⁻⁴ 通过强化学习的视频时间定位的数据集和方法

Ruizhe Chen^{1,2*}, Zhiting Fan², Tianze Luo¹, Heqing Zou¹, Zhaopeng Feng², Guiyang Xie¹, Hansheng Zhang¹, Zhuochen Wang¹, Zuozhu Liu², Huaijian Zhang¹

¹ Bytedance

² Zhejiang University

Link: Code & Model

视频时序定位(VTG)旨在根据自然语言查询定位视频中的相关时间段。尽管大型视觉语言模型 (LVLMs)和指令微调取得了近期进展,但现有方法常常面临时间感知有限和泛化能力欠佳的问题。 在这项工作中,我们引入了一种结合监督微调和强化学习(RL)的双阶段训练框架,以提高 VTG 模型的准确性和鲁棒性。我们的方法首先利用高质量精心挑选的冷启动数据进行 SFT 初始化,随后 通过难度控制的 RL 进一步增强时间定位和推理能力。针对多个 VTG 基准的全面实验表明,我们 的方法在具有挑战性和开放域的场景中,始终优于现有模型。我们对训练策略和数据集编排进行了深 入分析,强调了高质量冷启动数据和难度控制的 RL 的重要性。为促进进一步的研究和工业应用,我 们向社区发布了所有中间数据集、模型和代码。

1. 介绍

随着社交媒体平台的增多,视频内容已成为捕捉和传达日常体验的信息最丰富和多样的媒介。因此, 基于用户查询高效地识别视频中特定时刻的任务——被称为视频时态定位(VTG)——已成为从智能 视频检索到工作流程优化及自动事件监测等各种工业应用的核心能力 [16, 19, 27]。VTG 使从业人员 可以迅速定位大型视频中相关的片段,大大减少了人工审核的工作量,并支持实时决策 [39]。

近年来,大型视觉-语言模型(LVLMs)的进展推动了端到端时间定位框架的发展。像 TimeChat [35]、VTimeLLM [23]和 LITA [24]这样的指令调优模型将时间定位重新表述为文本生成任务,而像 Momentor [34]和 VTG-LLM [18]这样的模型则引入了专门的模块或词汇以改善时间感知。尽管取 得了显著进展,现有的方法仍受到监督微调固有局限的制约,难以精准地掌握时间和实现泛化。

为了解决这些挑战,我们提出了一种新颖的双阶段训练框架,将监督微调(SFT)与强化学习(RL) 相结合,以显著提高开源模型在 VTG 任务中的性能和泛化能力。我们的框架首先利用高质量的整理 数据通过 SFT 为模型提供稳健的冷启动初始化,随后通过一个难度控制的 RL 阶段进一步增强时序 定位能力和推理能力。

我们在多个 VTG 基准上进行了广泛的实验,系统地评估了每个训练阶段的贡献。我们的研究结果 突出了高质量冷启动数据和受控 RL 训练的关键重要性,为在实际工业场景中的实用部署提供了可操 作的见解。此外,为了促进未来的研究和应用,我们将所有中间结果和代码作为开源资源发布。

这项工作的主要贡献是:

- 我们引入了一个两阶段的训练框架,将 SFT 和 RL 结合在一起,以推进开源的 LVLMs 用于视频时间定位。
- 我们在多个基准上进行了全面评估,验证了我们方法的有效性和可扩展性。
- 我们开源所有中间数据集、模型和代码,以支持进一步的研究和工业应用。

^{*}Work done during his internship at Bytedance.



Fig. 1: 所提出的视频时间定位训练流程概述(VTG-R1)。该框架首先使用经过精心选择的高质量冷启动数据进行监督微调(SFT)以初始化基础模型,然后通过强化学习(RL)进一步增强时间定位能力。

2. 相关工作

视频时序定位(VTG)旨在根据自然语言查询在未剪辑的视频中定位相关的时间片段 [9,16,20,27,44] 。早期的研究,如 CTRL 和 MCN,引入了基础方法,利用滑动窗口和双流网络来生成文本视频匹配 的候选片段 [15,19],为后续的进步奠定了基础。

随着大视觉-语言模型(LVLMs)的出现,最近的研究转向了利用指令调优和文本生成的端到端 VTG 框架。TimeChat [35]、VTimeLLM [23]和 LITA [24]等模型将时间定位重新表述为一个序 列生成任务,而 Momentor [34]则通过引入时间感知模块解决时间量化误差问题。其他方法,包括 Grounded-VideoLLM 和 VTG-LLM [18,41],扩展了模型词汇表以促进时间嵌入学习,进一步提高 定位精度。

VTG 技术在多个领域显示出实际价值。在制造业中,VTG 支持自动化工作流程分析和异常检测,以提高运营效率 [26]。在安全监控方面,VTG 能够快速检索关键事件,支持实时监控和事后调查 [39]。在医疗保健领域,VTG 促进大型手术视频中关键步骤的高效识别,有利于临床分析和教育 [40]。

尽管取得了这些进展,但主要依赖于监督微调(SFT)通常会限制模型的时间感知和泛化能力,特别是在开放领域或具有挑战性的场景中。为了解决这些限制,我们提出了一个两阶段的训练框架,将监督微调与强化学习结合起来,旨在提高 VTG 模型的准确性和泛化能力。为了支持进一步的研究和应用,我们将所有中间数据、模型和代码作为开源资源发布。

Task	# Original Samples	Source Datasets	# Coldstart Samples	# RL Samples
Instance Grounding (Moment Retrieval)	40K	HiREST [47] (4K), QuerYD [32] (33K), TACoS [14] (10K), DiDeMo [1] (33K), InternVid-VTime [43] (54K)	10K	13K
Query Grounding	16K	Grounded-VLLM [41] (16K)	3К	$5\mathrm{K}$
Total	56K	-	13K	18K

Table 1: 源数据集以及过滤后的冷启动和 RL 数据集的统计信息。

3. 数据集和食谱

在本节中,我们通过一个两阶段训练流程展示构建 VTG-R1 的详细过程,包括数据收集、整理和具体的训练步骤。

3.1. 数据收集和整理

高质量的冷启动和强化学习数据集对于提升多模态大型语言模型(MLLMs)的时间视频定位能力至关重要。这里,我们描述了我们收集源数据的方法,并为强化学习训练策划了 TVG-RL-18K 数据集,为基于监督微调的冷启动策划了 TVG-Coldstart-13K 数据集。

我们汇总了来自各种公共数据集的数据,包括用于时刻检索和查询定位任务的数据集,仔细抽样和平衡每个子集的比例。TVG-RL-18K和 TVG-Coldstart-13K的原始源数据分布进行了分类和汇总,见表格 1。

CoT 标注和数据过滤。为了实现有效的监督微调(SFT)冷启动,我们使用 Gemini-2.5-Pro 为源样本生成连锁思维(CoT)推理。CoT 生成所用的提示模板如下,并在 SFT 和 RL 阶段保持一致应用。然后我们根据最终的交并比(IoU)分数筛选注释样本: IoU 为 > ϵ_1 的样本被视为高质量,并保留其 CoT 推理用于冷启动,形成 TVG-Coldstart-13K 子集。相反, IoU 为 < ϵ_2 的源样本被视为低质量——通常是由于过于困难或注释错误——并在 RL 阶段被排除。剩余样本构成 TVG-RL-18K 子集。

Prompt Template for TVG-R1

system You MUST reason based on the temporal changes and visual evidence in the video to determine the precise time period related to the query. The reasoning MUST reflect how the content evolves over time, not general logic. The reasoning process MUST BE enclosed within $\langle \text{ think } \rangle \langle /\text{think } \rangle$ tags. The specific time period MUST BE in the format [start time, end time] in seconds enclosed within $\langle \text{ time } \rangle \langle /\text{time } \rangle$ tags. user { query / instance }

3.2. 监督微调阶段

在我们的训练流程的第一阶段,我们采用监督微调(SFT)来为模型提供一个高质量的初始化,被称为冷启动阶段。这个过程使模型从一开始就具备强大的多模态对齐和结构化推理能力,为后续的强化学习阶段奠定了坚实的基础。

Madal	NExTGQA			RexTime				
Model	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU
Qwen2.5-VL-7B thinking	25.81	14.73	8.72	17.74	12.16	7.17	2.71	10.17
Qwen 2.5-VL-7B	31.60	18.06	7.46	20.87	10.31	6.08	3.04	8.10
Qwen 2.5-VL-32B	37.96	22.26	9.98	25.35	16.83	9.99	5.10	13.02
VTimeLLM	37.90	20.20	9.71	24.40	28.84	17.41	7.19	20.14
TimeChat	34.10	17.90	6.24	20.60	14.42	7.61	3.06	11.65
VideoChat-TPO	41.20	23.40	8.15	27.70	34.53	19.26	9.81	25.23
TVG-ColdStart	21.74	11.54	5.24	15.09	13.57	7.82	4.34	10.18
TVG-R1	41.65	20.78	10.01	29.25	41.04	24.54	11.07	28.20

Table 2: 在 NExTGQA 和 RexTime 基准上的性能比较。可以观察到, VTG-R1 优于使用大规模数据训练的 现有 SFT 方法。

3.3. 强化学习 (RL) 阶段

3.3.1. 奖励建模

奖励 r_i 在指导模型的学习目标中起着关键作用。为了通过显式推理促进有效的时间定位,我们使用了一个由两个组成部分构成的复合奖励函数: IoU 奖励 r_{tIoU} 和推理格式奖励 r_{form}。

时间戳感知 IoU 奖励 $r_{tIoU}(\cdot)$ 在 TVG 任务中, 预测时间段 $[t_s, t_e]$ 的质量主要使用交并比 (IoU) 度量进行评估, 该度量衡量预测段与真实段 $[t'_s, t'_e]$ 之间的重叠。IoU 计算如下:

$$r_{\text{tIoU}} = \frac{[t_s, t_e] \cap [t'_s, t'_e]}{[t_s, t_e] \cup [t'_s, t'_e]}$$

其中 ∩ 和 ∪ 表示预测与真实区间的交集和并集。

推理格式奖励 *r***form**(·) 为了明确地鼓励模型生成具有结构化推理的响应,我们引入了一种基于格式的奖励 *r***form**,该奖励验证输出是否遵循预期的推理格式。具体来说,我们要求模型将推理过程包括在 <**think**>...</**think**>标签内,并将最终答案包括在 <**answer**>...</**answer**>标签内。奖励定义为:

 $r_{\rm form} = \mathbb{1}_{\{ < \text{think}, </ \text{think}, < \text{answer}, </ \text{answer} \} \subseteq \text{output}}$

其中 1. 表示指标函数。

最终奖励 r_i 最终的奖励 r_i 被定义为两个组件的加权和:

 $r_i = \lambda_{tIoU} \cdot r_{tIoU} + \lambda_{form} \cdot r_{form}$

,其中 λ_{tIoU} 和 λ_{form} 是超参数。

我们采用组相对策略优化(GRPO) [37]进行强化学习,这是一种近端策略优化(PPO) [36]的变体。与依赖于学习到的评论者的 PPO 不同,GRPO 直接比较一组候选响应,消除了对评论者模型的需求,从而减少了计算开销。

对于一个给定的查询 q , GRPO 从策略中抽取 G 个不同的候选响应 $o = \{o_1, \ldots, o_G\}$ 。每个响应 的奖励如 Sec. 3.3.1 中描述的那样分配,产生 $\{r_1, \ldots, r_G\}$ 。这些分数随后在组内进行标准化,每个 响应的优势定义为:

$$A_i = \frac{r_i - \mu}{\sigma}, \quad \text{where } \mu = \frac{1}{G}\sum_{j=1}^G r_j, \sigma = \sqrt{\frac{1}{G}\sum_{j=1}^G (r_j - \mu)^2}.$$

其中, A_i 表示 *i*-th 响应的标准化优势。GRPO 鼓励模型在组内为相对更好的响应分配更高的概率。 最终的训练目标还包括一个 KL 散度正则项,以防止更新后的策略 π_{θ} 从参考策略 π_{ref} 发生显著偏离。完整的目标由以下公式给出:

$$\mathcal{L}_{\text{GRPO}} = \mathbb{E}_{o \sim \pi_{\theta_{\text{old}}}(p)} \left[\sum_{i=1}^{G} \frac{\pi_{\theta}(o_i)}{\pi_{\theta_{\text{old}}}(o_i)} \cdot A_i - \beta \cdot D_{\text{KL}} \left(\pi_{\theta} \parallel \pi_{\text{ref}} \right) \right],$$

其中 β 是控制与参考策略偏差的正则化系数。



Fig. 2: 比较带有高质量冷启动和不带冷启动的 RL 训练曲线。TVG-R1 具有高质量冷启动,收敛到更高的分数,这表明冷启动的好处在于解锁模型的潜力并增强其推理能力,正如训练期间响应长度的增加所示。

Model	Filter	NExTG.	RexT.	Charad.			
TVG-Coldstart-13k Dataset							
Qwen 2.5-VL-7B	-	20.87	8.10	46.14			
TVG-ColdStart	-	26.14	26.26	42.19			
TVG-R1-U	×	23.92	29.14	29.57			
TVG-R1	\checkmark	30.41	26.38	48.78			
TVG-R1-Zero	-	27.76	26.00	48.75			

Table 3: 验证高质量冷启动数据的有效性。TVG-R1-U 指的是在未筛选数据上进行冷启动。结果表明,TVG-R1 优于 TVG-R1-U,突显了高质量 SFT 数据的重要 性。

Model	Filter	NExTG.	RexT .	Charad.			
TVG-RL-18k Dataset							
Qwen2.5-VL-7B	-	20.87	8.10	46.14			
TVG-R1	×	27.88	25.91	46.96			
TVG-R1	\checkmark	30.41	26.38	48.78			
TVG-R1-Zero	×	5.49	24.18	20.32			
TVG-R1-Zero	\checkmark	27.76	26.00	48.75			

Table 4: 验证 RL 数据有效性。TVG-R1-Zero 指的是 跳过 SFT 冷启动,直接进行 RL 训练。结果显示, RL 数据过滤提高了模型性能,尤其是在没有冷启动的情 况下。

4. 实验

4.1. 实验设置

我们在三个基准上进行综合实验,以评估我们方法的有效性。具体来说,我们报告了在 ReXTime [22], NExT-GQA [28],和 Charades-STA [21]数据集上的结果。为了进行评估,我们采用了用于时间视频定位 (TVG)的 R1@m 指标。R1@m 表示在预测的前 1 段中,交并比 (IoU)大于阈值 *m* 的实例 比例,其中 $m \in 0.3, 0.5, 0.7$ 。此外,我们报告了所有样本的平均 IoU (mIoU) 作为 TVG 准确性的整体指标。

我们将我们的方法与几个强力的基线进行比较,包括经过指令微调的时间定位模型,如VTimeLLM [23], TimeChat [35],和 VideoChat-TPO [45],以及通用多模态大型模型,如 Qwen2.5-VL 7B 和 32B [2]。对于标有"thinking"的模型,我们采用 TVG-R1 提示模板来指导时间定位。

所有实验都是在 16 个 NVIDIA H100 (80GB) GPU 上进行的。在训练和推理中,我们将视频帧 数量限制为 64 帧,每帧处理的分辨率为 $128 \times 28 \times 28$ 像素。主干模型为 Qwen2.5-VL-7B [2] 。超 参数 $\epsilon_1 \ \pi \epsilon_2 \ D$ 别设定为 0.8 和 0.4。我们首先在 TVG-Coldstart-13K 数据集上进行一轮的监督微调 (SFT),以获得 TVG 冷启动模型。接下来,我们在 TVG-RL-18K 数据集上应用强化学习 (RL)以 获得最终的 TVG-R1 模型,其中 KL 散度项中 GRPO 算法的超参数 β 设置为 0.0。最大响应长度设 置为 2048 个标记,损失权重 λ_{tou} 和 $\lambda_{form} \ D$ 别设定为 0.9 和 0.1。由于计算资源的限制, RL 训练 被限制在 600 步。更多细节可以在附录中找到。

Model	R@0.3	R@0.5	R@0.7	mIoU
Base Base thinking	$68.98 \\ 36.48$	$\frac{48.18}{21.83}$	$22.87 \\ 9.76$	$46.14 \\ 23.48$
VTimeLLM TimeChat VideoChat-TPO	$55.3 \\ 51.5 \\ 58.3$	$34.3 \\ 32.2 \\ 40.2$	$14.7 \\ 13.4 \\ 18.4$	34.6 - 38.1
TVG-ColdStart TVG-R1	42.23 70.75	$29.38 \\ 50.46$	$14.95 \\ 23.92$	$28.91 \\ 46.73$

Table 5: Charades 数据集上的性能比较

4.2. 主要结果

如表 2 和 5 所示,我们在三个不同的基准上进行的实验表明 VTG-R1 在视频时间定位任务上的有效 性。可以得出两个关键观察结果。

VTG-R1 的卓越表现: VTG-R1 在大多数基准测试中始终优于以前的模型,突显了在解决视频时间定位难题中明确推理的重要性。这些结果进一步强调了强化学习在提高模型性能方面的影响。

强化学习的重要性:基于 SFT 的模型 TVG-ColdStart 在性能提升上并不一致,有时甚至在 SFT 后表现出轻微下降,这可能是由于过拟合或对未见场景的泛化能力有限所致。相比之下,VTG-R1 在 经过强化学习后实现了显著的改进,这强烈表明 RL 对于开发能够有效泛化的稳健推理能力至关重要。

4.3. 分析

为了更深入地了解不同变体的影响,我们展示了在其他配置下的实验结果。具体来说,我们分析了与 冷启动过程和 RL 数据选择相关的变体。

Max Length	NExTGQA	RexTime	Charades
$\begin{array}{c} 2048 \\ 1024 \end{array}$	$30.41 \\ 21.80$	$26.38 \\ 25.71$	$48.78 \\ 41.38$
512	24.09	24.91	46.31

Table 6: 冷启动长度对性能的影响。RL 训练后的结果显示,在冷启动期间较长的响应长度对模型的优化更有利。

发现 1: 高质量的冷启动数据至关重要。如图 2 所示,我们比较了 TVG-R1 和 TVG-R1-Zero 的 RL 训练曲线,其中 TVG-R1-Zero 指的是跳过 SFT 冷启动而直接进行 RL 训练。可以观察到,无论是训练期间的总奖励还是测试集的表现,TVG-R1 都收敛到更高的分数,这表明高质量的冷启动有助于在 RL 阶段释放模型的潜力。此外,如图 2 (b) 所示,冷启动初始化的模型在开始时表现出更高的响应 长度,并且在整个训练过程中增加更为显著。这表明冷启动增强了模型的推理能力,使其能够通过更 详细的推理推导出正确的答案。

我们进一步通过限制 Gemini-2.5-Pro 的最大输出长度,来检查冷启动响应长度对模型性能的影响。 我们重新标注了不同的冷启动数据集,并在表格 6 中报告了 RL 训练后的最终结果。结果表明,在冷 启动阶段较长的响应长度对模型优化更有利。

此外,如表 3 所示,我们比较了 TVG-R1 和 TVG-R1-U,其中 TVG-R1-U 表示使用未经过滤的 56K 数据集进行冷启动,然后进行 RL。请注意,所有 RL 操作均在 TVG-RL-18K 数据集上进行。结果显示 TVG-R1 明显优于 TVG-R1-U,这表明选择高质量的冷启动数据比单纯增加训练数据的数量 更有效于学习稳健的推理能力。

发现二:控制 RL 训练数据的难度是必要的。如表 4 所示,我们比较了在 TVG-R1 和 TVG-R1-Zero 设置下,有无数据过滤的 RL 训练结果。注意,TVG-R1 是通过 TVG-Coldstart-13K 数据集初始化 的。结果表明,在没有冷启动的情况下,模型在未经过滤的数据上进行训练会很难学习,而数据过滤 则会显著提高性能。此外,对于通过冷启动初始化的模型,过滤 RL 数据进一步有利于模型优化。这 些发现表明,如果训练数据在早期阶段过于困难或令人困惑,模型可能很难学习并达到收敛。

在这项工作中,我们提出了一种新颖的两阶段训练框架用于视频时间定位(VTG),以增强大型视觉语言模型的能力。对多个基准的广泛实验表明,高质量的冷启动数据和难度控制的 RL 训练对于提高模型性能和泛化能力都至关重要。我们的方法被证明是可扩展的,并且在实际部署中有效。

5.

局限性

尽管我们提出的框架在视频时间定位(VTG)方面显示了显著的改进,仍然存在一些限制。首先, 该方法严重依赖高质量的、经过精心准备的冷启动数据,这在某些领域或资源匮乏的情境中可能难以 获得。其次,强化学习阶段引入了相当大的计算开销,这可能会限制资源有限的组织或学术用户的可 访问性。未来的工作应该探索提高数据效率的方法,在资源有限的环境中优化强化学习,并扩大这种 训练范式在更复杂或多样化的多模态任务中的适用性。

References

- Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2017.
- [2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923, 2025.
- [3] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.
- [4] Ruizhe Chen, Jianfei Yang, Huimin Xiong, Jianhong Bai, Tianxiang Hu, Jin Hao, Yang Feng, Joey Tianyi Zhou, Jian Wu, and Zuozhu Liu. Fast model debias with machine unlearning. Advances in Neural Information Processing Systems, 36:14516–14539, 2023.
- [5] Ruizhe Chen, Tianxiang Hu, Yang Feng, and Zuozhu Liu. Learnable privacy neurons localization in language models. *arXiv preprint arXiv:2405.10989*, 2024.
- [6] Ruizhe Chen, Yichen Li, Jianfei Yang, Joey Tianyi Zhou, Jian Wu, and Zuozhu Liu. Identifying and mitigating social bias knowledge in language models. *arXiv preprint arXiv:2408.11843*, 2024.
- [7] Ruizhe Chen, Xiaotian Zhang, Meng Luo, Wenhao Chai, and Zuozhu Liu. Pad: Personalized alignment of llms at decoding-time. *arXiv preprint arXiv:2410.04070*, 2024.
- [8] Ruizhe Chen, Wenhao Chai, Zhifei Yang, Xiaotian Zhang, Joey Tianyi Zhou, Tony Quek, Soujanya Poria, and Zuozhu Liu. Diffpo: Diffusion-styled preference optimization for efficient inference-time alignment of large language models. arXiv preprint arXiv:2503.04240, 2025.
- [9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony M. H. Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. arXiv preprint arXiv:2305.06500, 2023.
- [10] Zhiting Fan, Ruizhe Chen, Tianxiang Hu, and Zuozhu Liu. Fairmt-bench: Benchmarking fairness for multi-turn dialogue in conversational llms. *arXiv preprint arXiv:2410.19317*, 2024.
- [11] Zhiting Fan, Ruizhe Chen, Ruiling Xu, and Zuozhu Liu. Biasalert: A plug-and-play tool for social bias detection in llms. *arXiv preprint arXiv:2407.10241*, 2024.
- [12] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. arXiv preprint arXiv:2503.21776, 2025.
- [13] Zhaopeng Feng, Shaosheng Cao, Jiahan Ren, Jiayuan Su, Ruizhe Chen, Yan Zhang, Zhe Xu, Yao Hu, Jian Wu, and Zuozhu Liu. Mt-r1-zero: Advancing llm-based machine translation via r1-zero-like reinforcement learning. arXiv preprint arXiv:2504.10160, 2025.

- [14] Tian Gan, Xiao Wang, Yan Sun, Jianlong Wu, Qingpei Guo, and Liqiang Nie. Temporal sentence grounding in streaming videos. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4637–4646, 2023.
- [15] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision* (ICCV), pages 5277–5285. IEEE, 2017. doi: 10.1109/ICCV.2017.563.
- [16] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, and Xingyu Liu et al. Ego4D: Around the world in 3,000 hours of egocentric video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 18995–19012, 2022.
- [17] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948, 2025.
- [18] Yongxin Guo, Jingyu Liu, Mingda Li, Dingxin Cheng, Xiaoying Tang, Dianbo Sui, Qingbin Liu, Xi Chen, and Kevin Zhao. VTG-LLM: Integrating timestamp knowledge into video LLMs for enhanced video temporal grounding. arXiv preprint arXiv:2405.13382, 2024.
- [19] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan C. Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5804–5813. IEEE, 2017.
- [20] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan C. Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5804–5813, 2017.
- [21] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan C. Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5804–5813, 2017.
- [22] Bin Huang and et al. Rextime: Temporal grounding benchmark for reasoning-intensive videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 14313–14323, 2024.
- [23] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 14271–14280, 2024.
- [24] De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. Lita: Language instructed temporal-localization assistant. In Computer Vision – ECCV 2024, Part XXX (Proc. 18th European Conference on Computer Vision), pages 202–218. Springer, 2024.
- [25] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. arXiv preprint arXiv:2503.06749, 2025.

- [26] Hao Li, Shuhui Wang, Ying Zhou, and et al. Vision-based abnormal event detection in industrial manufacturing processes: A review. *Computers in Industry*, 130:103469, 2021. doi: 10.1016/j. compind.2021.103469.
- [27] Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355, 2023.
- [28] Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. In arXiv preprint arXiv:2305.06355, 2023.
- [29] Ye Liu, Kevin Qinghong Lin, Chang Wen Chen, and Mike Zheng Shou. Videomind: A chain-oflora agent for long video reasoning. arXiv preprint arXiv:2503.13444, 2025.
- [30] Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.
- [31] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, et al. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. CoRR, 2025.
- [32] Andreea-Maria Oncescu, Joao F Henriques, Yang Liu, Andrew Zisserman, and Samuel Albanie. Queryd: A video dataset with high-quality text and audio narrations. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2265–2269. IEEE, 2021.
- [33] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35:27730– 27744, 2022.
- [34] Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat-Seng Chua, Yueting Zhuang, and Siliang Tang. Momentor: Advancing video large language model with fine-grained temporal reasoning. In Proceedings of the 41st International Conference on Machine Learning (ICML), 2024.
- [35] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14313–14323, 2024.
- [36] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. A simpler and effective first-order alternative to TRPO via surrogate objective and clipping.
- [37] Zhihong Shao et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. Introduces Group Relative Policy Optimization (GRPO) as a PPO variant enhancing reasoning ability and memory efficiency.

- [38] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. Advances in neural information processing systems, 33:3008–3021, 2020.
- [39] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 6479–6488, 2018.
- [40] Ap Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel de Mathelin, and Nicolas Padoy. Endonet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE Transactions on Medical Imaging*, 36(1):86–97, 2017. doi: 10.1109/TMI.2016.2593957.
- [41] Haibo Wang, Zhiyang Xu, Yu Cheng, Shizhe Diao, Yufan Zhou, Yixin Cao, Qifan Wang, Weifeng Ge, and Lifu Huang. Grounded-videollm: Sharpening fine-grained temporal grounding in video large language models. arXiv preprint arXiv:2410.03290, 2024.
- [42] Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, et al. Time-r1: Post-training large vision language model for temporal video grounding. arXiv preprint arXiv:2503.13377, 2025.
- [43] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. arXiv preprint arXiv:2307.06942, 2023.
- [44] Zhenzhi Wang, Limin Wang, Tao Wu, Tianhao Li, and Gangshan Wu. Negative sample matters: A renaissance of metric learning for temporal grounding. In *Proceedings of the AAAI Conference* on Artificial Intelligence, volume 36, pages 2613–2623, 2022.
- [45] Ziang Yan, Zhilin Li, Yinan He, Chenting Wang, Kunchang Li, Xinhao Li, Xiangyu Zeng, Zilei Wang, Yali Wang, Yu Qiao, et al. Task preference optimization: Improving multimodal large language models with vision task alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29880–29892, 2025.
- [46] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. arXiv preprint arXiv:2503.14476, 2025.
- [47] Abhay Zala, Jaemin Cho, Satwik Kottur, Xilun Chen, Barlas Oguz, Yashar Mehdad, and Mohit Bansal. Hierarchical video-moment retrieval and step-captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 23056–23065, 2023.
- [48] Yaowei Zheng, Junting Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, and Yuwen Xiong. Easyr1: An efficient, scalable, multi-modality rl training framework. https://github.com/ hiyouga/EasyR1, 2025.

A. 实现细节

我们使用 gemini-2.5-pro-preview-05-06 API 进行标注,并将最大长度设置为 8192。对于冷启动,选择重叠度(IoU)大于 0.8 的样本。

冷启动阶段 我们在 TVG-Coldstart 数据集上微调基础模型。微调在 8 个 H100 GPU 上进行, 批量 大小为 8, 训练 1 个周期。学习率设置为 1e-6。

RL 阶段 我们基于 Easy-R1 [48] 实现进行强化学习训练。最大响应长度设置为 2048。批量大小设置为 128,并训练 600 步。GRPO 样本数 G 设置为 8。

A.1. 实验

评估使用官方的 VideoMind [29] 实现进行。最大响应长度设置为 2048 个标记,所有其他推理超参数 保持在 transformers 库提供的默认值。

B. 定性结果

我们在图 3 中为 TVG-R1 提供了定性案例。这些全面的数据记录了视频片段理解和定位任务中时间 定位模型的推理过程和预测结果。每条数据包括基本视频信息、查询、标注的时间跨度、模型的逐步 推理过程和预测的时间跨度。推理内容通常提供了视频中关键事件和动作节点顺序的详细描述,帮助 模型明确目标片段的起始和结束点。预测结果通过使用如 IoU 等指标与真实时间跨度进行比较来评 估。此类数据不仅强调了多步推理和时间线索的结合,还反映了模型在具体案例中的定位能力,为性 能评估和分析提供了坚实的基础。



(a) Case Study 1: Find the time period when the little girl runs back.

(b) Case Study 2: Find the time period when the person takes off his coat.

Fig. 3: TVG-R1 评价案例。

B.1. TVG-Coldstart 数据集案例

我们在图 4 中提供了 TVG-Coldstart 数据集的定性案例。这些冷启动数据样本采用逐步推理的方法 来记录模型在时间定位任务中的决策过程。数据涵盖了视频片段中关键动作和事件节点的识别,清楚 地展示了模型如何分析每个片段并筛选事件以精确确定查询所需的时间段。此类数据强调多步推理结合时间线索,为后续的视频理解模型训练和评估提供高质量的推理样本。



(a) Case Study 1: Find the time period the boy climbs the steps.



(b) Case Study 2: Find the time period writing cash registers.

Fig. 4: TVG-Coldstart 数据集案例。

C. 其他相关工作

早期研究表明,通过人类反馈的强化学习(RLHF)可以有效地使大型语言模型(LLMs)符合人类的偏好,主要是确保大型模型遵循人类的意图和价值观 [3-8,10,11,33,38]。最近的研究更加关注于用于具有确定性答案的任务的可验证奖励强化学习(RLVR) [13,30,46]。

作为一个开创性的开源大型语言模型, DeepSeek-R1 [17] 采用了生成奖励驱动策略优化(GRPO) [37] 来增强其推理性能,利用精心设计的规则奖励,整合了推理模板和最终结果。在大型视觉语言模型的背景下,最近的方法已将 GRPO 应用于多模态图像推理任务,从而显著改善了图像理解 [25,31] 和视频理解 [12,42]。