

抽取（MORE） [10] 是一项新提出的关系抽取任务，更能反映现实世界场景，也契合了互联网日益增长的多模态信息处理需求（主要是图像和文本） [25]。MORE任务旨在抽取图像中给定对象与文本中给定实体之间的关系，要求模型同等关注两种模态，建立对象与实体之间的对应关系，并执行跨模态关系推理。如图1所示，为了准确抽取图像中蓝色边界框所框选对象与文本中实体“热火队”之间的关系，模型需要理解该对象代表文本中提到的“凯尔特人队”的一名篮球运动员，同时理解文本中的实体“热火队”是一个篮球队名称，并对应于图像中身穿黑色球衣、号码为“22”的球队成员。模型需要推断出这两支球队彼此对抗，才能准确预测“相反于”关系。因此，MORE任务对模型复杂多模态推理能力提出了更高要求。它为跨模态检索等各种下游应用 [18] 以及多模态知识图谱构建 [4]提供了重要支持。

现有的大多数MORE工作 [10,7,22,11,17,16]都采用基于分类的方法：它们基于使用BERT文本编码器和基于ViT的视觉编码器的小规模预训练视觉语言模型（VLMs）构建。这些工作主要致力于增强对象与实体的细粒度跨模态表征，随后使用分类器层将特征映射到一组预定义的离散关系类别中。尽管取得了进展，基于分类的方法仍存在两个方面的不足。（1）可扩展性较差。这类方法只能识别属于预定义类别集合的关系。当出现新的关系类型时，模型需要重新设计分类器层并重新训练其

网络参数。（2）无法处理MORE任务中的复杂关系抽取场景。分类方法对不同关系类别标签的语义表达能力和区分能力有限，无法有效理解和区分容易混淆的关系，例如“同伴”和“情侣”。

近年来，大规模预训练视觉语言模型发展迅速。其中，融合视觉编码器与大型语言模型（LLM）的大型视觉语言模型（LVLM），例如 LLaMA3.2-Vision [8] 和 Qwen2.5-VL[1], ，在各种多模态理解与推理任务中取得了令人瞩目的成果。为了使 LVLM 适应 MORE 任务，一种直接的方法是设计任务特定提示，或采用指令微调，直接迫使模型输出关系标签；这属于图1中的基于生成（无推理）的方法。尽管这种方法在一定程度上具有可扩展性，但近期研究 [23,24] 表明，通过提示学习或直接使用关系抽取标签进行微调的大语言模型和 LVLM 在各种关系抽取任务中的表现不佳。这些研究将此局限归因于大语言模型和 LVLM 在预训练及指令微调阶段缺乏关系抽取相关数据，以及离散关系标签语义复杂度较高，导致模型难以有效理解并遵循关系抽取专用指令 [2,24]。因此，这种方法仍不足以应对 MORE 所需的复杂推理。此外，直接生成关系标签无法揭示模型的决策过程，从而损害了可解释性和推理透明度。

近期，OpenAI-o1 [13] 和 Deep- seek-R1 [9] 等先进的大型推理模型，在编程和数学等诸多具有挑战性的任务中展现出卓越的能力。这些模型通过引入更长的思维链（CoT）推理，并利用强化学习（RL）提升大语言模型的推理能力。中间输出使大语言模型的决策过程更加透明，从而得出更可靠的结论 [6]。受这些模型在复杂推理方面取得成功的启发，我们提出 MORE-R1，这是一种两阶段训练模型，旨在引导大型视觉语言模型有效完成 MORE任务。MORE-R1采用先进的 Qwen2.5-VL [1]模型作为大型视觉语言模型的骨干模型，以实现通用多模态理解。我们首先采用监督微调（SFT）进行冷启动训练，使大型视觉语言模型掌握适用于 MORE任务的初步逐步推理模式。随后，我们采用强化学习（RL）优化大型视觉语言模型的推理轨迹，从而提升其整体推理能力。

在初始冷启动训练阶段，我们选取一小部分训练数据，并设计一种方法以自动生成高质量的推理数据。具体而言，我们将MORE任务重构为逐步渐进的推理过程。通过精心设计细粒度的引导指令，我们利用专家模型 [12] 生成推理示例。通过对生成数据进行监督微调，我们引导大型视觉语言模型学习 MORE任务的基本推理范式。在随后的强化学习阶段，我们采用先进的组相对策略优化（GRPO）算法 [21]进行高效的强化学习训练。为稳定训练并提升复杂情形下的推理能力，我们进一步设计了课程式渐进样本混合策略。随着GRPO训练

的推进，每个小批次中困难样本的比例逐渐增加，引导模型逐步掌握困难实例的正确推理路径。我们的主要贡献总结如下：

– 我们提出了 MORE-R1，这是一种以大型视觉语言模型为骨干、采用显式推理的生成式方法，用于 MORE 任务。MORE-R1 整合了两阶段的监督微调+强化学习训练框架，提升了模型的可扩展性、推理透明度和复杂场景能力。– 在阶段1中，我们设计了一种高效的自动数据构建策略，为 MORE 任务生成高质量的逐步推理数据，以用于监督微调，使模型能够有效学习推理范式。– 在阶段2中，我们引入渐进式样本混合策略，逐步提高模型对困难样本的关注度，从而有效稳定训练，并增强模型在挑战性案例上的推理表现。– 据我们所知，我们是首个成功将大型视觉语言模型适配到 MORE 任务的团队。全面的实验表明，MORE-R1 在 MORE基准上取得了最先进的结果。

2 相关工作

2.1 多模态关系抽取

早期的多模态关系抽取研究 [26,14] 将图像视为辅助信息，但仍专注于抽取文本实体之间的关系。MORE [10] 是近期提出的一项任务，对模型的跨模态推理能力提出了更高要求。近期 MORE 任务的研究进展主要聚焦于增强小规模视觉语言模型中的细粒度跨模态交互与多模态融合，并在输出层使用分类头来完成 MORE 任务。MOREformer [10] 通过融入从图像中提取的物体属性特征和深度图特征，增强多模态表征。HVFormer [17] 利用混合专家架构，有效整合多粒度的跨模态交互语义。CGI-MRE [22]借鉴生物信息学，并采用遗传算法挖掘跨模态共享信息。MMIB [7]利用信息瓶颈理论过滤模态噪声，进一步增强跨模态融合。FocalMRE[11] 设计了一种新颖的焦点注意力和门控机制，以增强图像区域交互。REMOTE [16] 提出了一种多层次最优传输机制，以有效融合多模态特征。尽管做出了这些努力，这些方法仍面临第1节中提出的基于分类的方法所带来的挑战。

2.2 大语言模型与大型视觉语言模型的强化学习

近期研究 [21,5,6] 表明，将 RL 应用于大语言模型和大型视觉语言模型的微调，能够显著提升其应对复杂任务的推理能力。针对这些模型的 RL 算法可根据训练样本是否由当前策略模型生成，分为离策略和同策略。

诸如 DPO 的离策略算法 [19]通过使用预先收集的数据，将训练数据与当前更新的策略模型解耦，而无需在强化学习训练期间生成新样本，从而确保较高的样本效率。然而，学习过程可能不稳定，并可能导致分布偏移。

在策略算法将训练数据与策略模型紧密结合，数据需要由当前策略模型实时生成。这确保了训练的稳定性，因为学习目标和采样数据来自同一个策略模型；然而，这会导致样本效率低下且计算成本高昂。PPO [20] 训练奖励模型和价值模型，以估计每次生成的优势，并在受限范围内更新策略模型，从而提高训练稳定性。GRPO [21] 用组内相对优势计算取代价值模型，以指导策略模型更新，进一步简化训练过程并提高计算效率。此外，近期研究 [9] 指出，未经初步冷启动训练便直接对大语言模型应用强化学习，可能导致输出不稳定，例如产生重复、可读性差或语言混杂的推理结果。

3 方法

3.1 任务表述与总体框架

给定一个输入样本 x，其中包含一幅图像 I，图像中有一个由边界框标出的对象 o；同时给定一段文本描述 T，其中包含一个实体 e。MORE任务 [10] 旨在识别对象与实体之间的关系标签 $R = MORE_Model$（I、o、T、e）。关系标签由 o 和 e 的实体类型及其语义关系构成，例如 /人物/组织/对立_于。可能的实体类型包括人物（per）、组织（org）、地点（loc）和其他（misc）。通过组合 o和 e的不同实体类型，MORE任务包含21种可能的关系类型，包括无关系类型，如图1所示。

如图所示。 2，MORE-R1 以大型视觉语言模型为骨干网络，并采用两阶段训练。阶段1为冷启动训练：我们使用专家模型构建少量高质量推理数据，然后对大型视觉语言模型进行监督微调，使其具备基本的 MORE 推理能力。阶段2为强化学习：我们使用先进的 GRPO [21] 作为核心强化学习算法，并结合渐进式样本混合策略。在强化学习训练过程中，随着强化学习训练轮次增加，每个小批次中简单样本与困难样本的比例会逐步调整。简单样本与困难样本的划分依据是大型视觉语言模型在阶段1冷启动训练后的推理表现。

我们采用先进的开源大型视觉语言模型 Qwen2.5-VL [1]作为骨干模型，它已广泛应用于近期的多模态研究 [6,15]。Qwen2.5-VL 可接受多张不同分辨率的图像作为输入，使用视觉编码器 $\phi_I(\cdot)$将视觉输入映射到文本空间，然后与经过分词的文本结合，并输入 Qwen2.5大语言模型 $f(\cdot)$以生成自然语言响应。为使 Qwen2.5-VL 适应 MORE 场景，我们裁剪出图像中的对象区域，并将其与原始

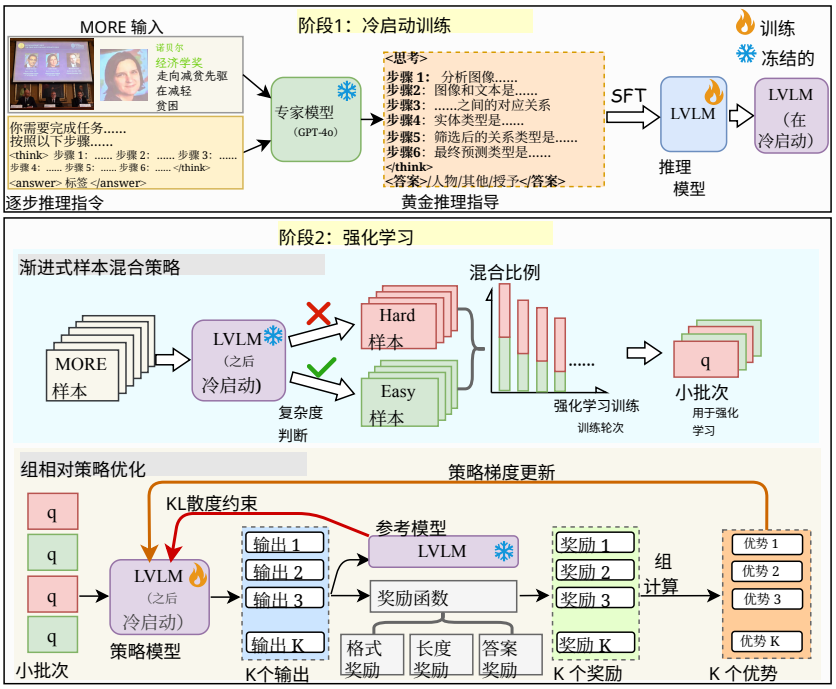


图2. MORE-R1的整体框架，该框架采用两阶段训练框架。阶段1中，冷启动训练使大型视觉语言模型学习针对MORE任务设计的基础推理范式。阶段2中，强化学习进一步提升经阶段1训练的大型视觉语言模型的推理能力。带火焰图标的模型表示其参与训练，而雪花图标表示参数被冻结。

图像组成一对图像输入。因此，在 MORE-R1 中采用 Qwen2.5-VL 作为骨干模型执行 MORE任务，可表述为：

$$Rea, Ans = f(\phi_I(I), \phi_I(o), Ins(T, e); \Theta) \tag{1}$$

其中， Rea 和 Ans 分别表示大型视觉语言模型的中间推理过程和最终预测标签。 I 是原始给定图像， o表示用于关系抽取的对象对应的裁剪图像区域。 Ins（·）是文本输入，其中包含与任务相关的指令，以及用于关系抽取的给定文本 T 和实体 e 。Θ 表示大型视觉语言模型中大语言模型的参数。

3.2 冷启动训练阶段

如图 2 上半部分所示，在阶段1中，我们通过使用小规模示例推理数据进行监督微调，引导大型视觉语言模型获得 MORE任务所需的基本逐步推理能力。由于人工标注推理过程示例极其耗时费力，我们设计了一种高效的自动推理示例构建策略。通过精心设计细粒度的标注提示，我们利用 GPT-4o [12]作为专家模型，其具备强大的

指令遵循能力，用于生成推理过程示范数据。专家模型的标注提示由三部分组成：基本任务描述、分步推理指令和实际答案提示[1]。

基本任务描述：简要介绍输入模态及 MORE任务的基本目标，列出所有候选关系标签，并指示模型将中间推理步骤输出在 <think> </think> 标签中，将最终预测标签输出在 <answer> </answer> 标签中。

逐步推理指令：包含细粒度的逐步推理指令，并提供六个顺序子步骤的指导，使模型能够逐步理清 MORE 输入中复杂的相互依赖关系：

– 步骤 1：图像与对象分析。引导模型初步理解图像所描绘的内容及其中对象的作用，为后续推理奠定基础。– 步骤 2：跨模态相关性评估。引导模型评估图像与文本的语义是否相关，为随后推断二者之间的细粒度对应关系做好准备。– 步骤 3：跨模态对齐。引导模型建立视觉对象与文本实体之间的对应关系，从而为后续推理子步骤提供帮助。例如，如图1所示，如果模型能够识别出给定图像中的对象对应于文本中提到的实体“凯尔特人队”，便能够准确推断出图像中的对象正在与来自“热火队”的一名队员对战。– 步骤 4：实体类型识别。引导模型将对象和实体判定为以下类型之一： 人物、 组织、地点、 其他，以加深模型对其语义的理解，并进一步缩小候选关系的范围。– 步骤 5：初步关系类型筛选。由于关系标签包含实体类型，此步骤引导模型缩小候选关系集合，从而降低任务复杂度。例如，如果模型判定对象和实体分别属于 人物 和 组织 类型，那么在当前类型组合下，21 个候选标签中只有 4 种关系（对抗、 领导者、成员、 无关系）是可能的。– 步骤 6：精确关系类型判定。基于前述步骤中的多模态理解和跨模态桥接，此步骤引导模型从筛选后的候选集中准确确定最终关系类型。

实际答案提示：为确保专家模型通过逻辑连贯的中间推理得出正确关系，我们将真实答案附加到完整提示的末尾，并要求模型生成能够推导出该标签的推理。对于非*无关系* 关系样本，我们提供金标准实体类型和关系类型。对于*无关系* 关系，我们提示模型给定的对象与实体之间不存在关系。

由于 GPT-4o 是闭源的，且只能通过付费 API 访问[2],为平衡 API 使用成本，同时获得相对充足的冷启动训练

[1]专家模型的完整提示模板见 https://github.com/MartinYuanNJU/MORE-R1

[2]https://platform.openai.com/docs/models/gpt-4o

数据，我们从训练集的每个关系类别中抽取 25% 的样本。使用专家模型生成推理数据后，我们设计了一种简单的基于正则表达式的筛选策略，以移除不符合推理结构或包含错误答案的样本。随后，使用筛选后的数据进行阶段 1 监督微调，此时仅向大型视觉语言模型提供基本任务描述。我们使用标准的下一词元预测损失 [3]进行阶段 1 训练。

3.3 强化学习阶段

通过阶段1，大型视觉语言模型掌握了执行 MORE任务所需的基本推理能力。为了进一步提升模型在 MORE任务复杂场景中的推理表现，我们在阶段2进行强化学习，以增强模型探索中间推理路径并发现处理复杂案例的有效策略的能力。具体而言，我们的强化学习训练采用先进的 GRPO算法 [21], ，该算法已成功应用于其他多模态理解任务，例如视觉问答 [6] 和医学图像分析 [15]。在强化学习训练期间，我们引入渐进式样本混合策略，引导大型视觉语言模型有效掌握针对复杂案例的推理。

组相对策略优化 GRPO [21] 是一种基于策略的规则强化学习算法，它通过免去训练价值模型来计算奖励信号，简化了奖励机制，从而降低整体计算复杂度。该算法利用直接的、基于规则的奖励函数评估输出正确性，引导模型在解空间中自主探索推理路径，而不依赖人工标注的偏好数据。在正确答案定义明确且细粒度标注有限的任务中，该算法具有显著优势 [6]。因此，GRPO非常适用于MORE任务，因为最终答案是定义明确的标签，而细粒度的黄金推理路径极难标注，且耗费大量人力。

如图2下半部分所示，GRPO 实现了一种简单而有效的计算流程：它直接评估针对一个输入问题生成的多个可能输出的相对优势。具体而言，给定一个输入问题 q，GRPO 首先使用旧策略 $\pi_{\theta_{\text{old}}}$ 生成 K 个不同的响应$\{o_1, o_2, ..., o_K\}$（称为一个组），然后使用基于规则的奖励函数 $r(\cdot)$计算当前组中每个响应的奖励 K ，从而得到 $\{r(o_1)\ r(o_2)$ 等对应的奖励$r(o_K)\}$。随后，GRPO 通过组内标准化计算每个响应 A_i 的相对优势分数 o_i ：

$$A_i = \frac{r(o_i) - \text{mean}(\{r(o_1), r(o_2), ..., r(o_K)\})}{\text{std}(\{r(o_1), r(o_2), ..., r(o_K)\})} \tag{2}$$

根据所有响应的相对优势分数，GRPO 通过最大化以下目标，在奖励最大化与策略稳定性之间取得平衡，以优化当前策略模型 π_θ:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \frac{1}{K}\sum_{i=1}^{K}\left(\min\left(\xi_i A_i, \text{clip}\left(\xi_i, 1-\epsilon, 1+\epsilon\right) A_i\right) - \beta\mathbb{D}_{KL}\left(\pi_\theta||\pi_{\text{ref}}\right)\right) \tag{3}$$

其中 $\xi_i = \frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}$ 用于量化策略变化，ϵ 而 β 分别控制裁剪阈值和散度惩罚的强度。π_{ref} 是用于约束策略模型 π_θ的参考策略模型，以限制其探索，避免过度散度并防止灾难性遗忘。该约束通过方程 3中的 Kullback–Leibler（KL）散度项来实现：

$$\mathbb{D}_{KL}(\pi_\theta||\pi_{\text{ref}}) = \frac{\pi_{\text{ref}}(o_i|q)}{\pi_\theta(o_i|q)} - \log\frac{\pi_{\text{ref}}(o_i|q)}{\pi_\theta(o_i|q)} - 1 \tag{4}$$

GRPO中用于MORE任务的奖励函数：MORE任务的奖励函数r(·)是一个基于规则的函数，与MORE-R1探索合理推理路径以得出正确关系预测的学习目标相一致。我们设计了一个综合奖励，由三个部分组成：格式奖励、长度奖励和答案奖励。

– 格式奖励 r_{format}：该奖励旨在强制策略模型遵循推理模板格式以完成推理探索。该格式描述为： <think> 步骤1：$[r1]$ 步骤2：$[r2]$ 步骤3：$[r3]$ 步骤4：$[r4]$ 步骤5：$[r5]$ 步骤6：$[r6]$ </think> <answer>$[label]$</answer>，其中 $[r1] \sim [r6]$ 为不受约束的推理内容，且预测的$[label]$ 必须是21个候选关系标签之一。若策略模型的输出符合该格式，则奖励为1.0；否则为0.0。– 长度奖励 r_{length}：我们期望策略模型在每个步骤中充分思考，避免仓促判断，从而形成相对较长的思维链推理，以获得更准确的最终预测。若输出文本长度超过1,024，则该奖励为1.0；否则为0.0。– 答案奖励 r_{answer}：由于MORE任务对关系标签有明确的标准答案，该奖励旨在使模型探索能够导向正确答案的中间推理路径。若预测的关系标签正确，则奖励为1.0；否则为0.0。

在此复合奖励函数下，大型视觉语言模型可在逐步推理范式的引导下进行充分探索，从而实现更准确的抽取。因此，给定响应 o_i 的奖励计算如下：

$$r(o_i) = r_{\text{format}}(o_i) + r_{\text{length}}(o_i) + r_{\text{answer}}(o_i) \tag{5}$$

渐进式样本混合策略 在阶段2中，一种直观的做法是使用训练集中的所有样本，但排除用于阶段 1 监督微调的样本，以进行 GRPO 训练（即原始训练集剩余的 75%）。然而，在阶段 1 训练后，我们直接将这些样本用于大型视觉语言模型的 GRPO 训练时，只观察到模型性能略有提升，且答案奖励在 GRPO 训练期间没有显著增长。因此，我们提出了一个假设：剩余 75% 的样本大多是“简单样本”，也就是说，经过阶段 1 训练后，大型视觉语言模型已经能够正确回答这些样本。这类样本对强化学习训练帮助不大，而在一个小批次中混入过多此类样本，会导致模型对简单模式过拟合，削弱其

算法1 采用渐进式样本混合策略的 GRPO

输入 冷启动训练后的大型视觉语言模型 $\pi_{\theta_{\text{init}}}$；原始 GRPO 训练集 $\mathcal{D}$；奖励函数 $r(\cdot)$；超参数 ϵ, β, μ, α；小批量大小 B

1：使用 $\pi_{\theta_{\text{init}}}$ 将 $\mathcal{D}$ 划分为简单样本集 $\mathcal{D}_{\text{easy}}$ 和困难样本集 $\mathcal{D}_{\text{hard}}$
2：参考模型 $\pi_{\text{ref}} \leftarrow \pi_{\theta_{\text{init}}}$
3：策略模型 $\pi_\theta \leftarrow \pi_{\theta_{\text{init}}}$
4：**对于** 训练轮次 $t = 1, 2, \ldots\ldots, E$ **执行**
5：　**对于** 步骤 $= 1$、2、……、M_t **执行**
6：　　样本 $\lceil \frac{\alpha^{t-1}B}{1+\alpha_{t-1}} \rceil$ 来自 $\mathcal{D}_{\text{easy}}$ 的数据，样本 $\lceil \frac{B}{1+\alpha^{t-1}} \rceil$ 来自 $\mathcal{D}_{\text{hard}}$的数据
7：　　混合采样数据以形成一个小批次 $\mathcal{D}_b$
8：　　更新旧策略模型 $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$
9：　　采样 K 输出 $\{o_i\}_{i=1}^{K} \sim \pi_{\theta_{\text{old}}}$ 对于每个问题 $q \in \mathcal{D}_b$的 $(\cdot \mid q)$
10：　　使用 $r(\cdot)$ 为每个采样输出 o_i 计算奖励 $\{r_i\}_{i=1}^{K}$
11：　　为每个输出计算相对优势分数 $\{A_i\}_{i=1}^{K}$
12：　　**对于** GRPO 迭代 $= 1, 2, \ldots\ldots, \mu$ **执行**
13：　　通过最大化目标函数 $\mathcal{J}_{\text{GRPO}}(\theta)$ 来更新策略模型 π_θ

GRPO 训练后的大型视觉语言模型策略模型 π_θ

探索复杂场景的能力。因此，我们进行了进一步探索。我们使用阶段 1 训练后的大型视觉语言模型对剩余 75% 的样本进行推理，并根据预测标签是否正确，将原始 GRPO 训练数据划分为简单样本和困难样本。结果显示，原始 GRPO 训练样本中有 79% 是简单样本，这与我们的假设一致。那么，一种直接的做法就是仅使用困难样本进行 GRPO 训练。然而，我们观察到，随着训练推进，答案奖励几乎没有提升，模型性能甚至有所下降。我们进一步分析发现，如果训练开始时没有简单样本提供的预热指导，模型就无法为复杂案例找到有效的探索轨迹，导致策略更新效果不佳。

基于上述观察，我们提出了一种在强化学习训练过程中使用的渐进式样本混合策略。如图 2 中间部分所示，我们首先由阶段1训练后的大型视觉语言模型进行复杂度判断，将原始 GRPO 训练样本划分为困难样本和简单样本；随后在 GRPO 训练期间，随着训练轮次增加，逐步调整每个小批次中简单样本与困难样本的比例。具体而言，逐步提高困难样本的占比，使模型能够平稳地过渡到从更具挑战性的样本中学习，从而有效增强其推理能力。在 GRPO 训练的初始轮次，我们将每个小批次中简单样本与困难样本的比例设为 1:1。随着训练轮次 t增加，简单样本的比例逐渐降低，公式如下：

$$ratio_{\text{easy}}(t) : ratio_{\text{hard}}(t) = \alpha^{t-1} : 1 \tag{6}$$

其中 $t \geq 1$ 表示当前训练轮次，$\alpha \in (0, 1]$ 表示衰减因子。我们在算法 1 中展示了阶段2的整体计算逻辑。

4 实验

4.1 实验设置

数据集与评估：我们在常用的 MORE 基准 [10]上进行评估。MORE 数据集使用在线多模态新闻数据作为资源，并标注了 20,264 个样本。每个样本都包含一张图像和一段文本，其中图像中给定一个对象（以边界框标示），文本中给定一个实体，并提供对象与实体之间关系的标准答案。MORE 基准提供了 21 种可能的关系类型（包括 无关系类型），并将所有样本划分为 15,486 个训练样本、1,742 个开发集样本和 3,036 个测试集样本。遵循以往研究 [14,10,17,11,22,7,16],，我们报告准确率（ Acc ）、精确率（P）、召回率（R）和 F1分数（$F1$）来评估性能。如 [16], 所述，由于 MORE 是一项类别不平衡任务， $F1$ 是最重要的指标，因为它能减轻长尾效应并提供稳健的评估。

基线方法：我们将 MORE-R1 与当前用于 MORE任务的最先进方法进行比较，包括 IFAformer [14], MOREformer [10], HVFormer [17],MMIB [7],CGI-MRE [22], FocalMRE [11], REMOTE [16]。这些工作均基于分类，其详细信息已在第 2.1 节介绍。我们设计了一种大型视觉语言模型基线 Qwen2.5-VL-SFT，直接使用 MORE 训练集对与 MORE-R1 相同的大型视觉语言模型主干进行监督微调，迫使模型直接输出关系标签，这属于基于生成（无推理）的方法。我们还报告了具有代表性的大视觉语言模型的零样本推理结果：直接提示大型视觉语言模型输出关系标签，而不进行监督微调。

实现细节：我们使用 ms-swift 3.8.0 框架进行整体训练，所有训练均在 8 块英伟达 A100 80GB GPU 上进行。我们使用广泛采用的 Qwen2.5-VL-7B-Instruct 作为大型视觉语言模型主干。在所有训练中，我们仅微调大型视觉语言模型中的大语言模型部分，并冻结视觉编码器和投影层。对于阶段1训练，我们从训练集的每个关系类别中抽取 25% 的样本，并使用 GPT-4o 生成推理标注。过滤掉不符合输出要求的回答后，我们获得了 3,865 个训练样本。在阶段1中，MORE-R1 训练 20 轮，每块 GPU 的批次大小为 8，学习率为 1e-4，梯度累积步数为 2，预热比例为 0.05，最大长度为 2096。阶段1训练耗时约 2 小时。

我们使用阶段1训练后的大型视觉语言模型对剩余训练集进行推理，并根据预测标签正确性将样本划分为 2,370 个困难样本和 8,853 个简单样本。在阶段2中，我们设置 4 个训练轮次，并将渐进式样本混合的衰减因子 α 设为 0.5。在每个 GRPO 训练轮次中，数据由所有困难样本以及一定数量的简单样本混合而成，简单样本的数量由公式 6定义的比例决定。采样简单样本时，我们确保 无关系 和非无关系 类别样本的占比与原始训练集一致。GRPO 在阶段1训练后的大型视觉语言模型上进行训练，GPU批次大小为 16，学习率为 1e-6，梯度累积步数为 1，预热比例为 0.05，温度为 0.8，最大补全文本长度为 1024。 ϵ 、 β 、 μ 在 GRPO 算法 1 中分别设为 0.2、0.001、2。阶段2训练耗时约 24 小时。所有阶段均采用 bfloat16精度进行训练。推理时，

表 1。MORE-R1 及其变体和其他最先进的基线方法在 MORE基准上的结果。最佳结果以**粗体**标出，次佳结果以下划线标出。“(- zs)”表示未经过训练的零样本结果。“Qwen2.5-VL-SFT(25%)”表示在阶段1中使用训练集相同的25%数据进行标签监督微调，不进行推理训练。“MORE-R1（阶段1）”仅保留冷启动训练，不包含阶段2强化学习。

模型	准确率	精确率	召回率	F1分数
基于分类的方法				
IFAformer [14]	79.28	55.13	54.24	54.68
MOREformer [10]	83.50	62.18	63.34	62.75
HVFormer [17]	82.31	58.81	62.84	60.76
MMIB [7]	82.41	60.15	62.28	61.17
CGI-MRE [22]	81.92	57.44	63.01	60.09
FocalMRE [11]	82.44	60.75	62.89	61.81
REMOTE [16]	<u>83.64</u>	<u>63.21</u>	64.63	<u>63.91</u>
基于生成（无推理）的方法				
LLaMA3.2-Vision-11B(-zs) [8]	6.69	5.09	18.66	8.01
Qwen2.5-VL-7B(-zs) [1]	23.94	25.81	23.95	24.84
Qwen2.5-VL-SFT	82.64	55.08	<u>64.84</u>	59.56
Qwen2.5-VL-SFT(25%)	80.20	51.85	54.11	52.96
基于生成（含推理）的方法				
MORE-R1（阶段1）	83.33	62.53	62.84	62.69
MORE-R1	**84.91**	**65.88**	**69.83**	**67.80**

我们使用 vLLM 引擎加速，并将温度设为 0，以获得稳定且可复现的输出。在整个测试集上进行推理，在单张 A100 GPU 上约需 7 分钟。

4.2 主要结果

表1 展示了我们的 MORE-R1 与当前最先进基线方法在 MORE基准上的性能。结果表明，我们的方法显著优于当前最先进的基线方法。与最先进的基于分类的方法 REMOTE 相比 [16], , 完整的 MORE-R1 在 Acc、P、R、$F1$ 上分别高出 1.5%、4.2%、8.0% 和 6.1%。与基于生成（无推理）的基线方法 Qwen2.5-VL-SFT 相比，MORE-R1 在 Acc、P、R、$F1$上分别高出 2.7%、19.6%、7.7% 和 13.8%。这些结果证明，MORE-R1 能够在生成最终抽取答案前进行显式推理，并采用两阶段训练范式来增强大型视觉语言模型在 MORE任务上的推理能力。此外，我们观察到：（1）Qwen2.5-VL-SFT 是使用全部训练集直接训练大型视觉语言模型输出关系标签的基线方法，其性能显著低于当前最先进模型 REMOTE， P低 12.9%，且 $F1$低 6.8%；（2）在零样本提示场景下，大型视觉语言模型的表现非常差（表中的“(-zs)”结果由 [16]报告）。这表明，仅通过简单的提示学习或在 MORE任务上直接微调，大型视觉语言模型无法取得理想结果，也进一步凸显了我们工作的先进性：成功地将大型视觉语言模型适配到 MORE任务，并取得最先进的性能。

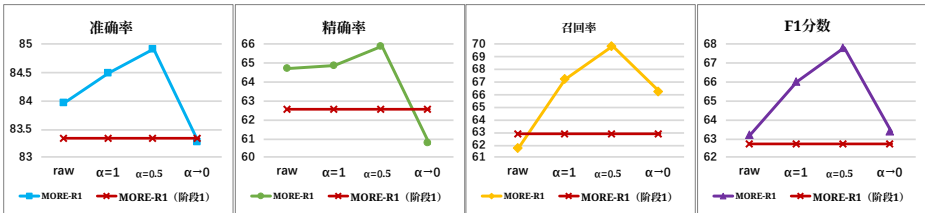


图 3. 不同样本混合策略在四项评估指标上的结果。“MORE-R1（阶段1）”是仅进行阶段1训练的基线。“MORE-R1”报告了采用不同策略进行阶段2训练后的结果。

4.3 消融研究

我们开展了多项消融研究，以验证方法中各组成部分的有效性。对于阶段1冷启动训练，我们在表1 中报告其性能，标记为“MORE-R1（阶段1）”。可以观察到，经过阶段1训练后，MORE-R1 的表现已接近最先进结果 REMOTE，[16], 在 Acc 和 $F1$ 上 Acc 分别 $F1$低 0.4% 和 1.9%。与 Qwen2.5-VL-SFT 相比，MORE-R1（阶段1）在 Acc 和 $F1$上分别高出 0.8% 和 5.3%；而且，与 REMOTE 和 Qwen2.5-VL-SFT 相比，我们仅使用了 25% 的训练数据。这表明，通过监督微调，大型视觉语言模型获得了完成 MORE任务所需的基本推理能力。经过进一步的阶段2强化学习训练后，MORE-R1 的推理能力显著提升，大幅领先所有基线方法。

此外，我们还观察到，与 Qwen2.5-VL-SFT 相比，MORE-R1（阶段1）在 P 上显著优于前者，高出 13.5%，但在 R 上略有下降，低了 3.1%。我们对此解释如下：由于 Qwen2.5-VL-SFT 仅使用最终关系标签作为监督信号，对训练集中的所有数据直接进行监督微调，模型未能有效学习无关系 类别与其他 20 个包含实际关系语义的非无关系 类别之间的语义差距，而是将全部 21 个类别视为离散且权重相同的标签。因此，对于那些实际上不包含任何关系的复杂样本，模型倾向于预测为非无关系 关系。总体而言，这导致模型倾向于预测 20 种非无关系 关系之一，从而间接提高了 R，但显著降低了 P。与 Qwen2.5-VL-SFT 相比，仅使用 25% 的样本，MORE-R1（阶段1）便能通过逐步推理引导，更稳健地区分无关系 与非无关系 关系，使整体评估的 $F1$ 提高了 5.3%。此外，为了进行更公平的比较，我们报告了使用阶段1中相同 25% 样本进行仅标签监督微调的 Qwen2.5-VL-SFT(25%)，并发现 MORE-R1（阶段1）在所有指标上均显著优于它。

为验证渐进式样本混合策略在阶段2中的有效性，我们设计了三种变体：（1）“原始”：在阶段1之后，每个 GRPO 训练轮次都直接使用剩余的 75% 原始数据，将所有简单样本与困难样本混合。（2）“$\alpha = 1$”：每个训练轮次采样与困难样本数量相等的简单样本，确保混合比例为 1:1，且不进行渐进式衰减。（3）“$\alpha \to 0$”：每个训练轮次仅使用困难样本进行训练。如图 3所示，在各项评估指标上，我们完整的 MORE-R1

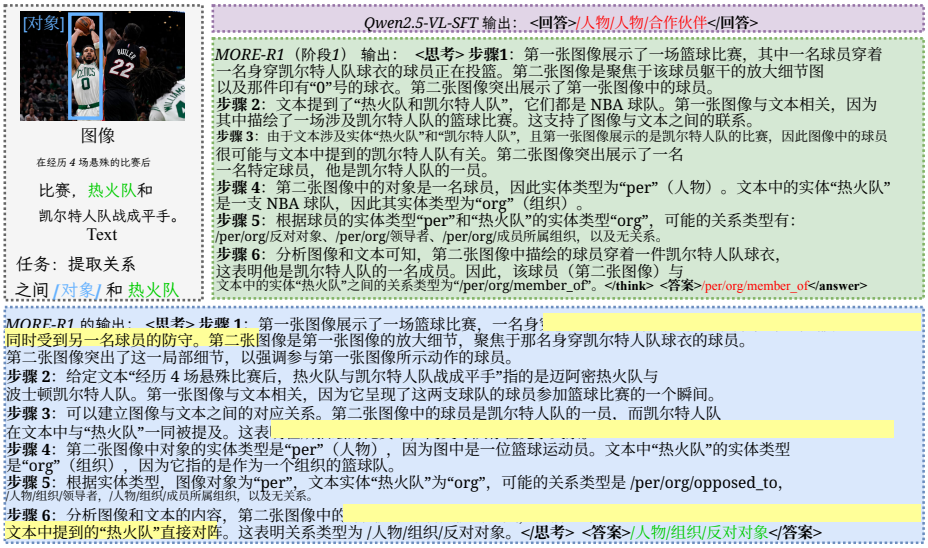


图4. Qwen2.5-VL-SFT（紫色框）、MORE-R1（阶段1）（绿色框）和 MORE-R1（蓝色框）的实际输出示例。只有 MORE-R1 给出了正确答案。

采用衰减因子 $\alpha = 0.5$ 进行渐进式简单样本递减，在所有变体中表现最佳。"原始"和"$\alpha \to 0$"表现欠佳的主要原因已在第 3.3节中分析。我们还发现，对于"$\alpha \to 0$"，Acc 和 P 的表现甚至低于强化学习训练前；对于"原始"，R 指标则低于 MORE-R1（阶段1）。我们对这些现象的解释如下。简单样本中包含大量 无关系 类型的数据（8,853 个样本中有 7,417 个），而大多数困难样本都属于非 无关系类型（2,370 个样本中有 1,574 个）。在强化学习训练期间，将所有简单样本与困难样本混合（"原始"）会导致模型在大量信息量较低的 无关系 样本上进行优化，从而削弱其对非 无关系 案例的推理能力，并导致R下降。同样，仅保留困难样本进行强化学习训练（"$\alpha \to 0$"）会严重削弱模型区分 无关系 与非 无关系 案例的能力，从而降低其 Acc和 P。原始渐进式样本混合策略（$\alpha = 0.5$）取得的优异结果证明了其有效性，因为它有助于大型视觉语言模型形成稳定的探索行为，并逐步适应更复杂的 MORE案例。

4.4 案例研究

如图所示。4，我们展示了测试集中的一个示例，并使用三个模型进行推理：Qwen2.5-VL-SFT、MORE-R1（阶段1）和完整的 MORE-R1。我们可以看到，Qwen2.5-VL-SFT 的预测有误，甚至在实体类型判断上也出现错误，表明它无法处理复杂场景。MORE-R1（阶段1）具备基本的逐步推理能力：它能正确识别实体类型并筛选潜在的候选关系。然而，由于缺乏进一步的强化学习推理增强，该模型难以推断

跨模态实体之间隐含的对抗语义。经过渐进式样本混合强化学习训练后，完整的 MORE-R1 在复杂场景中展现出强大的推理能力。如图4黄色高亮部分所示，MORE-R1 能够揭示实体之间的对抗关系，并据此识别出正确的“相反于”关系。可视化结果进一步证实了 MORE-R1 处理 MORE 任务的有效性。

5 结论

我们提出了 MORE-R1，这是一种采用两阶段训练的基于生成的方法，以大型视觉语言模型为骨干，并通过显式推理来解决 MORE 任务。在 MORE-R1 中，我们首先利用自动构建的细粒度逐步推理数据，对大型视觉语言模型骨干进行冷启动训练和监督微调。这一过程使模型习得专为 MORE 任务设计的推理模式。随后，我们基于 GRPO算法进行强化学习训练，进一步增强大型视觉语言模型针对特定任务的推理能力。在强化学习过程中，我们设计了一种渐进式样本混合策略，以稳定推理优化，并增强模型处理困难案例的推理能力。在 MORE基准上进行的大量实验表明，我们的方法有效且优越。

参考文献

1. Bai, S., Chen, K., Liu, X., 等：Qwen2.5-vl 技术报告。arXiv 预印本 arXiv:2502.13923（2025） 2. Chen, F., Feng, Y.：面向多模态命名实体识别和多模态关系抽取的思维链提示蒸馏。arXiv 预印本 arXiv:2306.14122（2023） 3. Chen, L., Wang, Z., Ren, S., 等：迈向多模态智能的下一词元预测：一篇综合综述。arXiv 预印本 arXiv:2412.18619（2024） 4. Chen, X., Zhang, J., Wang, X., 等：持续多模态知识图谱构建。载于：第三十三届国际人工智能联合会议论文集。第6225–6233页（2024） 5. Chen, Y., Yang, Z., Liu, Z., 等：Acereason-nemotron：通过强化学习推进数学与代码推理。arXiv 预印本 arXiv:2505.16400（2025） 6. Chu, X., Chen, X., Wang, G., 等：Qwen look again：引导视觉语言推理模型重新关注视觉信息。arXiv 预印本 arXiv:2505.23558（2025） 7. Cui, S., Cao, J., Cong, X., 等：利用变分信息瓶颈增强多模态实体与关系抽取。IEEE/ACM音频、语音与语言处理汇刊 **32**，1274–1285（2024） 8. Grattafiori, A., Dubey, A., Jauhri, A., 等：Llama 3 模型群。arXiv 预印本 arXiv:2407.21783（2024） 9. Guo, D., Yang, D., Zhang, H., 等：Deepseek-r1：通过强化学习激励大语言模型的推理能力。arXiv 预印本 arXiv:2501.12948（2025） 10. He, L., Wang, H., Cao, Y., 等：MORE：一个多模态对象-实体关系抽取数据集及其基准评测。载于：第31届ACM国际多媒体会议论文集。第4564–4573页（2023）

11. He, L., Wang, H., Wu, Z., 等：聚焦与门控：一种用于揭示噪声社交媒体中关系的多模态方法。载于：第32届ACM国际多媒体会议论文集。第1379–1388页（2024） 12. Hurst, A., Lerer, A., Goucher, A.P., 等：GPT-4o 系统卡。arXiv 预印本 arXiv:2410.21276（2024） 13. Jaech, A., Kalai, A., Lerer, A., 等：OpenAI o1 系统卡。arXiv 预印本 arXiv:2412.16720（2024） 14. Li, L., Chen, X., Qiao, S., 等：关于分析图像在视觉增强关系抽取中作用的研究（学生摘要）。载于：AAAI人工智能会议论文集。第16254–16255页（2023） 15. Lin, Q., Zhu, Y., Pu, B., 等：一种用于胸部X光片解读的基础模型：通过在线强化学习实现基于依据的推理。arXiv 预印本 arXiv:2509.03906（2025） 16. Lin, X., Xu, Y., Tang, M., 等：REMOTE：一种结合多层最优传输与专家混合的统一多模态关系抽取框架。载于：第33届ACM国际多媒体会议论文集。第121–130页（2025） 17. Liu, X., Hu, C., Zhang, R., 等：通过层次化视觉上下文学习器的专家混合实现多模态关系抽取。载于：2024年ACM网络会议论文集。第4283–4294页（2024） 18. Liu, Y., Yuan, X., Li, H., 等：Semscene：通过语义一致性增强的多层场景图匹配实现图文检索。ACM多媒体计算、通信与应用汇刊 **20**(8)，1–28（2024） 19. Rafailov, R., Sharma, A., Mitchell, E., 等：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） 20. Schulman, J., Wolski, F., Dhariwal, P., 等：近端策略优化算法。arXiv 预印本 arXiv:1707.06347（2017） 21. Shao, Z., Wang, P., Zhu, Q., 等：Deepseekmath：推动开放语言模型数学推理能力的极限。arXiv 预印本 arXiv:2402.03300（2024） 22. Wei, P., Huang, Z., Ouyang, H., 等：CGI-MRE：一种用于多模态关系抽取的综合遗传启发模型。载于：2024年国际多媒体检索会议论文集。第524–532页（2024） 23. Xing, F., Wang, Z., Wang, W., Zhang, H.：基准评测并提升大型视觉语言模型从多媒体文档中进行事件抽取的能力。arXiv 预印本 arXiv:2509.12876（2025） 24. Yang, X., Wu, W., Feng, S., 等：MM-BigBench：评测多模态模型在多模态内容理解任务上的表现。arXiv 预印本 arXiv:2310.09036（2023） 25. Zhao, X., Deng, Y., Yang, M., 等：关系抽取综合综述：近期进展与新前沿。ACM计算调查 **56**(11)，1–39（2024） 26. Zheng, C., Feng, J., Fu, Z., 等：利用高效图对齐进行多模态关系抽取。载于：第29届ACM国际多媒体会议论文集。第5298–5306页（2021）