arXiv:2603.06243v1 [cs.IR] 6 Mar 2026

MLLMRec-R1：激励大语言模型在多模态序列推荐中的推理能力

Yu Wang 合肥工业大学 中国合肥
w
angyu20001162@gmail.com

Yonghui Yang 新加坡国立大学 新加坡
yyh.hfut@gmail.com

Le Wu 合肥工业大学 中国合肥
lewu.ustc@gmail.com

Jiancan Wu 中国科学技术大学 中国合肥
wujcan@gmail.com

Hefei Xu 合肥工业大学 中国合肥
hefeixu604@gmail.com

Hui Lin 中国电子信息技术研究院 中国北京
linhui@whu.edu.cn

摘要

组相对策略优化（GRPO）已成为提升大语言模型推理能力和偏好对齐能力的标准后训练范式，近期也展现出在基于大语言模型的推荐系统中的显著效果。然而，将基于 GRPO 的推理流程扩展到采用多模态大语言模型（MLLM）的多模态序列推荐（MSR）面临着根本性障碍。首先，MSR 需要联合编码历史交互和多个候选物品的视觉内容，导致视觉词元占据大量输入，并使基于组的采样开销随历史长度和候选集大小而增长，从而令基于 GRPO 的训练成本高得难以承受。其次，现有的思维链（CoT）监督在推荐场景中存在奖励膨胀问题：更高的训练奖励并不能可靠地转化为更好的排序性能，还可能诱发捷径学习。为应对这些挑战，我们提出 MLLMRec-R1，这是一种面向多模态序列推荐的高效、稳定的基于 GRPO 的推理框架。MLLMRec-R1 在线下将视觉信号转化为文本，以消除开销高昂的视觉词元，同时保留多模态语义；并通过优化和置信度感知评估，构建高质量的多模态思维链监督。此外，混合粒度数据增强策略在保留标准训练数据的同时，有选择地注入可靠的思维链样本，从而缓解奖励膨胀并提升泛化稳定性。在三个基准数据集上的大量实验表明，MLLMRec-R1 的表现始终优于最先进的方法，证明了基于 GRPO 的方法在实践中的有效性

用于多模态序列推荐的推理流程。代码可通过此链接获取[1]。

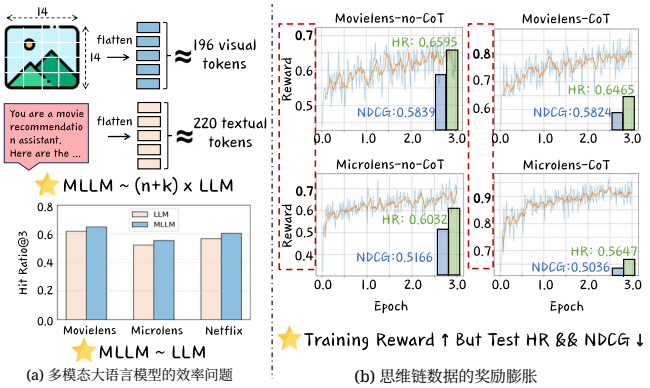
CCS 概念

• **信息系统** →**推荐系统**。

关键词

序列推荐、大语言模型、强化学习

[1]https://github.com/wangyu0627/MLLMRec-R1



(a) 多模态大语言模型的效率问题　(b) 思维链数据的奖励膨胀

图1：(a) 视觉词元会使计算量随历史长度和候选集大小增长，但相较于大语言模型，性能提升往往有限，暴露出效率瓶颈。(b) 一些思维链数据可能提高训练奖励分数，却损害测试性能，暴露出捷径学习和泛化能力差的问题。

1 引言

随着短视频和电子商务平台的快速发展，多模态序列推荐（MSR）[24, 28, 57] 利用多模态信号（例如图像 [61], 、音频 [50], 和视频 [52]）推断用户未来的兴趣。得益于多模态大语言模型（MLLM） [1, 44, 46], 强大的多模态理解与生成能力，近期研究 [48, 58, 72]越来越多地采用这类模型来提升 MSR 的性能。然而，面向推荐的推理与通用问答 [56]有所不同：MSR 必须处理结构化输入 标识、标题、图像中的长交互历史和多个候选项。如图1 (a)所示，视觉编码会将单张图像转换为 196 个视觉词元，其数量通常与文本提示的规模相当。词元数量会随着历史长度和候选集大小大致线性增长，从而导致较高的内存使用量和推理延迟。在许多情况下，相较于纯文本模型所带来的性能提升不足以抵消这项开销。**我们能否更高效地表示视觉语义，从而实现基于多模态大语言模型的 MSR 的可扩展训练与推理?**

一种自然的方法是将视觉信号离线压缩为文本，使模型在训练和推理期间以接近纯文本的

令牌成本获得多模态优势 [6, 58]。此外，将多模态信息转化为稳定的推荐决策仍然颇具挑战。先前的工作 [4, 12, 22, 67]采用监督微调（SFT） [5, 7, 18] 和直接偏好优化（DPO） [12, 36, 48] 等偏好学习方法来对齐输出，但这些方法主要拟合最终响应，往往无法组织历史记录中的多模态证据以进行细粒度的偏好推理 [27]。群体相对策略优化（GRPO） [14, 22, 30]则在相同上下文下采样多个响应，并进行相对更新，从而鼓励对推理路径的探索。**这引出了第一个挑战：要激活推理能力，就需要可学习的推理轨迹，例如思维链（CoT） [20, 49, 70], ，它能够组织历史证据、提取偏好线索并比较候选项，同时为 GRPO 提供更稳定的训练信号 [7, 19, 62]**。

为了生成更准确、信息更丰富的多模态思维链轨迹 [60, 65], ，数据构建可能会无意间引入与下一个物品画像相关的信号 [9, 47], ，从而在训练过程中造成奖励膨胀的假象 [31]。如图1(b)所示，这可能会抬高训练奖励，却损害测试时泛化。**这引出了第二个挑战：模型可能会在推理文本中利用这类线索走捷径，造成奖励膨胀，进而损害泛化能力**。

为解决这些问题，我们提出 MLLMRec-R1，这是一种简单而有效的推理范式，将思维链构建与强化学习（RL） [7, 34] 相结合，用于序列推荐。针对**挑战 1**，我们在离线阶段将视觉信号压缩为细粒度的文本描述，避免在训练和推理过程中进行昂贵的视觉词元计算，同时保留多模态语义增益。此外，我们构建了一个高质量的多模态思维链推荐数据集：MLLMRec-R1 使用现有的多模态大语言模型处理图像—标题序列，生成具有明确视觉依据和逐步推理过程的伪思维链文本，以文本形式解释交互。随后，我们将丰富的推理文本和物品描述输入纯文本推理模型 DeepSeek-R1 [14] ，以提取质量更高的思维链。针对**挑战 2**，我们提出一种混合粒度数据增强方法，将高置信度思维链样本与标准样本混合，以减少捷径学习。具体而言，我们使用多模态大语言模型嵌入标题–图像对，并计算模态一致性分数来估计输入质量。我们还评估思维链文本与预测物品之间的预测一致性，以衡量输出可靠性。结合这两种信号，我们筛选出高置信度思维链样本，并将其加入训练流。**简而言之，MLLMRec-R1 使基于多模态大语言模型的多模态推荐能够融入 GRPO 流程：文本化视觉线索提升了效率，而混合粒度训练策略则缓解了奖励膨胀**。

- 我们探究了多模态大语言模型在序列推荐中的应用，并通过压缩视觉信号、同时保留多模态语义，构建了高质量的多模态思维链数据，为强化学习提供可靠的监督信号。
- 我们进一步提出一种混合粒度数据增强方法，结合基于置信度的样本筛选与随机混合，以减少强化学习训练中的投机行为。
- 在三个基准数据集上进行的大量实验表明，我们的方法始终优于现有基线方法。

此外，我们全面分析了基于GRPO的思维链构建和混合粒度策略对多模态序列推荐性能的贡献。

2 预备知识

本节首先介绍一般序列推荐问题，然后介绍我们如何通过监督微调（SFT）和 GRPO 两种方式整合大语言模型，以增强序列推荐能力。

2.1 问题定义

在一般的序列推荐场景中，我们考虑用户集合 U 和物品集合 V，其中 $|\mathrm{U}|$ 和 $|\mathrm{V}|$ 分别表示用户和物品的总数。目标是利用用户的历史交互行为，对用户兴趣随时间的动态变化进行建模。形式化地，对于用户 $u \in \mathrm{U}$，给定其交互序列 $S^u = \{v_1^u, v_2^u, \cdots v_t^u\}$，其中 v_t^u 表示用户在时间戳 t 交互的物品，t 表示序列长度，目标是学习一个推荐函数 f_θ，用于预测用户最有可能交互的下一个物品 v_{t+1}^u。

我们使用交互历史 S^u 对序列推荐进行形式化。为了构建大语言模型指令，我们将最近的 K 次交互作为上下文，并构造大小为 k 的候选集 $\mathrm{C}^u = \{v_{t+1}^u\} \cup \mathrm{N}^u$，其中 N^u 包含从未交互物品中采样的负样本。随后，我们将 (S^u, C^u) 序列化为一个**指令提示**，其中包含四个组成部分：(i) 角色说明，(ii) 最近观看列表，(iii) 候选列表，以及 (iv) 严格的输出约束。如图1(a)所示，每个物品均由规范标记 [ITEM id] 及其电影标题表示，并且**输出响应** 被约束为遵循这一格式。接下来，我们介绍一个包含监督微调和 GRPO 的两阶段训练范式。我们首先使用监督微调使预训练大语言模型适应推荐指令，然后应用 GRPO 在相同提示下采样多个响应，并利用相对优势更新策略，从而进一步改进序列推荐中的排序决策。

2.2 监督微调（SFT）

对于纯文本大语言模型 π，用户偏好学习通常从指令微调阶段开始，从而得到一个微调模型 π_ϕ [4]。它为后续基于 GRPO 的后训练初始化一个稳定的策略分布。具体而言，该阶段将预训练大语言模型改造为面向推荐任务的指令跟随模型，并在数据集 $\mathrm{D} = \{(x^u, y^u)\}_{u=1}^{|\mathcal{D}|}$ 上对其进行优化，其中 x 表示**指令提示**，而 y 表示**输出响应**。形式上，SFT 通过给定提示 x，最大化生成目标响应 y 的似然：

$$\mathcal{L}_{\mathrm{SFT}} = -\mathbb{E}(x, y) \sim \mathcal{D} \left[\log, \pi_\phi(y \mid x)\right]. \tag{1}$$

2.3 组相对策略优化（GRPO）

经过监督微调后，我们进一步通过强化学习改进指令遵循型推荐模型 π_ϕ，得到一个记为 π_θ 的策略。与依赖构造明确的优选—拒选样本对的成对偏好优化 [36] 不同，GRPO 直接利用在相同提示下采样得到的**组内** [41] 响应优化模型。给定一个**指令提示** x，我们采样一组 G 个响应 $\{o_1, o_2, \ldots\ldots, o_G\} \sim \pi_{\theta_{\mathrm{old}}}(\cdot \mid x)$，并进行评分

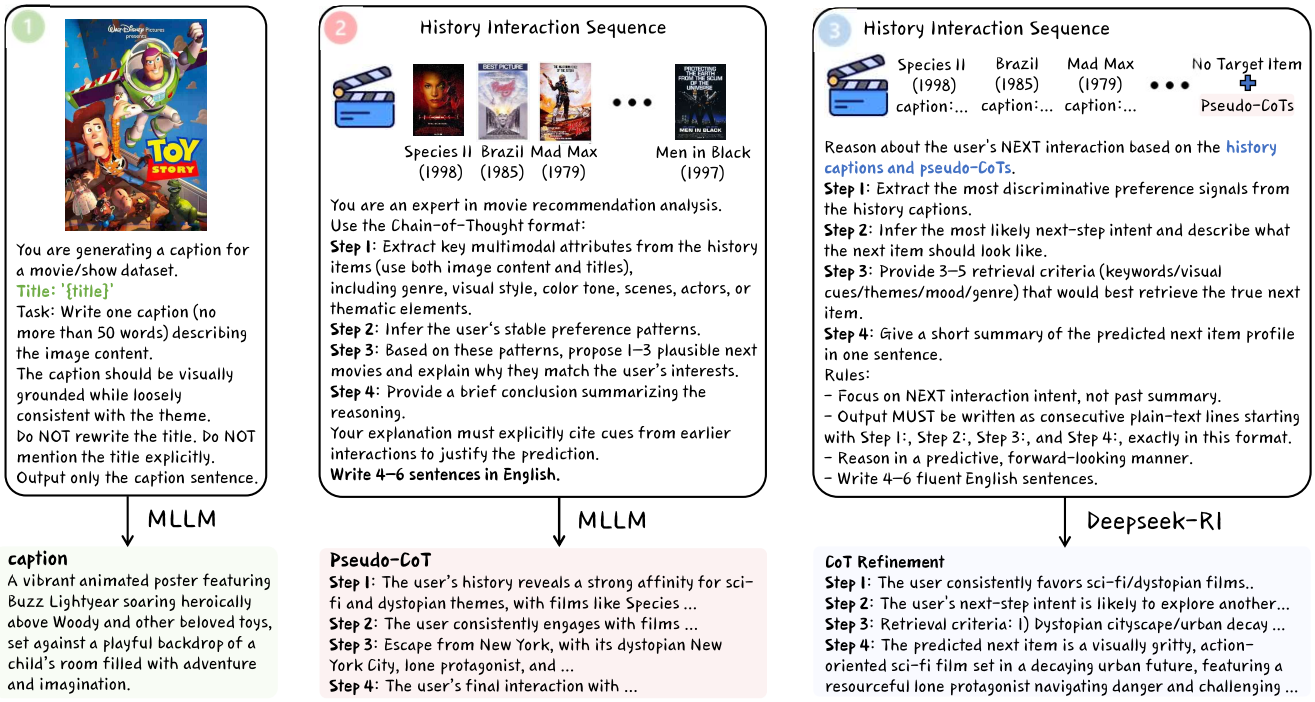


图2：多模态思维链构建流程。(1) 标题生成。我们使用多模态大语言模型为每个物品图像生成细粒度标题，将成本高昂的视觉信号离线压缩为可重复使用的文本。**(2) 伪思维链构建**。给定交互历史，我们提示多模态大语言模型生成结构化思维链推理依据，以文本形式呈现普通标题未能涵盖的关键视觉线索。**(3) 思维链优化**。我们将这些视觉线索与推理轨迹相结合，并将其输入 DeepSeek-R1 [14]，以获得高质量的多模态思维链监督信号。

使用任务奖励 $r(x,\ o_i,\ y)$ 为每个响应评分。GRPO 将组内的相对优势作为训练信号：

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{x\sim P(Q),\,\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(O|q)}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\right.$$
$$\underbrace{\min\left\{\frac{\pi_\theta(o_{i,t}\mid x,\ o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}\mid x,\ o_{i,<t})}\hat{A}_{i,t},\right.}_{\text{policy ratio}}$$
$$\underbrace{\left.\left.\text{clip}\left(\frac{\pi_\theta(o_{i,t}\mid x,\ o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}\mid x,\ o_{i,<t})},\ 1-\varepsilon,\ 1+\varepsilon\right)\right\}\hat{A}_{i,t}\right]}_{\text{clipped ratio}}$$
$$\underbrace{-\ \beta\ D_{\text{KL}}\left[\pi_\theta\,\|\,\pi_\phi\right]}_{\text{stability term}} \tag{2}$$

其中，$P(Q)$表示指令提示的经验分布。$\pi_{\theta_{\text{old}}}$ 是用于采样 o_i的冻结策略，π_θ是更新策略。同时，ε控制 PPO 式 [40] 裁剪范围，以实现稳定更新；β则对 KL散度加权，以将 π_θ 约束在 SFT参考策略π_ϕ附近。此外，我们使用组均值基线，根据序列级奖励计算词元级优势 Ai, t：对于 $t \in 1, \ldots, |o_i|$，$\hat{}A_{i,t} = r(x, o_i, y) - \frac{1}{G}\sum_{j=1}^{G} r(x, o_j, y)$，并将相同的$\hat{}A_{i,t}$应用于 o_i中的所有词元。最后，任务奖励函数 $r(x, o_i, y)$的设计将在下一节详述。

3 方法

我们的 MLLMRec-R1 专注于高质量训练信号，以稳定 GRPO 后训练。在离线阶段，我们使用多模态大语言模型将封面图像压缩为细粒度标题，并生成序列级伪思维链。随后，在不提供目标物品的情况下，使用更强的文本推理模型进行思维链优化，以避免标签泄漏。在在线阶段，我们根据一致性筛选低置信度样本，并将少量高质量思维链与标准数据混合进行增强。最后，我们采用轻量级奖励规则来支持 GRPO 中的组相对更新，从而提升排序性能和泛化能力。

3.1 多模态思维链（思维链）

尽管GRPO无需大规模思维链监督即可激活潜在推理，但它对训练信号的质量高度敏感，尤其是在多模态噪声 [19]下。因此，我们描述如何构建高质量的MSR推理数据，以支持基于GRPO的后训练。具体而言，我们的全自动流程包含三个步骤，如图2所示。首先，对于每个物品 v，我们使用多模态大语言模型（例如Qwen-VL [2]）将其视觉内容转换为与封面对齐的细粒度图注。设多模态输入为 $m_v = \{I_v, T_v\}$，其中 I_v为封面图像，T_v为标题。图注由多模态大语言模型生成如下：

$$c_v \sim p_\theta(c \mid m_v) = p_\theta(c \mid I_v,\ t_v)\,, \tag{3}$$

其中 c_v 是物品 v的文本表示， p_θ 是多模态大语言模型的条件生成分布。

接下来，我们构建**伪思维链**，以展现序列级的多模态偏好线索，且不使用任何目标物品信息。仅给定交互历史的多模态上下文 $\{m_{v_1^u},\ m_{v_2^u},\ \ldots,\ m_{v_t^u}\}$，我们使用带有结构化步骤约束的提示模板，引导多模态大语言模型生成逐步推理轨迹：

$$r^u \sim p_\theta\Big(r \mid m_{v^u_{1:t}},\ \Omega_{\text{step}}\Big), \tag{4}$$

其中， r^u 表示用户 u的伪思维链文本， Ω_{step} 表示对推理过程施加的逐步约束（图 2 的第 2 部分）。伪思维链并非旨在模仿复杂的人类推理，而是在数据构建中充当模态桥接步骤。它引导多模态大语言模型生成更丰富的、以历史记录为条件的描述，减少将原始多模态输入转换为文本表示时的信息损失 [9]。

最后，**思维链优化**将图像描述和伪思维链输入能力更强的纯文本推理模型，以去除噪声、补全并强化推理轨迹，从而为 GRPO 提供高质量监督信号。在此步骤中，我们仍避免暴露任何目标物品，并提示 DeepSeek-R1 仅使用文本化的历史记录 $\{c_{v_1^u},\ c_{v_2^u},\ \ldots,\ c_{v_t^u}\}$ 和伪思维链 r_u来生成用于下一意图预测的精炼思维链（图 2 第 3 部分）。形式化地，设 p_ϕ表示 DeepSeek-R1 的生成分布。精炼思维链的生成方式如下：

$$\tilde{r}^u \sim p_\phi\Big(\tilde{r} \mid c_{v^u_{1:t}},\ r^u,\ \tilde{\Omega}_{\text{step}}\Big), \tag{5}$$

其中， $\tilde{\Omega}_{\text{step}}$ 表示第三步的约束和规则。这种设计避免了依赖已知 v^u_{t+1}的事后归因“解释”，并能提供比原始伪思维链更细粒度的偏好线索。

一个自然的问题是，为什么这种构建方式有效？更简单的替代方案是否也能取得相当的结果？我们对思维链的三个组成部分进行了消融研究，表 2 中的结果表明，移除任何一个组成部分都会导致性能下降。为进一步控制虚假监督信号，我们在伪思维链构建中施加无目标约束，并在思维链优化过程中增加一个清理步骤，以过滤任何意外提及的物品标识符或未来交互（例如，可能泄露下一物品信息的电影标题/ID）。**值得强调的是，我们并未声称所提出的思维链构建流程是唯一或最优的选择。相反，我们将高质量多模态思维链数据构建视为 GRPO 式后训练的关键前提 [9, 19]**。具体而言，我们在附录 D 中提出了一套自动化的思维链数据质量评估流程。如图1所示，我们将每条生成的推理文本视为一个评估实例，并根据五个预定义维度对其评分：模态一致性、预测一致性、信号密度、泄漏风险和覆盖难度。我们要求评估者输出一个取值为 $[1, 3, 5]$ 的整数评分，以及一句仅引用原始推理文本作为依据的证据句。实验结果表明，思维链优化能够提升多模态对齐与推理一致性，同时显著降低潜在的标签泄漏风险。该评估协议可直接应用于由其他来源策略生成的思维链数据。

3.2 混合粒度数据增强

本节首先介绍一种过滤方案，该方案使用**（i）模态一致性**和**（ii）预测一致性**来移除低质量推理样本，从而减少 GRPO 更新中的噪声梯度。具体而言，对于用户u的历史序列 S^u中的每个交互物品 v，我们将其多模态输入 $m_v = \{I_v,\ T_v\}$映射为语义嵌入$\mathbf{e}_v$，$\mathbf{z}_v = f_m(I_v,\ T_v)$。然后，我们计算标题–图像对之间的平均相似度：

$$\mathbf{s}_u^{'} = \frac{1}{t}\sum_{i=1}^{t} \text{sim}\Big(\mathbf{e}_{v_i^u},\ \mathbf{z}_{v_i^u}\Big), \tag{6}$$

其中，$\text{sim}(\cdot,\cdot)$表示余弦相似度。直观而言，模态一致性越低，标题歧义、视觉噪声或跨模态不匹配的风险就越高，这可能会降低后续训练中的推理和学习效果。接下来，我们衡量预测一致性。对于每个样本，我们将精炼思维链 $\tilde{r}^u$ 第 4 步预测的下一个物品画像记为$\hat{r}^u$。我们将其编码为文本嵌入$\mathbf{r}^u = f_m(\hat{r}^u)$，并将预测一致性定义为：$\mathbf{s}_u^{''} = \text{sim}(\mathbf{z}^{v^u_{t+1}}, \mathbf{r}^u)$。预测一致性衡量推理轨迹是否与最终预测相符。如果思维链中预测的下一个物品画像不准确，我们就将该样本视为低置信度。我们将这两个相似度信号简单合并为一个序列级分数：

$$\tau_p = \text{Quantile}(\{\mathbf{s}_u\}_{u\in\mathcal{U}},\ p), \qquad \mathbf{s}_u = \sigma(\mathbf{s}_u^{'} + \mathbf{s}_u^{''}) \tag{7}$$

其中， $\sigma(\cdot)$ 是 Sigmoid函数， p 表示保留率。最后，我们根据序列级分数对所有样本进行排序$\mathbf{s}_u$，并保留分数高于 τ_p 的样本作为高置信度数据 $\mathrm{R} = \{\tilde{r}^1, \tilde{r}^2, \ldots, \tilde{r}^k\}$，用于后续训练。**模态一致性可确保输入可靠，预测一致性可确保输出可靠**。**二者共同过滤由跨模态不匹配和幻觉推理导致的噪声学习信号**。

过滤后，我们构建混合粒度训练数据集 $\mathrm{D}_{\text{mix}} = \{(x^u,\ y^u,\ \tilde{r}^u)\}_{u=1}^{|\mathcal{D}|}$，其中 x^u 已包含序列化交互历史和候选集。对于每个样本，仅当其被选入 R时，我们才为其附加精炼思维链轨迹 $\tilde{r}^u$；否则，将 $\tilde{r}^u = \varnothing$ 设为（即 $\tilde{r}^u \notin \mathrm{R}$），此时输入退化为标准的无思维链提示。 p 通常设为 0.1，作为较低的混合比例。这种混合粒度数据增强使模型能够从少量高质量思维链样本中学习细粒度的多模态偏好线索和推理模式，从而提升排序能力。同时，保留大多数标准样本可减少模型对思维链中捷径线索的依赖，避免训练奖励虚高、泛化能力较差，并更好地匹配测试分布。

3.3 轻量级奖励规则

如图4所示，在数据增强之后，我们设计了一个轻量级奖励函数，以支持GRPO的组相对更新。由于序列推荐的输出遵循明确的结构约束，我们将奖励分解为**格式检查**和**命中检查**，从而在无需额外奖励模型的情况下提供稳定、低噪声的训练信号。对于在相同上下文下采样的每个响应 o，我们首先执行**格式检查**。如果 o 未遵循所需模板 [ITEM_xxxx] 标题 <think> ... </think>，我们赋予惩罚 $R_{\text{fmt}}(o) = -1.0$；否则，我们给予基础奖励 $R_{\text{fmt}}(o) = +0.3$ 以鼓励有效输出。然后我们执行

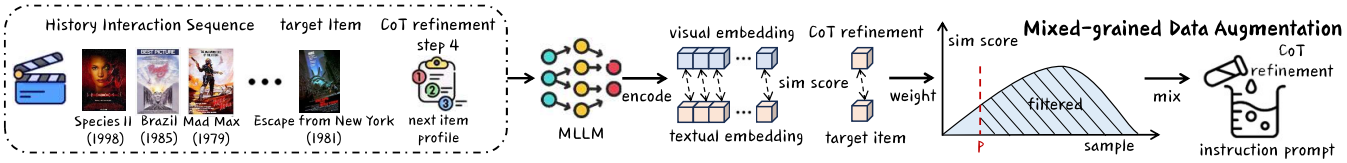


图3：我们计算标题–图像对之间的模态一致性，以及精炼思维链档案与目标物品之间的预测一致性，然后将经过筛选的思维链优化融入指令数据。

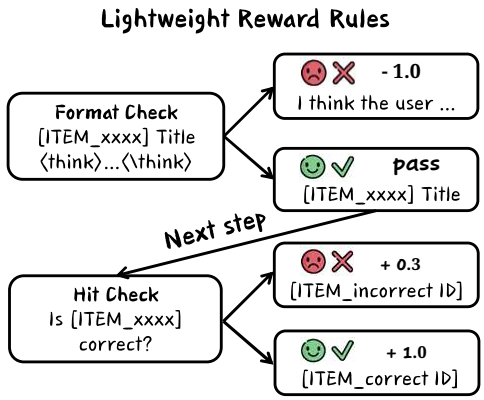


图4：GRPO 在相同上下文下执行组采样，并应用带有格式检查和命中检查的奖励规则，以提供稳定的训练信号。

命中检查：如果预测的 [ITEM_xxxx] 与真实下一物品 y匹配，我们赋予主要奖励 $R_{\text{hit}}(o, y) = +1.0$；否则，奖励保持较低。最终奖励可写为：

$$R(o) = R_{\text{fmt}}(o) + \mathbb{I}[\text{hit}(o, y)] \cdot 1.0, \tag{8}$$

其中，$\mathbb{I}$ 是指示函数。该奖励的复杂度为 O(1)，无需额外奖励模型的推理开销。我们允许模型在 <think>..... </think> 中生成解释，但不对此部分给予显式奖励。首先，没有外部验证时，自由形式推理难以可靠评估；对其给予奖励可能会鼓励模型通过模板化或幻觉解释进行奖励投机。其次，我们的过滤已能控制思维链质量，而 GRPO 的主要目标是改善排序决策。因此，将奖励集中在可验证格式和命中信号上，可以提高训练稳定性，并确保优化方向保持一致。

3.4 使用GRPO的模型优化

我们使用增强后的数据分两个阶段训练模型。在监督微调阶段，我们针对目标答案 y^u 进行最大似然训练。其他监督微调细节遵循预备知识部分，此处不再赘述。在此步骤中，我们遵循**"少而精"**原则，将过滤比例 p 设为 0.1 [14, 19]。在 GRPO 之前，监督微调主要用于冷启动初始化和行为规范化，使模型与基本的推荐指令保持一致，并减少可能在强化学习阶段占主导地位的无效输出。

随后，我们应用 GRPO。在此阶段，我们将过滤比例降低至 $p = 0.05$，因为 GRPO 使用在策略更新，对噪声信号更加敏感。如图 3 底部所示，对于每个上下文，我们从当前策略中采样一组响应，并将温度和 top-p 设置为 0.9，以增加多样性并鼓励探索，同时确保输出仍然可用。初步实验表明，这一设置能带来稳定收敛和持续的性能提升，因此我们不对温度或 top-p 进行大量调优。最后，我们对每个采样响应 o_i 应用轻量级奖励规则，以获得标量奖励 $r(x, o_i, y)$。参照公式 2 中的 GRPO 目标函数，优势项 $\hat{A}_{i,t}$ 根据每个组内推理输出对应的组级奖励 $\{r_1, r_2, \ldots, r_G\}$ 计算得出。这会鼓励模型进行更新，使响应表现优于组平均水平，同时 KL 正则化项则限制其偏离 SFT参考策略。为完整起见，我们总结了各阶段的时间复杂度，并在附录A中给出相应算法。

4 实验

本节中，我们进行了大量实验，并回答以下研究问题：

- **研究问题 1：**在性能方面，MLLMRec-R1 与当前最先进水平（SOTA）的框架相比如何?
- **研究问题 2：**我们的 MLLMRec-R1 中的关键组件是否带来了预期的性能提升?
- **研究问题 3：**模型性能更优的原因是什么?
- **研究问题 4：** 不同的超参数会如何影响 MLLMRec-R1?

4.1 实验设置

4.1.1 数据集。 我们在三个多模态数据集上评估模型：Microlens [33], Netflix [51], 和 Movielens-1M[2]。附录 B.1介绍了数据收集的详细信息。

4.1.2 基线方法。 我们将 MLLMRec-R1 与具有代表性的序列推荐、多模态推荐和基于大语言模型的推荐方法进行比较。附录 B.2介绍了基线方法的完整列表及其实现细节。

4.1.3 评估指标。 按照标准评估设置 [15]，我们采用命中率（HR） [63] 和归一化折损累积增益（NDCG） [3] 来评估模型性能。具体而言，将真实目标物品与 k 个随机采样的负样本物品组合，构成 k+1 个候选项。在我们的实验中，k默认固定为 9。为了进一步考察模型的大规模检索能力，我们还在第 4.2.2 节中研究了不同 k 对性能的影响。为确保公平比较，所有基线方法均采用相同的标准化候选集大小进行评估。

[2]https://grouplens.org/datasets/movielens/

表1：不同数据集上的推荐性能比较。上标*表示改进具有统计显著性，其中p值小于0.05。粗体表示最佳结果，下划线表示次佳结果。

	模型	Year	Movielens-1M				Microlens				Netflix			
			HR@3	HR@5	NG@3	NG@5	HR@3	HR@5	NG@3	NG@5	HR@3	HR@5	NG@3	NG@5
传统方法	**GRU4Rec** [16]	ICLR'16	0.5507	0.7131	0.4762	0.5236	0.5080	0.6440	0.4369	0.4756	0.5152	0.7152	0.3659	0.4494
	SASRec [21]	ICDM'18	0.6073	0.7498	0.5090	0.5638	0.5243	0.6597	0.4439	0.4958	0.5560	0.7570	0.4042	0.4786
	LightGCN [15]	SIGIR'20	0.5808	0.7167	0.4894	0.5330	0.5138	0.6604	0.4344	0.4913	0.5430	0.7420	0.3928	0.4702
	CL4SRec [55]	ICDE'22	0.6221	0.7557	0.5306	0.5861	0.5482	0.6826	0.4694	0.5186	0.5788	0.7873	0.4374	0.5091
多模态	**LATTICE** [66]	MM'21	0.6131	0.7558	0.5223	0.5762	0.5306	0.6612	0.4531	0.5036	0.5635	0.7955	0.4110	0.5038
	MoRec [61]	SIGIR'23	0.6059	0.7552	0.5074	0.5642	0.5267	0.6515	0.4537	0.4970	0.5742	0.7725	0.4277	0.4887
	BM3 [73]	WWW'23	0.6222	0.7618	0.5259	0.5788	0.5421	0.6836	0.4749	0.5291	0.5825	0.8155	0.4213	0.5358
	AB-Rec [53]	KDD'25	0.6194	0.7513	0.5219	0.5787	0.5426	0.6740	0.4645	0.5125	0.5865	0.8094	0.4390	0.5243
基于大语言模型	**TallRec** [4]	Recsys'23	0.6139	0.7324	0.5236	0.5613	0.5243	0.6597	0.4439	0.4958	0.5690	0.7885	0.4174	0.4967
	LLaRA [26]	SIGIR'24	0.6172	0.7490	0.5201	0.5681	0.5213	0.6799	0.4423	0.5080	0.5650	0.7700	0.4072	0.4854
	SPRec [12]	WWW'25	0.6273	0.7598	0.5290	0.5738	0.5473	0.6798	0.4690	0.5138	0.5765	0.7982	0.4250	0.5167
	RecZero [22]	NIPS'25	0.6595	0.7664	0.5639	0.6067	0.6032	0.7062	0.5166	0.5648	0.6520	0.8205	0.5195	0.5857
基于多模态大语言模型	**MSRBench** [72]	WWW'25	0.6272	0.7507	0.5288	0.5743	0.5543	0.6897	0.4739	0.5258	0.5720	0.7820	0.4295	0.5029
	MLLM-MSR [58]	AAAI'25	0.6261	0.7563	0.5313	0.5827	0.5447	0.6734	0.4655	0.5181	0.5910	0.8080	0.4429	0.5272
	MLLMRec-R1	Ours	**0.7630***	**0.8368***	**0.6524***	**0.6784***	**0.6627***	**0.7906***	**0.5845***	**0.6365***	**0.7150***	**0.8670***	**0.5902***	**0.6293***
	提升 %		15.69%	9.19%	15.82%	11.82%	9.86%	11.95%	13.14%	12.69%	9.66%	5.67%	13.61%	7.44%

4.1.4 实现细节。 为确保公平性，我们的实验和基线方法均在配备 8 张 RTX PRO 6000 GPU 的 Linux 服务器上进行。我们使用 Qwen3-VL-8B-Instruct[3] 作为思维链数据构建的多模态大语言模型骨干网络，并使用 DeepSeek-R1[4] 作为大型推理模型。在 MLLMRec-R1 的训练过程中，我们始终将低秩适配的秩设为 16、学习率设为 1e-5、梯度累积步数设为 8。在监督微调阶段，每台设备的批次大小设为 2：模型在 MovieLens 和 Netflix 上训练 3 个周期，在 MicroLens 上训练 5 个周期。在 GRPO 阶段，每台设备的批次大小增至 4：模型在 MovieLens 和 MicroLens 上训练 3 个周期，在 Netflix 上训练 2 个周期。按照 [26], 的设置，根据序列数量将所有数据集按 7:1:2 的比例划分为训练集、验证集和测试集。由于数据集的序列长度有限，所有模型仅使用最近的 9 次交互，并将最后一次交互作为预测目标。此外，我们使用相同的数据来训练和测试所有传统基线方法。

4.2 性能比较（RQ1）

本节在两种评估方案下进行实验：标准候选集和更大候选集。

4.2.1 标准训练设置。 在评估指标部分，候选集大小为 10（一个目标物品和九个随机采样的负样本）。表 1 展示了 MLLMRec-R1 在三个公开数据集上的性能提升。我们进一步讨论结果中一些有趣的观察，并给出可能的解释。所有实验结果均为五次运行的平均值。

• **传统方法的局限性**。序列 [16, 21, 55]与协同过滤 [15] 模型依赖于丰富的交互，而我们的数据仍然高度稀疏（即使在 MovieLens上进行了截断）。尽管CL4SRec [55], 是其中最强的方法，并采用基于对比学习的序列增强来缓解定义较为宽泛的冷启动问题，它仍然落后于基于大语言模型的方法SPRec [12] 在三个数据集上的表现，凸显了稀疏场景下的性能瓶颈。

• **基于大语言模型的方法的局限性**。 现有的基于大语言模型的语义检索方法 [4, 12, 22, 26]通常采用监督微调和直接偏好优化，从交互历史和采样偏好对中学习。然而，这些方法往往忽视采样负样本中显著的难度不平衡，因此直接偏好优化会反复强化由简单负例主导的对比。此外，纯文本大语言模型无法利用图像等多模态信息，因此难以解决歧义标题带来的模态偏差 [64]。以 RecZero [22]为例，RecZero 和本文提出的 MLLMRec-R1 都使用 GRPO 进行后训练。然而，RecZero 未对多模态思维链进行建模，因此整体性能明显低于 MLLMRec-R1。

• **基于多模态大语言模型的方法的局限性**。 现有基于大语言模型的语义检索方法 [4, 12, 22, 26] 通常使用监督微调和直接偏好优化，从交互历史和采样偏好对中学习。然而，这些方法往往忽视采样负例之间显著的难度不平衡，因此直接偏好优化会反复强化由简单负例主导的对比。此外，纯文本大语言模型无法利用图像等多模态信息，因此难以解决歧义标题带来的模态偏差 [64]。以 RecZero [22] 为例，RecZero 和本文提出的 MLLMRec-R1 均使用 GRPO 进行后训练。然而，RecZero 未对多模态思维链进行建模，因此整体性能显著低于 MLLMRec-R1。

• **MLLMRec-R1 的优越性**。 我们的模型在所有评估指标上始终优于所有基线方法，在 MicroLens、Netflix 和 MovieLens 上的相对提升分别为 15.82%、13.14% 和 13.61%。这些结果表明，模型对细粒度用户偏好的建模能力更强。NG@3 上更大的提升进一步表明头部排名质量更高，同时相关性和多样性有所改善，流行度偏差也得到部分缓解。

[3]https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct [4] https://huggingface.co/deepseek-ai/DeepSeek-R1

表 2：模型变体在所有数据集上的消融研究。上标 * 表示改进具有统计显著性（p 值小于 0.05）。粗体表示最佳结果，下划线表示次佳结果。

变体	Movielens-1M				Microlens				Netflix			
	HR@3	HR@5	NG@3	NG@5	HR@3	HR@5	NG@3	NG@5	HR@3	HR@5	NG@3	NG@5
不使用全部组件	0.6172	0.7490	0.5201	0.5681	0.5019	0.6530	0.4283	0.4855	0.5650	0.7700	0.4072	0.4854
不使用GRPO	0.6139	0.7324	0.5236	0.5613	0.5191	0.6582	0.4397	0.4928	0.5690	0.8010	0.4174	0.5064
不使用MDA	0.6465	0.7593	0.5824	0.6035	0.5647	0.6799	0.5036	0.5441	0.6505	0.8145	0.5098	0.5771
不使用MCoT	0.6665	0.7767	0.5978	0.6287	0.6097	0.7051	0.5174	0.5625	0.6605	0.8280	0.5177	0.5899
不使用思维链优化	0.6924	0.7791	0.5933	0.6239	0.6371	0.7558	0.5418	0.5893	0.6870	0.8249	0.5314	0.5806
不使用伪思维链	0.7095	0.7864	0.6039	0.6267	0.6394	0.7526	0.5439	0.5887	0.6810	0.8280	0.5329	0.5872
不使用标题	0.7243	0.8021	0.6230	0.6435	0.6434	0.7623	0.5616	0.6109	0.6965	0.8415	0.5518	0.6067
MLLMRec-R1	**0.7630***	**0.8368***	**0.6524***	**0.6784***	**0.6627***	**0.7906***	**0.5845***	**0.6365***	**0.7150***	**0.8670***	**0.5902***	**0.6293***

4.2.2 更多候选项的标准设置。 此设置用于评估在更大的候选项集合下排序性能的鲁棒性，反映了更贴近实际且更具挑战性的检索场景。详细结果和分析见附录 C.1。

4.2.3 不同长度的标准设置。 该设置考察交互历史长度对模型性能的影响，揭示序列上下文如何影响多模态推理和偏好学习。详细结果见附录 C.2。

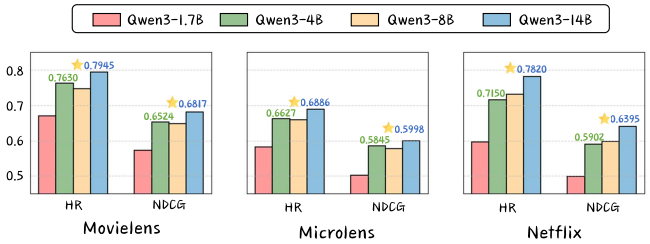


图 5：不同骨干网络规模在所有数据集上的 HR@3 和 NDCG@3 性能比较。

4.3 消融研究（研究问题 2）

4.3.1 关键组件的影响。 本节评估 MLLMRec-R1 中关键组件的影响，并对其进行可能的解释。我们评估以下模型变体：

- MLLMRec-R1 不使用全部组件：移除所有提出的组件，仅采用监督微调训练范式；
- MLLMRec-R1 不使用GRPO：移除基于GRPO的后训练，仅采用监督微调训练范式与混合粒度数据；
- MLLMRec-R1 不使用MDA：移除混合粒度数据增强，将所有多模态思维链加入训练数据；
- MLLMRec-R1 不使用MCoT：移除多模态思维链，仅在标准指令提示上进行"监督微调 +GRPO"；
- MLLMRec-R1 不使用思维链优化：仅使用伪思维链数据进行训练，不进行思维链优化；
- MLLMRec-R1 不使用伪思维链：在获取思维链优化数据时移除伪思维链；
- MLLMRec-R1 不使用标题：在获取思维链优化数据时移除图注；

表 2 展示了在所有数据集上的实验结果，并得出以下结论：移除 GRPO 会导致与移除所有组件相近的性能下降，因为二者都表现出显著的退化。这是因为，在没有用于组内偏好优化的强化学习阶段时，模型无法有效学习相对优势，从而导致其对多样化候选集的排序不可靠。尽管 SFT 能够提供基本的对齐和初始化，但它无法充分捕捉多模态思维链的推理模式和跨模态相关性。其次，不使用MCoT的性能显著差于完整的 MLLMRec-R1，这表明纯文本强化学习能够改善推荐效果，但在缺乏高质量多模态思维链数据时，其能力受到限制。同样，不使用MDA的性能略低于不使用MCoT，这表明，在没有 MDA 的情况下，模型倾向于依赖捷径，利用噪声和表面的思维链关联，这可能会损害训练效果。总体而言，MDA 对训练稳定性和泛化至关重要，因为它能够防止模型过度依赖思维链而产生偏差。

最后，我们对多模态思维链构建流程进行了细粒度消融实验。结果表明，图像描述和伪思维链均能提升性能，其中伪思维链的贡献更大，因为它提供了明确的偏好线索和可复用的推理模式，而图像描述主要用于压缩细粒度的视觉语义。结合图 8 中观察到的现象，移除思维链优化并仅使用伪思维链进行训练会导致性能大幅下降，这可能是因为潜在的信息冲突损害了测试时泛化能力。

4.3.2 不同大语言模型骨干网络的影响。 为评估不同多模态大语言模型作为骨干网络的影响，我们使用不同规模的 Qwen3 开展实验 [1]。如图 5 所示，在所有数据集上，规模更大的 Qwen3 模型表现始终优于较小的模型。这表明，借助多模态思维链数据增强后，大模型能更好地发挥其推理能力，从而改善决策和排序。相比之下，4B 和 8B 模型的表现不太稳定，这说明尽管它们具备一定的多模态推理能力，但利用多模态思维链数据的能力仍弱于更大的模型。这凸显了模型规模对于最大限度发挥多模态思维链监督信号潜力的重要性，尤其是在结合 GRPO 优化时。1.7B 模型表现较弱，可能是因为指令微调过于严格，导致其输出与 top-k 候选项的匹配度较低 [7]。我们正在评估其他大语言模型骨干网络（例如，

LLaMA [10] 或 Gemma [45]，因为这些模型需要调整提示。我们预计这些模型也会呈现类似趋势。

4.4 案例研究（RQ3）

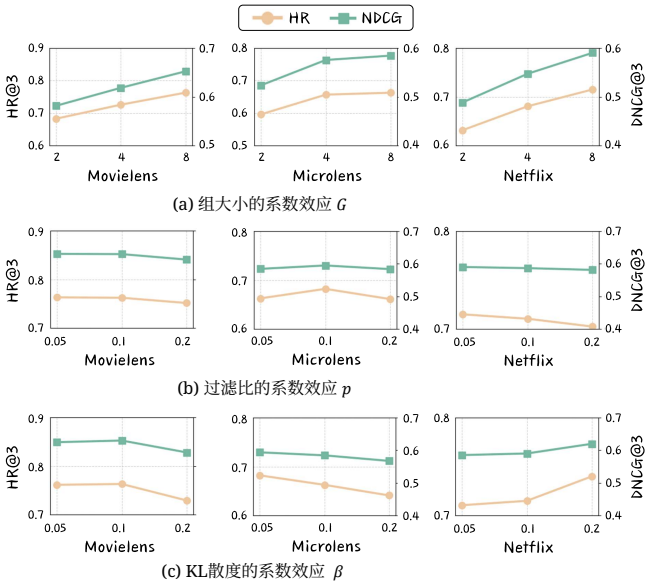
为进一步说明推理如何提升推荐准确率，我们在附录 E 中提供了案例研究。

4.5 超参数敏感性（RQ4）

我们将超参数分为三组：**GRPO 特定参数、奖励函数参数和数据过滤参数**。**（1）GRPO 特定参数**。该组包括组大小 G、采样温度、top-p 以及 KL散度系数 β。根据初步实验结果，我们将温度和 top-p 固定为 0.9，以增加组采样多样性，同时确保输出具有可用性。**（2）奖励函数**。如奖励设计部分所述，由于这一设置通过简单验证即可取得良好效果，我们不进行大规模网格搜索。**（3）数据过滤**。过滤比例控制 GRPO 中所使用样本的质量和噪声水平，是训练稳定性的关键因素。

总体而言，我们的敏感性研究聚焦于组大小 G、KL系数 β以及过滤比例 p，以评估它们对训练稳定性和最终推荐性能的影响。

- 如图 6 (a) 所示，将 G 从 2 增加到 8 会在全部三个数据集上带来命中率和 NDCG 的整体提升，这表明更大的组能够提供更稳定且信息更丰富的相对优势信号，从而改善排序质量。
- 从图中可以观察到，过于激进的过滤（更小的 p）会在部分数据集上造成轻微下降。这表明过滤能够减少噪声梯度，但过度剔除会缩小训练覆盖范围，并丢弃困难却有用的样本。
- 最后，在 (c) 部分，对 β的敏感性在不同数据集之间差异明显，但中等取值始终表现最佳。这反映了在限制策略漂移（稳定性）与允许探索性更新（有效性）之间的权衡。



(a) 组大小的系数效应 G

(b) 过滤比的系数效应 p

(c) KL散度的系数效应 β

图6：不同超参数下的性能比较。左侧的 y轴对应评估指标HR，右侧对应NDCG。

5 相关工作

面向推荐的大语言模型与多模态大语言模型。大语言模型因其先进的语言理解与推理能力，在推荐系统中受到了广泛关注。该领域的研究主要遵循三种范式：作为推荐器的大语言模型、增强器和编码器。作为推荐器的大语言模型[26, 63, 68]：这些方法将用户交互历史作为提示，引导大语言模型从候选集中选择推荐目标。TallRec [

4] 在构建的推荐数据集上微调 Llama，增强大语言模型在推荐中的决策能力。RosePO [25] 采用平滑个性化偏好优化来微调大语言模型，在提升性能的同时确保推荐系统保持“有帮助且无害”。RecZero [22] 通过使用结构化推理提示的强化学习来训练大语言模型进行推荐。作为增强器的大语言模型 [35, 47, 69]：RLMRec[38] 提出一个利用大语言模型驱动的表征学习框架，结合对比式和生成式对齐方法来改进推荐。AlphaRec [42] 用语言嵌入替代基于 ID 的嵌入，并结合图卷积网络和对比学习，构建简单而有效的推荐方法。作为编码器的大语言模型 [13, 17]：EasyRec [37] 利用协同信息以及来自用户和物品的文本数据，重新训练用于推荐的语言模型，在零样本场景中取得了出色的性能。为将基于大语言模型的推荐系统拓展至多模态场景，研究人员探索了多模态大语言模型（MLLM），通过整合视觉和文本信息，更好地捕捉用户偏好。Rec-GPT4V [29] 引入视觉摘要推理机制，以改进用户建模并减少动态图像冗余。MSRBench [72] 对不同的 MLLM 集成策略进行了系统评估，并指出计算效率低下是一项关键挑战。MLLM-MSR [58]设计了一个两阶段用户建模框架，将图像摘要与序列建模相结合，以微调多模态大语言模型。

大语言模型的后训练。作为一种广泛使用的后训练方法，监督微调（SFT） [5, 67] 对指令—示范数据进行最大似然学习，从而实现稳定的指令遵循能力和基本的推理格式。然而， [7] 表明，SFT 往往会记忆训练样本，导致其在分布外场景中的泛化能力较弱，因此人们开始采用基于强化学习（RL） [14, 59, 62]的后训练，以提升领域泛化能力。近端策略优化（PPO） [40]在限制策略偏离参考模型的同时最大化期望奖励，从而改进策略，并有效增强推理能力 [5]。基于这一强化学习基础，近期研究主要沿着两条路线展开：（1）基于直接偏好优化（DPO）的方法 [11, 32, 46, 54]；（2）基于组相对策略优化（GRPO）的方法 [30, 41, 59, 71]。DPO 从成对偏好中学习，无需像 PPO 那样进行显式奖励建模和同策略采样，从而提升了稳定性和效率。SimPO [32] 进一步对目标函数进行重新参数化，使其能够直接从偏好数据中学习，无需训练单独的奖励模型。AlphaDPO [54]

通过加强偏好建模和训练鲁棒性，降低噪声偏好和样本难度不均带来的偏差。相比之下，GRPO [41] 沿用 PPO 的同策略强化学习设置，但会根据同一提示下的多个样本计算分组相对优势，因此更适用于面向推理的优化 [14]。DAPO [59] 通过将裁剪与动态采样解耦，进一步提升稳定性和样本效率。GSPO [71]将重要性加权从词元级转移到序列级，从而减少词元级比率估计带来的不稳定性。

6 结论

本文探讨了如何有效激励大语言模型在多模态序列推荐中的推理能力。我们提出了 MLLMRec-R1，这是一个将基于GRPO的后训练与高质量多模态思维链监督相结合的统一框架。该框架离线将视觉线索转化为文本，以降低视觉token开销；构建高质量的多模态思维链监督信号；并采用混合粒度数据增强来筛除噪声轨迹、缓解捷径学习。借助轻量级奖励规则，MLLMRec-R1 在 MovieLens、MicroLens 和 Netflix 上均取得了稳定提升，展现出其有效性。

参考文献

[1] Shuai Bai、Yuxuan Cai、Ruizhe Chen、Keqin Chen、Xionghui Chen、Zesen Cheng、Lianghao Deng、Wei Ding、Chang Gao、Chunjiang Ge、Wenbin Ge、Zhifang Guo、Qidong Huang、Jie Huang、Fei Huang、Binyuan Hui、Shutong Jiang、Zhaohai Li、Mingsheng Li、Mei Li、Kaixin Li、Zicheng Lin、Junyang Lin、Xuejing Liu、Jiawei Liu、Chenglong Liu、Yang Liu、Dayiheng Liu、Shixuan Liu、Dunjie Lu、Ruilin Luo、Chenxu Lv、Rui Men、Lingchen Meng、Xuancheng Ren、Xingzhang Ren、Sibo Song、Yuchong Sun、Jun Tang、Jianhong Tu、Jianqiang Wan、Peng Wang、Pengfei Wang、Qiuyue Wang、Yuxuan Wang、Tianbao Xie、Yiheng Xu、Haiyang Xu、Jin Xu、Zhibo Yang、Mingkun Yang、Jianxin Yang、An Yang、Bowen Yu、Fei Zhang、Hang Zhang、Xi Zhang、Bo Zheng、Humen Zhong、Jingren Zhou、Fan Zhou、Jing Zhou、Yuanzhi Zhu和Ke Zhu。2025。Qwen3-VL技术报告。*arXiv*预印本 *arXiv:2511.21631*（2025）。[2] Shuai Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wenbin Ge、Sibo Song、Kai Dang、Peng Wang、Shijie Wang、Jun Tang、Humen Zhong、Yuanzhi Zhu、Mingkun Yang、Zhaohai Li、Jianqiang Wan、Pengfei Wang、Wei Ding、Zheren Fu、Yiheng Xu、Jiabo Ye、Xi Zhang、Tianbao Xie、Zesen Cheng、Hang Zhang、Zhibo Yang、Haiyang Xu和Junyang Lin。2025。Qwen2.5-VL技术报告。*arXiv*预印本 *arXiv:2502.13923*（2025）。[3] Keqin Bao、Jizhi Zhang、Wenjie Wang、Yang Zhang、Zhengyi Yang、Yanchen Luo、Chong Chen、Fuli Feng和Qi Tian。2025。面向推荐系统中大语言模型的双步定位范式。*ACM*推荐系统汇刊（*TORS*）3, 4（2025），1–27。[4] Keqin Bao、Jizhi Zhang、Yang Zhang、Wenjie Wang、Fuli Feng和Xiangnan He。2023。TallRec：一种将大语言模型与推荐对齐的高效调优框架。见*RecSys*。1007–1014。[5] Hardy Chen、Haoqin Tu、Fali Wang、Hui Liu、Xianfeng Tang、Xinya Du、Yuyin Zhou和Cihang Xie。2025。监督微调还是强化学习？对训练类R1推理型大型视觉语言模型的早期探究。*arXiv*预印本 *arXiv:2504.11468*（2025）。[6] Siran Chen、Boyu Chen、Chenyun Yu、Yuxiao Luo、Ouyang Yi、Lei Cheng、Chengxiang Zhuo、Zang Li和Yali Wang。2026。VRAgent-R1：通过强化学习利用基于多模态大语言模型的智能体提升视频推荐。见*AAAI*。[7] Tianzhe Chu、Yuexiang Zhai、Jihan Yang、Shengbang Tong、Saining Xie、Dale Schuurmans、Quoc V Le、Sergey Levine和Yi Ma。2025。监督微调负责记忆，强化学习负责泛化：基础模型后训练的比较研究。见*ICML*。[8] Junyoung Chung、Caglar Gulcehre、KyungHyun Cho和Yoshua Bengio。2014。门控循环神经网络在序列建模上的实证评估。*arXiv*预印本 *arXiv:1412.3555*（2014）。[9] Yuhao Dong、Zuyan Liu、Hai-Long Sun、Jingkang Yang、Winston Hu、Yongming Rao和Ziwei Liu。2025。Insight-V：利用多模态大语言模型探索长链视觉推理。见*CVPR*。9062–9072。[10] Abhimanyu Dubey、Abhinav J auhri、Abhinav Pandey、Abhishek Kadian、Ahmad Al-Dahle、Aiesha Letman、Akhil Mathur、Alan Schelten、Amy Yang、Angela Fan等。2024。Llama 3模型群。*arXiv*预印本 *arXiv:2407.21783*（2024）。[11] Kawin Ethayarajh、Winnie Xu、Niklas Muennighoff、Dan Jurafsky和Douwe Kiela。2024。将模型对齐视为前景理论优化。见*ICML*。

[12] Chongming Gao、Ruijun Chen、Shuai Yuan、Kexin Huang、Yuanqing Yu 与 Xiangnan He。2025。SPRec：通过自我博弈消除基于大语言模型的推荐偏差。载于 WWW。5075–5084。[13] Yunfan Gao、Tao Sheng、Youlin Xiang、Yun Xiong、Haofen Wang 与 Jiawei Zhang。2023。Chat-Rec：迈向交互式、可解释的大语言模型增强推荐系统。arXiv 预印本 arXiv:2303.14524（2023）。[14] Daya Guo、Dejian Yang、Haowei Zhang、Junxiao Song、Peiyi Wang、Qihao Zhu、Runxin Xu、Ruoyu Zhang、Shirong Ma、Xiao Bi 等。2025。DeepSeek-R1 通过强化学习激励大语言模型中的推理。《自然》645, 8081（2025），633–638。[15] Xiangnan He、Kuan Deng、Xiang Wang、Yan Li、Yongdong Zhang 与 Meng Wang。2020。LightGCN：简化并增强用于推荐的图卷积网络。载于 SIGIR。639–648。[16] Balázs Hidasi、Alexandros Karatzoglou、Linas Baltrunas 与 Domonkos Tikk。2016。基于循环神经网络的基于会话的推荐。载于 ICLR。[17] Yupeng Hou、Jiacheng Li、Zhankui He、An Yan、Xiusi Chen 与 Julian McAuley。2024。弥合语言与物品以实现检索与推荐。arXiv 预印本 arXiv:2403.03952（2024）。[18] Edward J Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang、Weizhu Chen 等。2022。LoRA：大语言模型的低秩适配。载于 ICLR。[19] Wenxuan Huang、Bohan Jia、Zijie Zhai、Shaosheng Cao、Zheyu Ye、Fei Zhao、Zhe Xu、Yao Hu 与 Shaohui Lin。2026。Vision-R1：激励多模态大语言模型的推理能力。载于 ICLR。[20] Aaron Hurst、Adam Lerer、Adam P Goucher、Adam Perelman、Aditya Ramesh、Aidan Clark、AJ Ostrow、Akila Welihinda、Alan Hayes、Alec Radford 等。2024。GPT-4o 系统卡。arXiv 预印本 arXiv:2410.21276（2024）。[21] Wang-Cheng Kang 与 Julian McAuley。2018。自注意力序列推荐。载于 ICDM。197–206。[22] Xiaoyu Kong、Junguang Jiang、Bin Liu、Ziru Xu、Han Zhu、Jian Xu、Bo Zheng、Jiancan Wu 与 Xiang Wang。2025。推荐前先思考：自主推理增强推荐系统。载于 NeurIPS。[23] Yuhan Li、Xinni Zhang、Linhao Luo、Heng Chang、Yuxiang Ren、Irwin King 与 Jia Li。2025。G-Refer：用于可解释推荐的图检索增强大语言模型。载于 WWW。240–251。[24] Jiahao Liang、Xiangyu Zhao、Muyang Li、Zijian Zhang、Wanyu Wang、Haochen Liu 与 Zitao Liu。2023。MMMLP：用于序列推荐的多模态多层感知机。载于 WWW。1109–1117。[25] Jiayi Liao、Xiangnan He、Ruobing Xie、Jiancan Wu、Yancheng Yuan、Xingwu Sun、Zhanhui Kang 与 Xiang Wang。2024。RosePO：将基于大语言模型的推荐系统与人类价值观对齐。arXiv 预印本 arXiv:2410.12519（2024）。[26] Jiayi Liao、Sihang Li、Zhengyi Yang、Jiancan Wu、Yancheng Yuan、Xiang Wang 与 Xiangnan He。2024。LLaRA：大语言模型推荐助手。载于 SIGIR。1785–1795。[27] Jianghao Lin、Xinyi Dai、Yunjia Xi、Weiwen Liu、Bo Chen、Hao Zhang、Yong Liu、Chuhan Wu、Xiangyang Li、Chenxu Zhu 等。2025。推荐系统如何从大语言模型中受益：综述。ACM 信息系统汇刊（TOIS）（2025），1–47。[28] Qidong Liu、Jiaxi Hu、Yutian Xiao、Xiangyu Zhao、Jingtong Gao、Wanyu Wang、Qing Li 与 Jiliang Tang。2024。多模态推荐系统：综述。《计算调查》57, 2（2024），1–17。[29] Yuqing Liu、Yu Wang、Lichao Sun 与 Philip S Yu。2024。Rec-GPT4V：基于大型视觉语言模型的多模态推荐。arXiv 预印本 arXiv:2402.08670（2024）。[30] Zichen Liu、Changyu Chen、Wenjun Li、Penghui Qi、Tianyu Pang、Chao Du、Wee Sun Lee 与 Min Lin。2025。理解类 R1-Zero 训练：批判性视角。arXiv 预印本 arXiv:2503.20783（2025）。[31] Jinlong Ma、Yu Zhang、Xuefeng Bai、Kehai Chen、Yuwei Wang、Zeming Liu、Jun Yu 与 Min Zhang。2026。超越单模态捷径：多模态大语言模型作为面向扎根命名实体识别的跨模态推理器。arXiv 预印本 arXiv:2602.04486（2026）。[32] Yu Meng、Mengzhou Xia 与 Danqi Chen。2024。SimPO：基于无参考奖励的简单偏好优化。载于 NeurIPS，第37卷。124198–124235。[33] Yongxin Ni、Yu Cheng、Xiangyan Liu、Junchen Fu、Youhua Li、Xiangnan He、Yongfeng Zhang 与 Fajie Yuan。2023。大规模内容驱动的微视频推荐数据集。arXiv 预印本 arXiv:2309.15379（2023）。[34] Long Ouyang、Jeffrey Wu、Xu Jiang、Diogo Almeida、Carroll Wainwright、Pamela Mishkin、Chong Zhang、Sandhini Agarwal、Katarina Slama、Alex Ray、John Schulman、Jacob Hilton、Fraser Kelton、Luke Miller、Maddie Simens、Amanda Askell、Peter Welinder、Paul F Christiano、Jan Leike 与 Ryan Lowe。2022。训练语言模型遵循带有人类反馈的指令。载于 NeurIPS，第35卷。27730–27744。[35] Yingtao Peng、Chen Gao、Yu Zhang、Tangpeng Dan、Xiaoyi Du、Hengliang Luo、Yong Li 与 Xiaofeng Meng。2025。面向推荐的大语言模型去噪对齐。ACM 信息系统汇刊（TOIS）43, 2（2025），1–35。

[36] Rafael Rafailov、Archit Sharma、Eric Mitchell、Christopher D Manning、Stefano Ermon 与 Chelsea Finn。2023。直接偏好优化：你的语言模型其实是一个奖励模型。载于 NeurIPS，第36卷。53728–53741。[37] Xubin Ren 与 Chao Huang。2025。EasyRec：简单而有效的用于推荐的语言模型。载于 EMNLP。17728–17743。[38] Xubin Ren、Wei Wei、Lianghao Xia、Lixin Su、Suqi Cheng、Junfeng Wang、Dawei Yin 与 Chao Huang。2024。面向推荐的基于大语言模型的表征学习。载于 WWW。3464–3475。[39] Lei Sang、Yu Wang、Yiwen Zhang 与 Xindong Wu。2024。面向推荐的异构图预训练去噪框架。ACM 信息系统汇刊（TOIS）（2024）。[40] John Schulman、Filip Wolski、Prafulla Dhariwal、Alec Radford 与 Oleg Klimov。2017。近端策略优化算法。arXiv 预印本 arXiv:1707.06347（2017）。[41] Zhihong Shao、Peiyi Wang、Qihao Zhu、Runxin Xu、Junxiao Song、Xiao Bi、Haowei Zhang、Mingchuan Zhang、YK Li、Yang Wu 等。2024。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）。[42] Leheng Sheng、An Zhang、Yi Zhang、Yuxin Chen、Xiang Wang 与 Tat-Seng Chua。2025。语言表征可以成为推荐系统所需：发现与潜力。载于 ICLR。[43] Aaditya Singh、Adam Fry、Adam Perelman、Adam Tart、Adi Ganesh、Ahmed El-Kishky、Aidan McLaughlin、Aiden Low、AJ Ostrow、Akhila Ananthram 等。2025。OpenAI GPT-5 系统卡。arXiv 预印本 arXiv:2601.03267（2025）。[44] Zhiqing Sun、Sheng Shen、Shengcao Cao、Haotian Liu、Chunyuan Li、Yikang Shen、Chuang Gan、Liang-Yan Gui、Yu-Xiong Wang、Yiming Yang 等。2024。通过事实增强型RLHF对齐大型多模态模型。载于 ACL。13088–13110。[45] Gemma团队、Aishwarya Kamath、Johan Ferret、Shreya Pathak、Nino Vieillard、Ramona Merhej、Sarah Perrin、Tatiana Matejovicova、Alexandre Ramé、Morgane Rivière 等。2025。Gemma 3 技术报告。arXiv 预印本 arXiv:2503.19786（2025）。[46] Fei Wang、Wenxuan Zhou、James Y Huang、Nan Xu、Sheng Zhang、Hoifung Poon 与 Muhao Chen。2024。mDPO：面向多模态大语言模型的条件偏好优化。载于 EMNLP。8078–8088。[47] Yu Wang、Lei Sang、Yi Zhang 与 Yiwen Zhang。2025。面向推荐的基于大语言模型的意图表征学习。载于 SIGIR。1870–1879。[48] Yu Wang、Yonghui Yang、Le Wu、Yi Zhang 与 Richang Hong。2025。采用自适应偏好优化的多模态大语言模型用于序列推荐。arXiv 预印本 arXiv:2511.18740（2025）。[49] Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Fei Xia、Ed Chi、Quoc V Le、Denny Zhou 等。2022。思维链提示激发大语言模型中的推理。载于 NeurIPS，第35卷。24824–24837。[50] Wei Wei、Chao Huang、Lianghao Xia 与 Chuxu Zhang。2023。面向推荐的多模态自监督学习。载于 WWW。790–800。[51] Wei Wei、Xubin Ren、Jiabin Tang、Qinyong Wang、Lixin Su、Suqi Cheng、Junfeng Wang、Dawei Yin 与 Chao Huang。2024。LLMRec：结合图增强的大语言模型用于推荐。载于 WSDM。806–815。[52] Yinwei Wei、Xiang Wang、Liqiang Nie、Xiangnan He、Richang Hong 与 Tat-Seng Chua。2019。MMGCN：用于微视频个性化推荐的多模态图卷积网络。载于 MM。1437–1445。[53] Binrui Wu、Shisong Tang、Fan Li、Bing Han、Chang Meng、Jingyu Xiao 与 Jiechao Gao。2025。面向推荐的ID与多模态表征对齐与平衡。载于 KDD。5029–5038。[54] Junkang Wu、Xue Wang、Zhengyi Yang、Jiancan Wu、Jinyang Gao、Bolin Ding、Xiang Wang 与 Xiangnan He。2025。AlphaDPO：面向直接偏好优化的自适应奖励边界。载于 ICML。[55] Xu Xie、Fei Sun、Zhaoyang Liu、Shiwen Wu、Jinyang Gao、Jiandong Zhang、Bolin Ding 与 Bin Cui。2022。面向序列推荐的对比学习。载于 ICDE。1259–1273。[56] Zhihe Yang、Xufang Luo、Dongqi Han、Yunjian Xu 与 Dongsheng Li。2025。通过直接偏好优化缓解大型视觉语言模型中的幻觉：同策略数据是关键。载于 CVPR。10610–10620。[57] Yu Ye、Junchen Fu、Yu Song、Kaiwen Zheng 与 Joemon M Jose。2025。多模态嵌入对推荐真的有益吗？整体模态与单一模态的深入探究。arXiv 预印本 arXiv:2508.07399（2025）。[58] Yuyang Ye、Zhi Zheng、Yishan Shen、Tianshu Wang、Hengruo Zhang、Peijun Zhu、Runlong Yu、Kai Zhang 与 Hui Xiong。2025。利用多模态大语言模型进行多模态序列推荐。载于 AAAI，第39卷。13069–13077。[59] Qiying Yu、Zheng Zhang、Ruofei Zhu、Yufeng Yuan、Xiaochen Zuo、Yu Yue、Weinan Dai、Tiantian Fan、Gaohong Liu、Lingjun Liu 等。2025。DAPO：大规模开源大语言模型强化学习系统。arXiv 预印本 arXiv:2503.14476（2025）。[60] Tianyu Yu、Haoye Zhang、Yuan Yao、Yunkai Dang、Da Chen、Xiaoman Lu、Ganqu Cui、Taiwen He、Zhiyuan Liu、Tat-Seng Chua 等。2024。RLAIF-V：开源人工智能反馈带来超越 GPT-4V 的可信度。arXiv 预印本 arXiv:2405.17220（2024）。

[61] Zheng Yuan、Fajie Yuan、Yu Song、Youhua Li、Junchen Fu、Fei Yang、Yunzhu Pan 和 Yongxin Ni。2023。推荐系统的下一步该走向何方？重新审视基于 ID 与基于模态的推荐模型。载于*SIGIR*。2639–2649。[62] Yang Yue、Zhiqi Chen、Rui Lu、Andrew Zhao、Zhaokai Wang、Shiji Song 和 Gao Huang。2025。强化学习是否真的能让大语言模型的推理能力超越基础模型？载于*NeurIPS*。[63] Zhenrui Yue、Sara Rabhi、Gabriel de Souza Pereira Moreira、Dong Wang 和 Even Oldridge。2023。LlamaRec：使用大语言模型进行排序的两阶段推荐。载于*CIKM* 个性化生成式人工智能研讨会。[64] Jizhi Zhang、Keqin Bao、Yang Zhang、Wenjie Wang、Fuli Feng 和 Xiangnan He。2023。ChatGPT 是否公平地对待推荐？评估大语言模型推荐中的公平性。载于*RecSys*。993–999。[65] Jingyi Zhang、Jiaxing Huang、Huanjin Yao、Shunyu Liu、Xikun Zhang、Shijian Lu 和 Dacheng Tao。2025。R1-VL：通过逐步组相对策略优化，学习使用多模态大语言模型进行推理。载于*ICCV*。1859–1869。[66] Jinghao Zhang、Yanqiao Zhu、Qiang Liu、Shu Wu、Shuhui Wang 和 Liang Wang。2021。挖掘潜在结构以实现多媒体推荐。载于*MM*。3872–3880。[67] Yang Zhang、Fuli Feng、Jizhi Zhang、Keqin Bao、Qifan Wang 和 Xiangnan He。2025。CoLLM：将协同嵌入融入用于推荐的大语言模型。*IEEE* 知识与数据工程汇刊 *(TKDE)* 37，5（2025），2329–2340。[68] Yang Zhang、Wenxin Xu、Xiaoyan Zhao、Wenjie Wang、Fuli Feng、Xiangnan He 和 Tat-Seng Chua。2026。基于强化的潜在推理，用于基于大语言模型的推荐。载于*ICLR*。[69] Yi Zhang、Yiwen Zhang、Yu Wang、Tong Chen 和 Hongzhi Yin。2025。推动协同空间与语言空间之间的分布匹配，以实现生成式推荐。载于*SIGIR*。2006–2016。[70] Keyu Zhao、Fengli Xu 和 Yong Li。2025。Reason-to-Recommend：利用交互思维推理增强大语言模型推荐。*arXiv* 预印本 *arXiv:2506.05069*（2025）。[71] Chujie Zheng、Shixuan Liu、Mingze Li、Xiong-Hui Chen、Bowen Yu、Chang Gao、Kai Dang、Yuqiong Liu、Rui Men、An Yang 等。2025。组序列策略优化。*arXiv* 预印本 *arXiv:2507.18071*（2025）。[72] Peilin Zhou、Chao Liu、Jing Ren、Xinfeng Zhou、Yueqi Xie、Meng Cao、Zhongtao Rao、You-Liang Huang、Dading Chong、Junling Liu、Jae Boum Kim、Shoujin Wang、Raymond Chi-Wing Wong 和 Sunghun Kim。2025。当大型视觉语言模型遇上多模态序列推荐：一项实证研究。载于*WWW*。275–292。[73] Xin Zhou、Hongyu Zhou、Yong Liu、Zhiwei Zeng、Chunyan Miao、Pengwei Wang、Yuan You 和 Feijun Jiang。2023。用于多模态推荐的自举潜在表征。载于*WWW*。845–854。

附录

A 时间复杂度分析

我们的框架由离线多模态思维链构建阶段和在线后训练阶段组成。离线阶段通过混合粒度数据增强生成 N 个数据点。这一调用是一次性的离线过程，对训练效率的影响极小。在在线阶段，混合粒度数据集上的监督微调遵循标准指令微调，其复杂度为 $O(NT^2d)$，其中 T 是训练序列长度，d 是骨干网络的隐藏层维度。最后，GRPO 是主要的计算瓶颈：对于每个提示，我们采样 G 个输出，平均长度为 $|o|$；同策略更新需要进行采样和一次前向—反向传播，实际复杂度为 $O(NG|o|^2d)$。我们的轻量级奖励规则对每个样本执行常数时间的字符串/ID 操作，仅增加 $O(NG)$ 的开销；与 Transformer 项相比，这一开销可以忽略不计。因此，MLLMRec-R1 的总体复杂度可概括为：$O\left(NT^2d + NG|o|^2d\right)$。

算法1MLLMRec-R1 算法

输入： 用户序列 S^u、目标物品 v^u_{t+1}、物品元数据 $m_v = \{I_v、T_v\}$、混合粒度集合 $D_{mix} = \{(x^u、y^u、\bar{r}^u)\}$、策略 π_θ 以及 π_{ref}、组大小 G、解码参数（温度、top-p）、以及奖励规则 $R(\cdot)$。

输出： 思维链优化 $\widetilde{R} = \{\tilde{r}^u\}$ 以及优化后的策略 π_θ。

阶段一：多模态思维链构建
对于每个物品 $v \in V$ **执行**
  构建图像描述 c_v。 ▷ 公式 (3)
结束循环
对于每个具有历史记录 S^u 和目标 v^u_{t+1} 的用户 $u \in U$ **执行**
  获取伪思维链 r^u。 ▷ 公式 (4)
    在 r^u中掩蔽目标信息。
    获取思维链优化 $\tilde{r}^u$。 ▷ Equ (5)
结束 for
阶段二：GRPO 后训练
对于 每个训练样本 $(x^u, y^u, \bar{r}^u) \in D_{mix}$ **执行**
  采样 G 个响应$\{o_i\}_{i=1}^G \sim \pi_\theta(\cdot \mid x^u, \bar{r}^u)$，温度为(temperature，top-$p$)。
    使用 $r_i = R(o_i, y^u)$ 计算奖励 $\{r_i\}_{i=1}^G$。▷ 公式 (8)
    计算组相对优势 $\{\hat{A}_i\}_{i=1}^G$ 和 GRPO 损失并进行 KL 正则化，得到 π_{ref}。▷ 公式 (2)
    通过最小化 GRPO 损失来更新 π_θ。
结束循环

B 实验设置详情

B.1 数据集

我们在三个多模态数据集上评估了模型。表3展示了所有数据集的统计概览。

• **MicroLens** [33] 是一个短视频推荐数据集，包含标题、封面图像和视频信息。在本研究中，大语言模型仅使用标题和图像。用户交互的平均数量为 8.79。

• **Netflix** [51] 来自 LLMRec 在 Kaggle 上收集的数据，是一个基于用户的电影推荐数据集。不过，

表3：实验数据集的统计信息。

数据集	#用户	#物品	#交互	密度
MicroLens	25,411	41,081	223,263	2.1×10^{-4}
Netflix	13,187	17,366	68,933	3.0×10^{-4}
MovieLens	6,040	3,952	1,000,209	4.2×10^{-2}

该数据集非常稀疏，每位用户平均只与 5.23 个物品发生交互。

• **MovieLens** [39] 是一个经典的电影推荐数据集，其中每部电影都包含标题、年份和对应的海报。该数据集非常稠密，平均序列长度为 165.56。

B.2 基线方法

我们的 MLLMRec-R1 采用 Qwen 作为骨干网络，用作序列推荐器。因此，我们将基线方法分为传统方法、多模态方法、基于大语言模型的方法和基于多模态大语言模型的推荐方法。

传统推荐系统。 GRU4Rec [16] 使用门控循环单元 [8] 构建会话级序列模型，用于下一次点击预测。SASRec [21] 使用自注意力机制对用户的历史行为序列进行建模。LightGCN [15] 移除了图卷积网络中的特征变换和非线性激活，以实现轻量级推荐。CL4SRec [55] 率先将对比学习引入序列推荐。

多模态推荐系统。 LATTICE [66] 构建潜在物品-物品图，以提升多模态推荐性能。MoRec[61] 探索纯模态信息能否取代传统物品ID表示。BM3 [73]提出了一种自监督多模态模型，无需负采样即可学习表征。AB-Rec [53]对齐ID和多模态表征，以实现有效的多模态推荐。

基于大语言模型的推荐系统。 TallRec [4] 提出对大语言模型进行监督微调，使其适配推荐任务。LLaRA [26] 设计了一种推荐助手，将协同过滤信号与语言模型相结合。SPRec [12]利用大语言模型中的自博弈机制进行去偏，并引导模型产生更均衡的输出。RecZero [22] 通过带有结构化推理提示的强化学习来训练大语言模型以执行推荐。

基于多模态大语言模型的推荐系统。 MSRBench [72] 评估视觉-语言多模态大模型在序列推荐中的能力。MLLM-MSR [58] 对多模态大语言模型进行微调，使其专门用于序列推荐任务。

C 主实验的扩展

C.1 采用更多候选项的标准设定。

为评估大规模候选检索，我们通过采样 99 个负样本，将候选集增加到 100。**不进行重新训练，直接测试在标准设置下训练的模型**，并在表 4 中报告结果。我们纳入了几种经典方法作为基线方法，并排除现有的基于多模态大语言模型的方法，因为它们的视觉 token开销使推理难以扩展到更大的候选集。如结果所示，与排名第二的 RecZero 相比，MLLMRec-R1 在更大候选集下取得了更大的相对提升，尤其是在 NDCG 上。这一优势源于多模态思维链推理监督和 GRPO 的组相对优化：前者提供细粒度的偏好信号，并减少由幻觉或低置信度轨迹造成的噪声更新

表4：更多候选项下的性能比较。

数据集	模型	HR@5	HR@10	NG@5	NG@10
MovieLens	CL4SRec [55]	0.1427	0.1991	0.1084	0.1198
	TallRec [4]	0.1317	0.1930	0.0930	0.1108
	RecZero [22]	0.1963	0.2418	0.1432	0.1530
	MLLMRec-R1	**0.2444**	**0.2983**	**0.1806**	**0.1969**
	提升 %	24.50%	23.37%	26.12%	28.69%
Netflix	CL4SRec [55]	0.1080	0.1872	0.0671	0.0937
	TallRec [4]	0.0980	0.1790	0.0591	0.0836
	RecZero [22]	0.1963	0.2418	0.1432	0.1530
	MLLMRec-R1	**0.1715**	**0.2560**	**0.1199**	**0.1512**
	提升 %	38.87%	15.06%	56.91%	41.84%

表5：MovieLens数据集上不同历史长度的性能比较 t。

长度	模型	HR@5	HR@10	NG@5	NG@10
5	CL4SRec [55]	0.5919	0.7492	0.4822	0.5452
	TallRec [4]	0.5993	0.7566	0.4935	0.5592
	RecZero [22]	0.6161	0.7275	0.5436	0.5639
	MLLMRec-R1	**0.7119**	**0.8353**	**0.6077**	**0.6574**
	提升 %	15.55%	10.40%	11.79%	16.58%
10	CL4SRec [55]	0.6221	0.7557	0.5306	0.5861
	TallRec [4]	0.6139	0.7324	0.5236	0.5613
	RecZero [22]	0.6595	0.7664	0.5639	0.6067
	MLLMRec-R1	**0.7630**	**0.8368**	**0.6524**	**0.6784**
	提升 %	15.69%	9.19%	15.82%	11.82%
20	CL4SRec [55]	0.6224	0.7521	0.5469	0.5817
	TallRec [4]	0.5637	0.7199	0.4765	0.5391
	RecZero [22]	0.6325	0.7455	0.5613	0.5958
	MLLMRec-R1	**0.7433**	**0.8369**	**0.6404**	**0.6760**
	提升 %	17.52%	12.26%	14.09%	13.46%

，后者则学习到更具区分性的排序行为，即使负样本数量大幅增加，也能更好地将目标物品保持在靠前位置 [14, 65]。

C.2 不同长度的标准设置。

我们进一步考察了历史长度 t 对 MovieLens 上性能的影响，结果汇总于表5。对于基于大语言模型的模型，性能随着 t增加而提升，在 $t = 10$达到峰值，随后下降，这表明 GRPO 和思维链的收益并非单调递增：更长的历史记录可能会引入关联性较弱的交互和噪声，从而产生收益递减。相比之下，序列模型在 $t = 10$ 和 $t = 20$ 时的表现相近，这可能是因为它们更加依赖局部模式，并且从额外上下文中获得的收益较少。此外，设置 $t = 20$ 会移除历史记录较短的用户，从而减少训练数据并削弱监督信号。

D 数据质量评估详情

为评估构建的思维链数据的质量和潜在风险，我们设计了一套结合自动评估与人工评估的混合协议，如图10所示。我们从生成的思维链优化集和伪思维链集中随机抽取200段推理文本，并邀请 GPT-5.2 [43], Claude-4.5 [5], 和人工标注者 [23] 从五个维度对每个样本进行独立评分——模态一致性、预测一致性、信号密度、泄漏风险和覆盖难度。我们将结果汇总为图7中的雷达图。这样做的主要动机是，思维链构建可能会引入与目标相关的捷径信号，从而人为抬高训练奖励并损害泛化；因此，必须明确评估并控制潜在的信息泄漏和噪声。此外，我们的过滤和置信度建模强调模态一致性与预测一致性，以减轻跨模态不匹配和幻觉推理所导致的噪声学习信号。我们发现，多模态思维链在所有维度上的得分都高于伪思维链。结合性能提升，这表明质量更高的思维链监督能更有效地增强模型的推理能力 [19, 62]。

E 案例研究

本节展示了 MovieLens数据集中用户1127的案例研究。图7（左）展示了推理提示模板，其中仅包含交互历史和候选集，没有明确加入多模态思维链。图中间对比了纯文本GRPO与 MLLMRec-R1 使用 Qwen3“思考”格式生成的输出。纯文本 GRPO的 Top-3 结果中有两部电影来自同一个《十三号星期五》恐怖系列，这表明若缺少多模态思维链和 MDA，模型可能会依赖浅层共现信号（例如系列重复），从而错过真正的目标物品。相比之下，MLLMRec-R1 返回了更多样且语义一致的候选项，并将目标物品排在 Top-1。这表明，多模态思维链能够提供细粒度线索，帮助 MLLMRec-R1 更可靠地区分组内选项，从而提升排序推荐质量。

最后，右侧面板提供了更细致的对齐分析。我们将用户的历史封面序列作为多模态证据，观察到了一些共同的视觉线索，包括大多为黑白和低饱和度风格，以及严肃的人物中心构图。这些线索表明，用户偏好经典时代色调和道德驱动的人性冲突。基于这一用户画像，MLLMRec-R1 将候选项与推断出的用户表征进行匹配，并将《西线无战事》排在第 1 位。Metropolitan（1990）和 Blood Simple（1984）也名列前茅，因为它们分别通过以对话为核心的人际关系和冷硬黑色叙事，契合用户对严肃、以人物为驱动的故事的偏好。总体而言，多模态思维链提供了细粒度信号，并结合 GRPO 的组内学习，减少了系列作品重复和类型关键词等捷径，从而使排名更加稳定且具有更强的泛化能力。

F 奖励收敛

在前 0.2–0.5 个训练轮次内，平滑奖励迅速上升，表明策略很快学会与用户偏好对齐。之后，增益逐渐放缓，奖励在约 2–3 个训练轮次后趋于稳定。Netflix 起初的奖励较低，但提升幅度更大，波动也更明显，这表明其优化问题更具挑战性，但仍能从后训练中显著获益。相比之下，MovieLens 和 MicroLens 的收敛过程更加平稳，后期增益较小，说明随着策略接近局部最优，收益逐渐递减。

[5]https://claude.ai/

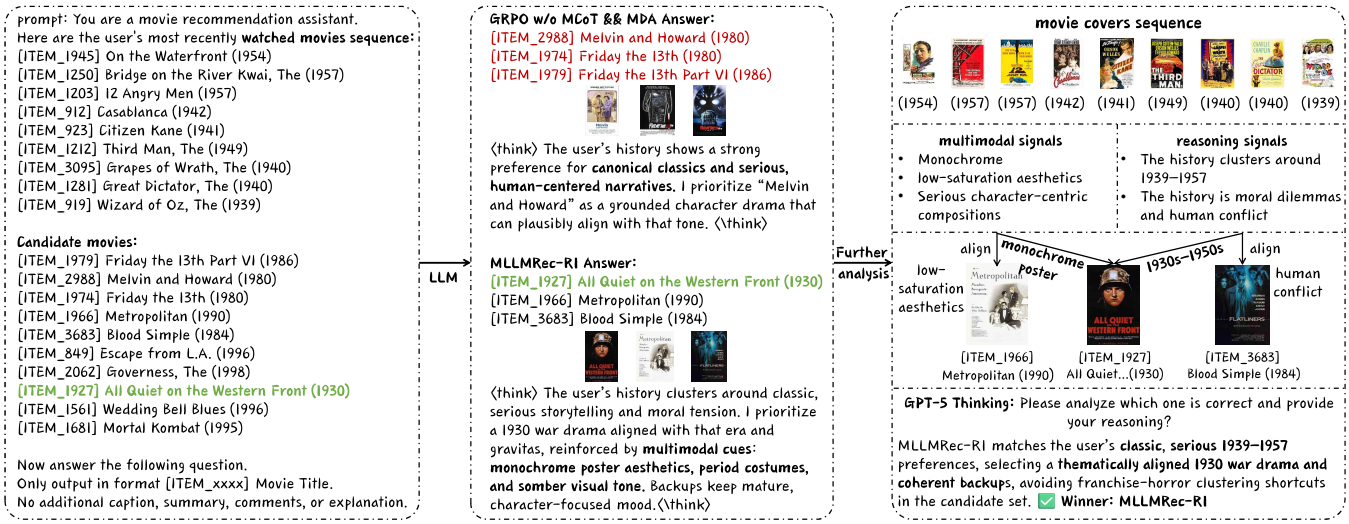


图 7：MovieLens 测试集中的序列 1127 案例研究。"GRPO w/o MCoT && MDA"指移除了多模态思维链推理和数据增强模块的训练变体。绿色电影为目标物品。

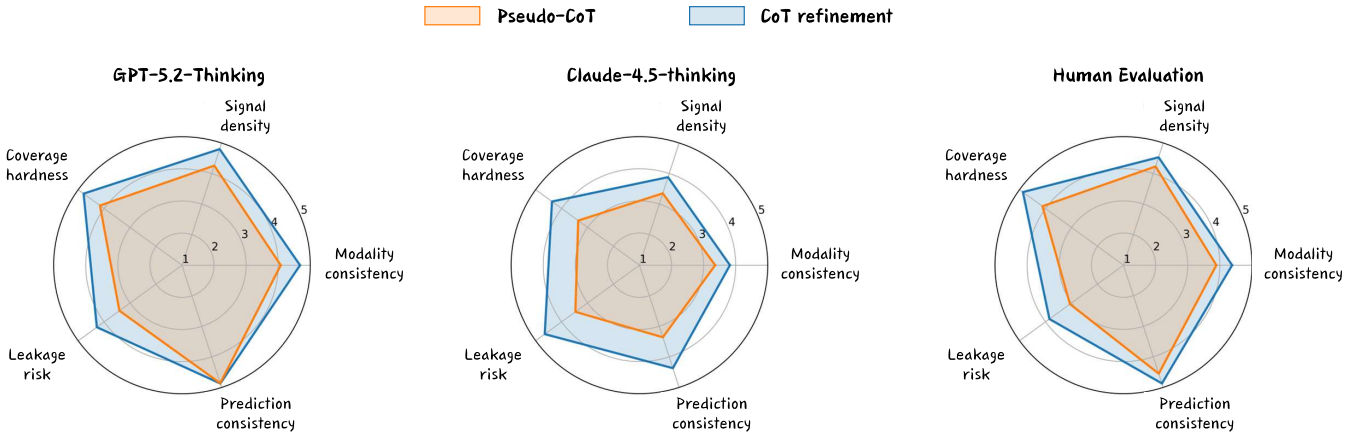


图 8：伪思维链与思维链优化在五个数据质量维度上的雷达图对比，分数为三个数据集上的平均值。我们报告了 GPT-5.2-thinking、Claude-4.5-thinking 和人工评估下的结果。

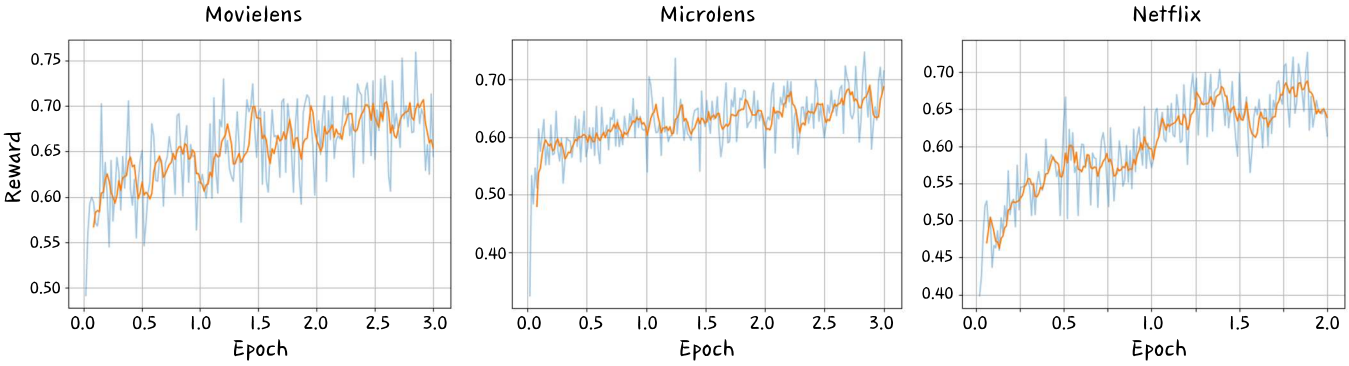


图 9：三个基准数据集上基于 GRPO 的训练奖励曲线。

多模态序列推荐数据质量评估提示

角色。 你是一名“多模态序列推荐数据质量”评估者。你将收到一段思维链（CoT）推理文本。你的任务是从五个数据质量维度对其进行评分，并为每项评分提供依据。

待评估对象。

- **输入：**一段推理文本。

评分要求。

- 为每个维度评定 1 到 5 的整数分数。
- 为每个维度提供恰好一句证据/理由，且只能引用给定的推理文本。
- 输出必须是严格的 JSON（不得添加额外说明、Markdown 或多个代码块）。

五个维度的评分标准。 +

(1) 模态_一致性。 评估推理中的偏好信号是否具体且能够以多模态输入（标题/图像/图注）为依据，避免含糊或无法验证的说法。

- **5:** 具体且可验证的描述，能够与图像相对应。
- **3:** 总体相关，但略显笼统；细节有限；部分内容难以验证。
- **1:** 大多是模糊或类似模板的陈述，或明显缺乏依据、相互矛盾的描述。

（2）预测_一致性。 评估各步骤是否在逻辑上支持结论，避免“跳至结论”。

- **5:** 结论明确承接了多个步骤中的检索标准。
- **3:** 主线逻辑一致，但存在缺失环节或论据薄弱的问题。
- **1:**存在矛盾，或结论与先前的偏好画像无关。

(3) 信号_密度. 评估推理是否提供了足够丰富、细粒度且可学习的偏好线索，而非通用模板。

- **5:**多个细 -粒度的偏好维度，对候选项具有较强的区分能力 .
- **3:** 有一些偏好线索，但较为粗略（例如，仅“喜欢科幻动作”）；区分度中等。
- **1:** 几乎没有可学习的偏好信息，或高度模板化/空洞的内容。

(4) 泄露_风险. 评估文本是否包含捷径信号，从而无需真正的偏好推理即可进行猜测。

- **5:** 没有明确的目标信息；停留在“个人资料/标准”层面；并未唯一指向某个答案 er.
- **3:**没有明确的标题/ID，但包含高度可识别的组合或过强的提示。
- **1:** 包含目标标题/ID、候选项位置提示，或任何对答案的明确泄露。

(5) 覆盖_难度. 评估该样本是否为训练提供有用的难度与覆盖（区分相似候选项、捕捉长尾偏好），而不是过于简单或冗余。

- **5:** 有助于区分同质候选项的细粒度约束；具有“难而有用”的训练价值。
- **3:** 中等难度，信息价值中等。
- **1:** 过于简单（候选项显然可分）或重复度过高，边际训练收益较低。

输出格式（严格 JSON）。

{ "modality_consistency": {"score": 1-5, "evidence": "..."}, "prediction_consistency": {"
score": 1-5, "evidence": "..."}, "signal_density": {"score": 1-5, "evidence": "..."}, "leakage_
risk": {"score": 1-5, "evidence": "..."}, "coverage_hardness": {"score": 1-5, "evidence": "..."} }

现在开始评估。 请仅根据以下推理文本进行评分： <REASONING_TEXT>

图10：用于自动评估多模态思维链数据质量的提示。