WHAT IS THE EVENT AND WHEN DOES IT HAPPEN?

Anonymous authors
Paper under double-blind review

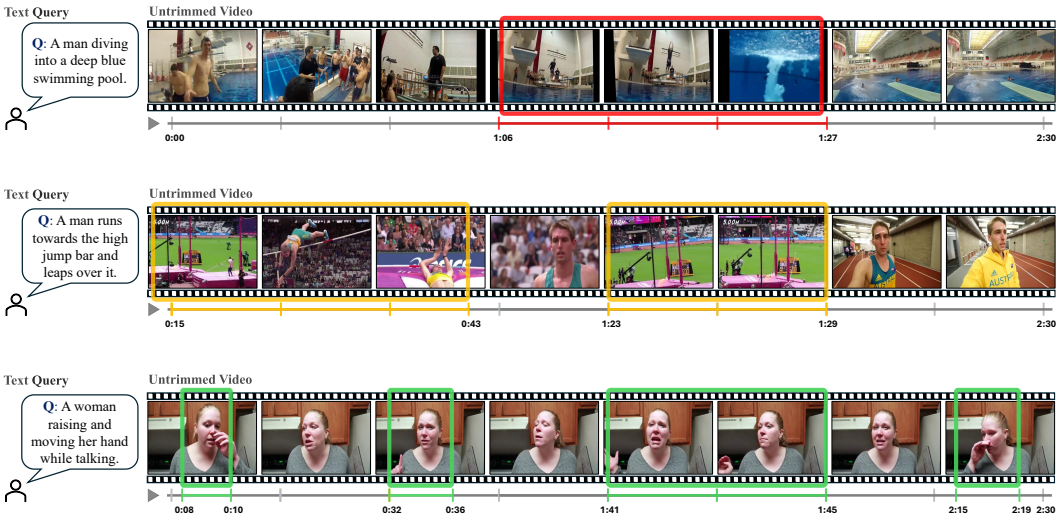


Figure 1: **M-R1** processes an untrimmed video and a natural-language query, showcasing versatile localization capabilities across varying temporal durations and numbers of relevant moments.

ABSTRACT

Video Moment Retrieval fundamentally demands mapping a continuous visual stream onto discrete temporal intervals. This distinctive characteristic requires a discriminative latent space capable of determining *what* event is being described, *when* it occurs, and *which* candidate is most plausible. However, existing approaches remain largely optimized through localization-oriented objectives that supervise final temporal predictions, leaving the latent space only indirectly constrained with respect to such discrimination. In this paper, we explicitly decompose multi-dimensional discriminability into a triad of core axes: *semantic*, *temporal*, and *relational discrimination*. Building on this view, we introduce **M-R1**, a reinforcement learning-based framework that rethinks moment retrieval as the joint resolution of semantic, temporal, and ranking ambiguities. Specifically, we first propose MM-GRPO, which encourages the model to identify modality-specific semantic evidence and selectively regulate its participation in cross-modal interactions. We further devise TA-MOD, which structures temporal information flow within the fused representation via query-conditioned affinity modulation to preserve temporal discriminability. Finally, we present Rank-GRPO, which models inter-candidate relations and optimizes the relative ordering of competing moments. Experimental results on large-scale VMR benchmarks demonstrate significant improvements, with M-R1 achieving average performance gains of 5.4% on multi-moment retrieval and 32.4% on short-moment localization over previous state-of-the-art methods, showing its robustness in challenging scenarios.

1 INTRODUCTION

Video understanding in continuous streams requires models to resolve not only what occurs, but also where semantically relevant events lie along the temporal axis. Video Moment Retrieval (VMR) embodies this challenge by localizing the temporal segment described by a natural-language query within an untrimmed video. Effective moment retrieval therefore hinges on both rich multimodal understanding and precise temporal localization, motivating advances along these dimensions.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Recently, LLM-based approaches have emerged as a promising direction for VMR by leveraging large-scale pretrained knowledge and reasoning capabilities (Yang et al., 2025; Zheng et al., 2026b; Zhang et al., 2026b; Zheng et al., 2026a). Such methods enable zero-shot generalization to unseen queries and scenarios by leveraging rich multimodal priors (Ren et al., 2024; Qian et al., 2024; Guo et al., 2025b; Jeon et al., 2026), but remain costly due to autoregressive inference and large model scale. These limitations have motivated continued research on non-LLM approaches (Li et al., 2023; Yang et al., 2024; Wei et al., 2025; Li et al., 2026), which provide more architectural flexibility while maintaining strong in-domain performance. QD-DETR, TR-DETR, and DualGround (Moon et al., 2023b; Sun et al., 2024; Kang et al., 2026) redesign their cross-modal interaction architectures to explicitly incorporate query information into video representations. R^2-Tuning and FlashVTG (Liu et al., 2024; Cao et al., 2025) further reshape temporal modeling through coarse-to-fine and multi-scale feature aggregation. BAM-DETR (Lee & Byun, 2024) extends this architectural flexibility to the prediction stage, replacing center-length estimation with a boundary-oriented formulation.

Despite these architectural advances, these methods still face several fundamental limitations. We argue that these limitations stem from a deeper issue: supervision on final localization outcomes only indirectly constrains how the latent space encodes retrieval-critical distinctions. Consequently, models can satisfy the localization objective without fully exploiting the information needed to distinguish the target moment. As illustrated in Fig. 2, predictions often remain largely unchanged even when the temporal order of the video is perturbed or the language query is altered, suggesting that the model may rely on shortcut cues rather than a semantic understanding of the event. Moreover, the inherent temporal redundancy of video makes temporal discrimination particularly difficult. As shown in Fig. 4, consecutive frames often exhibit high visual similarity and can induce excessive feature propagation through self-attention, weakening fine-grained temporal boundaries. At the ranking stage, candidate scores are primarily learned from localization or confidence supervision, leaving their relative ordering only indirectly constrained. Together, these observations show that localization supervision alone does not ensure a discriminative latent space for moment retrieval.

To address these limitations, we introduce **M-R1**, a reinforcement learning-based framework for non-LLM video moment retrieval. Our key idea is to decompose the discriminative process of moment retrieval into three core dimensions through semantic interaction, temporal propagation, and candidate comparison. Inspired by the relative optimization principle of Group Relative Policy Optimization (GRPO) (Shao et al., 2024), we design M-R1 to explicitly structure the retrieval-critical distinctions that localization supervision leaves under-constrained. However, GRPO is not directly compatible with non-LLM VMR, since deterministic retrieval models do not naturally produce the diverse rollouts available in autoregressive LLMs. To bridge this gap, we parameterize retrieval decisions as stochastic policies and formulate role-specific group-relative objectives.

For semantic discrimination, we introduce MultiModal-GRPO (MM-GRPO), which constructs counterfactual multimodal variations over event structure and action semantics. On the video side, we introduce two complementary corruptions, where intra-event perturbations shuffle the temporal order within the target event and inter-event perturbations modify its pre- and post-event context. In parallel, we construct an action corruption on the text side by removing the action predicate while preserving the remaining query content. A higher reward is assigned when successful localization exhibits stronger dependence on these counterfactual variations. This objective encourages the model to ground retrieval in event structure and action semantics rather than shortcut cues.

Building on the resulting multimodal representation, we introduce Temporal Affinity Modulation (TA-MOD) for temporal discrimination. TA-MOD modulates temporal affinities within the fused representation according to query relevance, selectively regulating information flow along the temporal sequence. By suppressing excessive propagation induced by the high visual similarity of consecutive frames, TA-MOD maintains distinct temporal boundaries and fine-grained distinctions.

Finally, we introduce Rank-GRPO for relational discrimination between competing moment candidates. Rank-GRPO first samples multiple candidate rankings from a stochastic policy, then constructs alternative orderings by prioritizing similarly scored moment candidates. We design a preference-aware objective that assigns higher rewards to orderings that preserve the relative priority of more relevant moments over competing alternatives. Comparing rewards across sampled rankings directly optimizes preference relations between candidates. This encourages the model to prioritize query-relevant moments, yielding a more discriminative ranking structure.

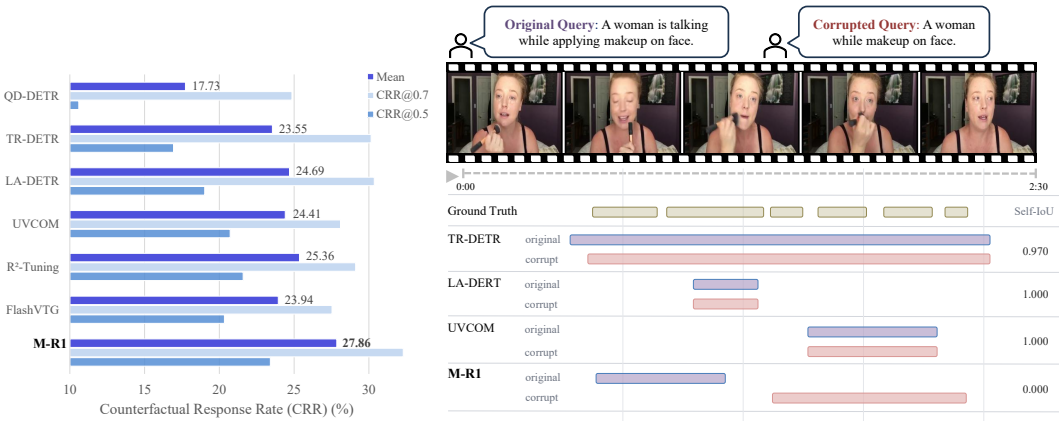


Figure 2: Counterfactual sensitivity analysis comparing candidate-level responses to semantic changes and prediction shifts after removing action semantics from the query.

Extensive experiments and analyses across VMR benchmarks validate M-R1 across semantic, temporal, and relational discrimination. Our contributions are summarized as follows.

1. To the best of our knowledge, we are the first to rethink moment retrieval by decomposing its discriminability into three dimensions: semantic, temporal, and relational discrimination.

2. We introduce **M-R1**, a novel reinforcement learning-based framework that enables GRPO for deterministic VMR models by parameterizing retrieval decisions as stochastic training policies.

3. Extensive experiments and analyses demonstrate consistent improvements across diverse VMR benchmarks and validate the effectiveness of MM-GRPO, TA-MOD, and Rank-GRPO.

2 Related Work

2.1 Video Moment Retrieval

LLM-Based. Recent advances in multimodal large language models have enabled new approaches to video temporal grounding through rich multimodal priors and reasoning capabilities. A series of studies has enhanced temporal awareness through time-aware representations (Ren et al., 2024; Guo et al., 2025b; Deng et al., 2025) and temporal instruction learning (Huang et al., 2024; Qian et al., 2024). Other approaches explicitly structure the grounding process through event-centric modeling, temporal identification, expert specialization, and decoupled grounding formulations (Guo et al., 2024; Wu et al., 2025; Yang et al., 2025; Zeng et al., 2025). More recent studies explore temporal search and reasoning, including recursive coarse-to-fine localization (Hannan et al., 2025), temporal chain-of-thought reasoning (Zhang et al., 2025), and difficulty-aware routing (Zhang et al., 2026b).

Non-LLM-Based. In parallel with recent advances in LLM-based approaches, non-LLM methods continue to be explored for their efficiency and architectural flexibility. Inspired by the success of DETR (Carion et al., 2020) in 2D object detection, VMR methods have extended DETR-style architectures to temporal localization (Sun et al., 2024; Xiao et al., 2024; Tang et al., 2025; Sun et al., 2025a; Zhang et al., 2026a). CG-DETR (Moon et al., 2023a) suppresses query injection into irrelevant clips using text-conditioned dummy tokens, while SE-DETR (Hu et al., 2026) improves semantic distinguishability through cluster-derived semantic proxies. Beyond DETR-style architectures, alternative methods have diversified temporal modeling through multi-scale representations and discrete moment modeling (Liu et al., 2024; Cao et al., 2025; Sun et al., 2025b; An & Grauman, 2026). KDA further extracts query-relevant knowledge through multi-stage feature processing and diffusion-based background suppression (Ran et al., 2025). Complementary to architectural design, recent methods improve localization by diversifying training samples through augmentation and synthesis (Wei et al., 2025; Park et al., 2026; Zhou et al., 2025). Despite increasingly sophisticated architectures and training strategies, these methods remain primarily optimized by matching predictions to ground-truth temporal intervals. Such localization supervision specifies where the target lies, while leaving the distinctions underlying reliable retrieval only indirectly constrained. In contrast to these approaches, M-R1 decouples the learning of semantic, temporal, and relational distinctions through dedicated learning objectives, allowing each to be optimized separately and explicitly.

2.2 Policy Optimization with Reinforcement Learning

Reinforcement learning (RL) provides a principled framework for optimizing decision-making policies based on reward signals. Policy optimization pursues this objective by directly updating a parameterized policy to maximize expected returns. Trust Region Policy Optimization (TRPO) (Schulman et al., 2015) introduced trust-region constraints to limit overly large policy updates, while Proximal Policy Optimization (PPO) (Schulman et al., 2017) adopted a simpler clipped surrogate objective. To reduce the memory overhead of the separate value model used in PPO, Group Relative Policy Optimization (GRPO) (Shao et al., 2024) estimates advantages from the relative rewards of multiple sampled outputs within each group. Based on these advances, RL-based policy optimization has been increasingly adopted in probabilistic models with sampling processes, including LLMs (Guo et al., 2025a; Yu et al., 2026), MLLMs (Huang et al., 2026; Liu et al., 2025; Meng et al., 2025; Tan et al., 2026) and generative models (Black et al., 2024; Ren et al., 2025; Liu et al., 2026; Yuan et al., 2025). Video-R1 (Feng et al., 2026) proposes T-GRPO, which assigns additional rewards to correct predictions that exhibit dependence on temporal information. Time-R1 (Wang et al., 2026) applies GRPO to temporal video grounding with task-specific rewards for timestamp accuracy and output formatting. Inspired by these advances, we introduce M-R1, the first policy optimization framework for non-LLM VMR. Unlike existing RL-based approaches built on probabilistic models with natural sampling processes, deterministic models do not inherently provide such rollout spaces. M-R1 bridges this gap by formulating retrieval decisions as stochastic policies and tailoring policy optimization to the distinct discriminative requirements of moment retrieval.

3 M-R1

We introduce **M-R1**, a reinforcement learning-based framework that decomposes moment retrieval discriminability into three complementary dimensions: semantic, temporal, and relational, with each dimension explicitly optimized. As illustrated in Fig. 3, these dimensions are instantiated by MM-GRPO, TA-MOD, and Rank-GRPO, respectively. We first introduce the formulation of Group Relative Policy Optimization and discuss its applicability to deterministic moment retrieval models (Sec. 3.1). MM-GRPO constructs counterfactual multimodal variations to encourage retrieval dependence on event structure and action semantics (Sec. 3.2), while TA-MOD modulates temporal affinities within the fused representation using query-conditioned relevance (Sec. 3.3). Rank-GRPO further organizes competing moment proposals into a shared group for relative ranking optimization (Sec. 3.4). The overall training strategy is described in (Sec. 3.5).

3.1 Background: Group Relative Policy Optimization

Group Relative Policy Optimization (GRPO) (Shao et al., 2024) optimizes a policy using the relative rewards of multiple sampled outputs without requiring a separate value model. For LLMs, the policy π_θ is the language model itself, which generates a response $o_i = (o_{i,1}, \ldots, o_{i,|o_i|})$ conditioned on an input q. GRPO samples a group of G responses $\{o_i\}_{i=1}^{G}$ from the old policy $\pi_{\theta_{\text{old}}}$ and assigns each response a reward r_i. The normalized advantage $\hat{A}_i$ and policy ratio $\rho_{i,t}$ are then computed as

$$\hat{A}_i = \frac{r_i - \text{mean}(\{r_j\}_{j=1}^{G})}{\text{std}(\{r_j\}_{j=1}^{G})}, \qquad \rho_{i,t} = \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}. \tag{1}$$

Based on the group-relative advantage and policy ratio, GRPO optimizes the policy by maximizing the following clipped surrogate objective with KL regularization toward a reference policy:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim P(Q),\, \{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{G}\sum_{i=1}^{G} \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \left\{ \min\left[\rho_{i,t}\hat{A}_i,\ \text{clip}(\rho_{i,t}, 1-\epsilon, 1+\epsilon)\hat{A}_i\right] - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \right\} \right] \tag{2}$$

Here, ϵ controls the clipping range for stable policy updates, while β scales the KL regularization that constrains deviation from the reference policy π_{ref}. GRPO consequently derives its learning signal from reward variations among multiple outputs sampled under the same input. In contrast to this stochastic generation process, deterministic VMR models map each video-query pair to a fixed retrieval prediction and thus provide no inherent rollout diversity for group-relative comparison. Bridging this discrepancy motivates a stochastic formulation of retrieval decisions.

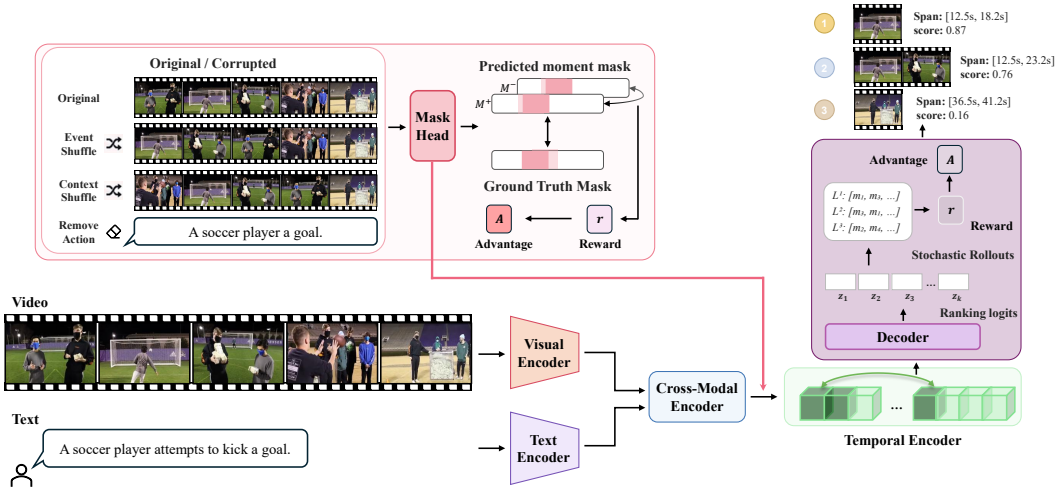


Figure 3: Overall pipeline of M-R1, consisting of MM-GRPO for semantic discrimination, TA-MOD for temporal discrimination, and Rank-GRPO for preference discrimination.

3.2 MultiModal Group Relative Policy Optimization

Semantic discrimination requires understanding the queried event through its visual dynamics and linguistic semantics. To this end, we introduce MultiModal Group Relative Policy Optimization (MM-GRPO), which encourages the model to distinguish and exploit visual and linguistic cues.

For temporal evidence selection to be directly optimized, we formulate a stochastic mask policy over the fused representation. Specifically, given a video-query pair $X = (V, Q)$, we obtain its fused representation and predict a temporal mask logit $\ell_{\theta,t}(X)$ at each temporal position t. Each logit is then converted into a selection probability, from which the temporal mask is sampled as

$$p_{\theta,t}(X) = \sigma\left(\frac{\ell_{\theta,t}(X)}{\tau_m}\right), \qquad m_t \sim \text{Bernoulli}(p_{\theta,t}(X)). \tag{3}$$

where σ is the sigmoid function, τ_m is the sampling temperature, and $\{m_t\}_{t=1}^{T}$ is the sampled mask.

The sampled mask separates temporal foreground from background, but does not guarantee that the mask prediction reflects an understanding of the queried event. Motivated by Video-R1 (Feng et al., 2026), which contrasts ordered and globally shuffled frames to strengthen temporal reasoning in video question answering, we construct counterfactual variations designed for moment retrieval.

We construct the visual counterfactual by locally shuffling either the intra-event temporal structure or the pre/post-event context while preserving the ground-truth interval. We also derive a linguistic counterfactual by selectively suppressing action-related semantics and retaining the remaining query context. We then pair the original input $X^+ = (V, Q)$ with one of these counterfactuals X^- and form a group of G rollouts. Within each rollout g, the original and counterfactual branches produce sampled masks $\{m^+_{g,t}\}_{t=1}^{T}$ and $\{m^-_{g,t}\}_{t=1}^{T}$, respectively, using the same random variable $U_{g,t} \sim \mathcal{U}(0,1)$ at each temporal position to suppress variation caused by independent mask sampling.

Based on these paired rollouts, we measure the IoU of $\{m^+_{g,t}\}_{t=1}^{T}$ and $\{m^-_{g,t}\}_{t=1}^{T}$ with the ground-truth interval Y. We define the reward r_g to increase when the original mask accurately captures the target region and exhibits sensitivity to the perturbed modality-specific semantics, as follows:

$$r_g = \text{IoU}\big(\{m^+_{g,t}\}_{t=1}^{T}, Y\big) \cdot s\big(\text{IoU}\big(\{m^+_{g,t}\}_{t=1}^{T}, Y\big) - \text{IoU}\big(\{m^-_{g,t}\}_{t=1}^{T}, Y\big)\big) \tag{4}$$

where $s(\cdot)$ is a zero-anchored monotonic function that maps the IoU margin to $[0, 1]$. The resulting rewards $\{r_g\}_{g=1}^{G}$ are converted into group-relative advantages following Eq. 1.

We treat each sampled temporal mask as a structured action and define its policy score $\ell_g(\theta)$ as

$$\ell_g(\theta) = \frac{1}{T}\sum_{t=1}^{T}\left[m^+_{g,t}\log p_{\theta,t}(X^+) + (1 - m^+_{g,t})\log\big(1 - p_{\theta,t}(X^+)\big)\right]. \tag{5}$$

The corresponding policy ratio is computed as $\rho_g = \exp(\ell_g(\theta) - \ell_g(\theta_{\text{old}}))$ and optimized using the GRPO objective in Eq. 2. Through this optimization, MM-GRPO encourages temporal mask decisions to reflect a multimodal understanding of the queried event. Notably, the counterfactual branch is used only for reward construction, and optimization is applied to the original mask action.

3.3 Temporal Affinity Modulation

Beyond semantic discrimination, temporal discrimination remains challenging because adjacent video features are inherently correlated. This correlation can cause interactions within the fused representation to smooth fine-grained distinctions across adjacent moments. To address this issue, we introduce Temporal Affinity Modulation (TA-MOD), which incorporates the query-conditioned relevance encoded in the MM-GRPO mask logits into interactions within the fused representation.

Specifically, we estimate query relevance at each temporal position from the attention mass on real query tokens, with learned dummy tokens as alternative targets (Moon et al., 2023a). Combined with the fused representation, this relevance produces foreground logits ℓ_t. TA-MOD deterministically transforms these logits into soft event-relevance scores and measures their local variation as

$$r_t = \sigma(\ell_t), \qquad \Delta_t = |r_{t+1} - r_t| \,. \tag{6}$$

We aggregate these local variations into a pairwise modulation matrix $D \in \mathbb{R}^{T\times T}$, where D_{ij} captures the relevance variations between temporal positions i and j. Given the affinity A_{ij}, we compute the relevance-aware affinity as $\tilde{A}_{ij} = A_{ij} - \alpha D_{ij}$, with $\alpha > 0$ controlling the modulation strength.

Accordingly, TA-MOD leverages text-video relevance variations to modulate temporal affinities, mitigating the smoothing of cross-modal cues that distinguish event boundaries.

3.4 Rank GRPO

At the ranking stage, moment candidates are typically scored independently, leaving their relative ordering only implicitly determined by their scores. We therefore convert the deterministic ranking process into a stochastic policy over candidate orderings for relative comparison.

Given N moment candidates $\{c_n\}_{n=1}^{N}$ with ranking logits $\{z_n\}_{n=1}^{N}$, we formulate a Plackett–Luce ranking policy (Plackett, 1975) that sequentially samples candidate orderings from these logits. In the h-th rollout, the probability of selecting candidate n at rank k is given by

$$p_{\theta,h,k}(n) = \frac{\exp(z_n/\tau_r)}{\sum_{m\notin i_{h,<k}} \exp(z_m/\tau_r)}, \qquad i_{h,k} \sim \text{Categorical}(p_{\theta,h,k}) \,. \tag{7}$$

Repeating this process yields a group of H ranking rollouts $\{L_h\}_{h=1}^{H}$, where $L_h = (i_{h,1}, \ldots, i_{h,K})$. To probe the relative preference within each sampled ranking, we construct a competing ordering $\tilde{L}_{h,k}$ by locally rearranging the candidate sequence associated with the k-th ranking decision. We then define a rank-dependent reward from L_h and $\tilde{L}_{h,k}$. This decomposition reflects the distinct retrieval roles of the top-ranked candidate and the subsequent ranking, corresponding to localization fidelity and listwise multi-moment retrieval, respectively. Accordingly, we adopt temporal IoU for the top-ranked candidate and average precision for the tail ranking, defining the reward as

$$R_{h,k} = \begin{cases} S_{\text{tIoU}}(L_h) \cdot \max\Big(S_{\text{tIoU}}(L_h) - S_{\text{tIoU}}(\tilde{L}_{h,k}), 0\Big) \,, & k = 1, \\ S_{\text{AP}}(L_h) \cdot \max\Big(S_{\text{AP}}(L_h) - S_{\text{AP}}(\tilde{L}_{h,k}), 0\Big) \,, & k > 1. \end{cases} \tag{8}$$

The multiplicative term scales the reward by the quality of the sampled ordering, while the positive margin preserves only preferences for which the competing ordering yields a lower retrieval score. Across the H sampled rankings, the resulting rewards are used to compute rank-wise advantages through the group-relative comparison in Eq. 1. For each sampled ranking decision, we compute the policy ratio as $\rho_{h,k} = p_{\theta,h,k}(i_{h,k})/p_{\theta_{\text{old}},h,k}(i_{h,k})$. The ranking policy is then optimized with the rank-wise advantages and policy ratios using Eq. 2. The competing ordering $\tilde{L}_{h,k}$ serves only as a reward reference, while policy updates are applied to the sampled ordering L_h, enabling Rank-GRPO to directly optimize inter-candidate preference relations through ranking comparisons.

3.5 Training Strategy

M-R1 follows a two-stage training pipeline. In the first stage, we optimize the model with a supervised localization objective. The mask head is simultaneously trained with moment-level mask supervision. Starting from the first-stage model, we jointly optimize MM-GRPO and Rank-GRPO through reinforcement learning in the second stage. To accommodate the two objectives within a single model, we adopt an alternating policy optimization strategy. MM-GRPO updates the encoder-side mask policy, whereas Rank-GRPO updates the decoder-side ranking policy. The stochastic policies are used only during training, while inference remains deterministic.

Table 1: Comparison of moment retrieval performance on the QVHighlights test and validation splits. The best and second-best results are shown in **bold** and underlined, respectively. † denotes results using InternVideo2 features, and ‡ denotes methods using task-specific data augmentation. Gray row denotes methods with trainable feature adaptation rather than fixed pre-extracted features.

Method	Test					Val				
	R1		mAP			R1		mAP		
	@0.5	@0.7	@0.5	@0.75	Avg.	@0.5	@0.7	@0.5	@0.75	Avg.
TR-DETR *AAAI'24*	64.66	48.96	63.98	43.73	42.62	67.10	51.48	66.27	46.42	45.09
TaskWeave *CVPR'24*	–	–	–	–	–	64.26	50.06	65.39	46.47	45.38
UVCOM *CVPR'24*	63.55	47.47	63.37	42.67	43.18	65.10	51.81	–	–	45.79
BAM-DETR *ECCV'24*	62.71	48.64	64.57	46.33	45.36	65.10	51.61	65.41	48.56	47.61
R^2-Tuning *ECCV'24*	68.03	49.35	69.04	47.56	46.17	68.71	52.06	–	–	47.59
FlashVTG *WACV'25*	66.08	50.00	67.99	48.70	47.59	69.03	54.06	68.44	52.12	49.85
RGTR *AAAI'25*	65.50	49.22	67.12	45.77	45.53	67.68	52.90	67.38	48.00	46.95
CDTR *AAAI'25*	65.79	49.60	66.44	45.96	44.37	68.03	52.68	66.62	46.97	45.85
Keyword-DETR *AAAI'25*	66.86	51.23	67.73	46.24	45.69	–	–	–	–	–
TD-DETR *ICCV'25*	64.53	50.37	66.21	47.32	46.69	65.88	53.67	66.43	49.86	49.05
Sim-DETR *ICCV'25*	67.64	50.91	67.81	47.59	46.93	69.48	54.06	69.70	51.11	49.50
KDA *ICCV'25*	66.70	50.88	67.57	46.31	45.67	69.11	53.46	68.17	48.04	47.41
MQVTG *ICCV'25*	66.28	50.00	67.98	48.55	47.08	67.94	53.03	68.54	51.48	48.81
MS-DETR *ACM MM'25*	64.72	48.77	66.41	44.91	44.89	66.90	51.68	67.19	46.05	46.00
DualGround *NeurIPS'25*	68.20	51.72	69.23	47.71	49.02	–	–	–	–	–
CG-DETR *PR'26*	65.40	48.40	64.50	42.80	42.90	66.70	50.90	65.60	45.50	44.80
SRF *AAAI'26*	68.09	52.20	67.81	46.74	46.54	70.00	55.68	–	–	48.14
ME-DETR *ECCV'26*	66.17	51.22	68.30	52.09	50.02	65.61	52.58	66.80	52.99	51.62
CVA‡ *CVPR'26*	70.05	55.32	69.49	48.45	47.49	–	–	–	–	–
LA-DETR‡ *WACV'26*	65.37	50.71	66.65	49.22	48.21	–	–	–	–	–
DE^2TR‡ *ECCV'26*	67.98	52.46	68.36	51.24	49.87	69.81	55.29	70.52	53.43	52.12
EaTR† *ICCV'23*	–	–	–	–	–	64.90	48.00	64.77	42.97	42.46
QD-DETR† *AAAI'24*	–	–	–	–	–	68.71	53.29	68.63	47.58	46.93
TR-DETR† *AAAI'24*	–	–	–	–	–	71.94	56.06	70.23	49.77	48.96
BAM-DETR† *ECCV'24*	–	–	–	–	–	68.71	55.42	69.69	53.71	51.76
MS-DETR† *ACMMM'25*	–	–	–	–	–	69.86	54.58	68.99	49.75	50.08
FlashVTG† *WACV'25*	70.69	53.96	72.33	53.85	52.00	73.10	57.29	72.75	54.33	52.84
DualGround† *NeurIPS'25*	71.87	56.94	72.41	54.38	52.73	73.48	58.97	72.99	56.35	53.26
SDST† *ICCV'25*	70.82	56.23	71.31	54.99	53.31	73.68	**60.90**	73.52	57.42	55.60
SG-DETR† *WACV'26*	**72.20**	56.60	73.20	55.80	**54.10**	–	–	73.52	**57.91**	**55.64**
CG-DETR† *PR'26*	–	–	–	–	–	69.55	54.06	69.02	49.39	48.33
ME-DETR† *ECCV'26*	70.10	55.13	71.41	55.52	53.87	70.71	57.16	71.18	56.88	55.41
M-R1	68.35	51.69	69.12	49.72	48.62	69.48	55.84	69.70	52.17	50.11
M-R1†	72.11	**57.00**	**73.66**	**56.52**	53.68	**74.06**	59.35	**73.69**	56.53	54.58

4 Experiments

4.1 Setup

Benchmarks and metrics. To comprehensively evaluate the effectiveness of M-R1 across diverse moment retrieval settings, we conduct experiments on three widely used benchmarks: QVHighlights (Lei et al., 2021), Charades-STA (Gao et al., 2017), and TACoS (Regneri et al., 2013). QVHighlights contains free-form queries with multiple relevant moments, providing a challenging setting that better reflects practical moment retrieval scenarios. To further assess M-R1 across different video domains, we evaluate on Charades-STA and TACoS, which cover indoor daily activities and fine-grained cooking activities, respectively. For fair comparison, we follow the evaluation protocols adopted in prior VMR works (Liu et al., 2024; Cao et al., 2025). For QVHighlights, we report R1@0.5, R1@0.7, mAP@0.5, mAP@0.75, and the average mAP over IoU thresholds from 0.5 to 0.95. For Charades-STA and TACoS, we report R1@0.3, R1@0.5, R1@0.7, and mIoU.

Implementation details. We build M-R1 upon a non-LLM VMR baseline (Cao et al., 2025), adding only a lightweight 1D convolutional mask head to the output of the cross-modal encoder. Following

Table 2: Comparison of moment retrieval performance on Charades-STA and TACoS. The best and second-best results are shown in **bold** and underlined, respectively. † denotes InternVideo2 features.

Method	Charades-STA				TACoS			
	R1@0.3	R1@0.5	R1@0.7	mIoU	R1@0.3	R1@0.5	R1@0.7	mIoU
Moment-DETR	65.83	52.07	30.59	45.54	37.97	24.67	11.97	25.49
QD-DETR	–	57.31	32.55	–	–	36.80	21.10	35.8
UniVTG	70.81	58.01	35.65	50.10	51.44	34.97	17.35	33.60
UVCOM	–	59.25	36.64	–	–	36.39	23.32	–
LLMEPET	–	–	36.49	–	52.73	–	22.78	36.55
R^2-Tuning	70.91	59.78	37.02	50.86	49.71	38.72	25.12	35.92
BAM-DETR	–	60.0	39.4	52.3	56.69	41.54	26.77	39.31
TR-DETR	–	69.73†	46.33†	–	–	–	–	–
FlashVTG	–	70.32†	49.87†	–	53.71	41.76	24.74	37.61
CG-DETR	–	70.40†	48.40†	–	52.23	39.61	22.23	36.48
SG-DETR†	–	70.20	49.50	59.1	–	44.70	29.90	40.90
DualGround†	–	70.67	50.33	–	–	–	–	–
SDST†	–	**72.00**	**52.60**	**61.20**	–	44.50	**32.30**	**42.20**
M-R1†	**81.02**	71.85	51.34	60.42	**59.71**	**44.86**	31.15	41.92

the standard protocol (Moon et al., 2023b), we use pre-extracted SlowFast (Feichtenhofer et al., 2019) and CLIP (Radford et al., 2021) features, and adopt InternVideo2 (Wang et al., 2024) features as in recent state-of-the-art methods (Tang et al., 2026; Kang et al., 2026). For MM-GRPO, we set $G = 16$ and the mask sampling temperature τ_m to 1.25. For each counterfactual pair, we sample intra-event visual, pre/post-event visual, and linguistic corruptions with probabilities 0.4, 0.3, and 0.3, respectively. We generate the linguistic corruption offline using spaCy (Honnibal et al., 2020) to remove the action predicate and re-encode the resulting query with the same text encoder. For Rank-GRPO, we use $H = 16$ ranking rollouts of length $K = 10$ with a policy temperature τ_r of 0.5. We set the KL coefficient β to 0.3 for both policy objectives. The second-stage optimization is conducted for 30 epochs. As Stage 2 involves stochastic policy sampling, we report M-R1 results averaged over three independent runs. All experiments are performed on a single NVIDIA A6000 GPU.

4.2 Comparison with State-of-the-Art

QVHighlights. Table 1 compares M-R1 with recent VMR methods using SlowFast and CLIP features as well as InternVideo2 features. With InternVideo2 features, M-R1 achieves the best test results for R1@0.7, mAP@0.5, and mAP@0.75, while attaining the second-best R1@0.5. Compared with the FlashVTG baseline (Cao et al., 2025), M-R1 consistently improves all reported test and validation metrics. The gains are particularly pronounced under stricter localization criteria, with improvements of 3.04 points in R1@0.7 and 2.67 points in mAP@0.75 on the test split. With SlowFast and CLIP features, M-R1 likewise improves FlashVTG across all reported metrics on both the test and validation splits. Notably, M-R1 also surpasses SDST (Pujol-Perich et al., 2025) across all test metrics using only fixed pre-extracted features, while SDST relies on trainable feature adaptation. These results show that M-R1 delivers consistent gains across feature configurations, with notable improvements when more precise alignment with the target moments is required.

Charades-STA and TACoS. Table 2 further evaluates M-R1 on Charades-STA and TACoS, covering distinct activity domains and temporal characteristics. On Charades-STA, M-R1 remains competitive across the commonly reported localization metrics, attaining second-best performance in R1@0.5, R1@0.7, and mIoU. For TACoS, M-R1 achieves the best R1@0.5 while remaining competitive in R1@0.7 and mIoU. Together with the QVHighlights results, these findings indicate that the performance gains of M-R1 persist across diverse video domains and localization settings.

4.3 Ablation Study

We ablate the components of M-R1 in Table 3. S1 denotes the TA-MOD-based Stage-1 model, V and T denote the video and text counterfactuals in MM-GRPO, and R denotes Rank-GRPO. Starting from S1, either visual or linguistic counterfactual optimization improves retrieval, while combining both consistently outperforms either single-counterfactual setting across all reported metrics. This suggests that semantic discrimina-

Table 3: Ablation on QVHighlights (val).

Setting	S1	V	T	R	R1@0.5	R1@0.7	Avg. mAP
(a)					73.10	57.29	52.84
(b)	✓				73.29	58.35	52.61
(c)	✓	✓			73.74	58.77	53.54
(d)	✓		✓		73.61	58.52	53.60
(e)	✓	✓	✓		73.87	59.03	54.00
(f)	✓			✓	73.29	58.65	54.21
(g)	✓	✓	✓	✓	**74.06**	**59.35**	**54.58**

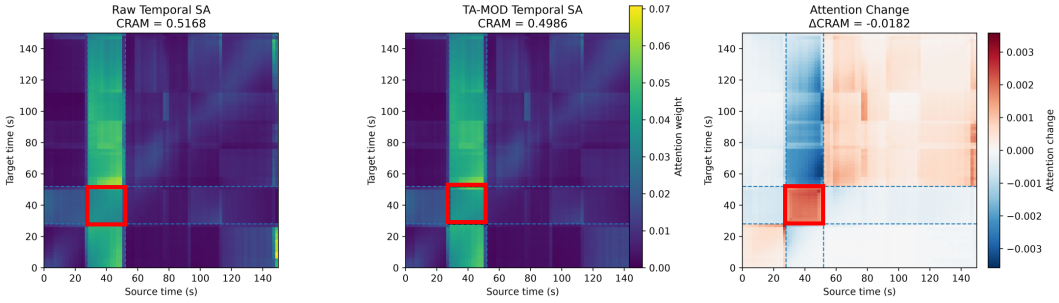


Figure 4: Attention maps show that TA-MOD reduces cross-region attention between the target moment and surrounding temporal context, thereby limiting temporal mixing across regions.

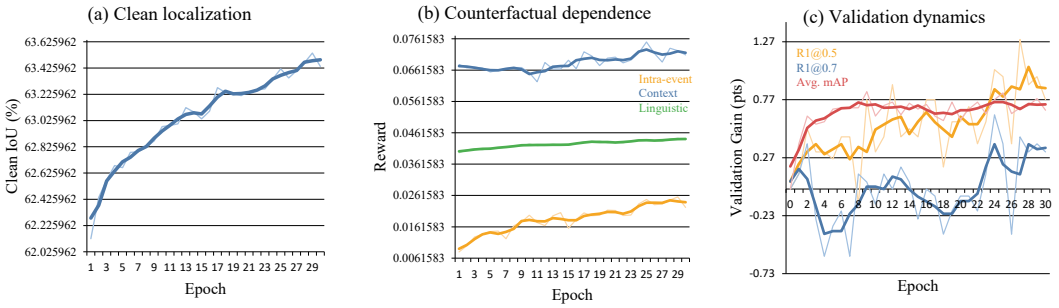


Figure 5: Training dynamics of MM-GRPO show preserved clean localization, increasing counterfactual dependence across visual and linguistic perturbations, and validation gains from Stage 2.

tion benefits from jointly capturing visual event structure and linguistic action semantics rather than relying on evidence from a single modality. Compared with the baseline, S1 primarily improves performance under the stricter IoU criterion while leaving the overall mAP nearly unchanged, a pattern consistent with TA-MOD's role in preserving fine-grained temporal distinctions. Rank-GRPO provides further gains, particularly in Avg. mAP, both when applied to S1 alone and when combined with MM-GRPO, supporting its complementary role in refining the ordering of competing moments. Together, these components yield the strongest performance across all reported metrics, supporting the complementary optimization of semantic, temporal, and preference discrimination.

4.4 Analysis

Does MM-GRPO promote semantic discrimination? We measure retrieval sensitivity to semantic changes using Counterfactual Response Rate (CRR), where CRR@τ denotes the fraction of top-k candidates unmatched at a temporal IoU threshold τ. M-R1 achieves the highest mean CRR in Fig. 2, while action removal substantially alters its prediction but leaves several baselines nearly unchanged. Training curves in Fig. 5 further show increasing counterfactual dependence, consistent with the gains from jointly optimizing visual and linguistic counterfactuals in Table 3. These results suggest that MM-GRPO encourages semantic discrimination through counterfactual sensitivity.

Does TA-MOD promote temporal discrimination? We quantify cross-region temporal mixing with Cross-Region Attention Mass (CRAM), which measures attention between temporal positions inside and outside the ground-truth moment. TA-MOD reduces CRAM for 95.9% of validation queries, while Fig. 4 shows reduced cross-region attention between the target moment and its surroundings. These results support TA-MOD's role in preserving fine-grained temporal distinctions.

5 Conclusion

In this work, we introduced **M-R1**, which rethinks moment retrieval through semantic, temporal, and relational discrimination. MM-GRPO encourages sensitivity to multimodal semantic evidence, TA-MOD preserves fine-grained temporal distinctions, and Rank-GRPO refines the ordering of competing moments. Experiments on QVHighlights, Charades-STA, and TACoS demonstrate consistent improvements across diverse retrieval settings. Notably, M-R1 achieves strong performance with only a lightweight mask head and training-time optimization, without introducing additional inference-time modules. These results establish multi-dimensional discrimination as a promising direction for accurate and efficient moment retrieval. We hope this work also encourages further exploration of reinforcement learning as an optimization paradigm for moment retrieval.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for language editing and proofreading. We have not used generative AI tools for research idea formulation, conceptualization, baseline model implementations, or experimental evaluations, and all other tasks requiring disclosure are not applicable to this work. We have reviewed all AI-assisted work. We have thoroughly reviewed all AI-assisted text revisions to ensure technical accuracy and maintain our intent. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

(This section is **recommended** and does not count toward the page limit.)

We are not aware of any specific ethical concerns associated with this work.

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.)

We provide the methodological formulation, training strategy, implementation details, and hyperparameters required to reproduce M-R1 throughout the main paper. We have aimed to describe the experimental settings and evaluation protocols with sufficient detail for reproducibility.

REFERENCES

Joungbin An and Kristen Grauman. Hieramamba: Video temporal grounding via hierarchical anchor-mamba pooling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16954–16965, 2026.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, volume 2024, pp. 4965–4987, 2024.

Zhuo Cao, Bingqing Zhang, Heming Du, Xin Yu, Xue Li, and Sen Wang. Flashvtg: Feature layering and adaptive score handling network for video temporal grounding. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 9226–9236. IEEE, 2025.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Andong Deng, Zhongpai Gao, Anwesa Choudhuri, Benjamin Planche, Meng Zheng, Bin Wang, Terrence Chen, Chen Chen, and Ziyan Wu. Seq2time: Sequential knowledge transfer for video llm temporal grounding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13766–13775. IEEE, 2025.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *2019 IEEE/CVF international conference on computer vision (ICCV)*, pp. 6201–6210. IEEE, 2019.

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. *Advances in Neural Information Processing Systems*, 38:99114–99137, 2026.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5277–5285. IEEE, 2017.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025a.

Yongxin Guo, Jingyu Liu, Mingda Li, Qingbin Liu, Xi Chen, and Xiaoying Tang. Trace: Temporal grounding video llm via causal event modeling. *arXiv preprint arXiv:2410.05643*, 2024.

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. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 3302–3310, 2025b.

Tanveer Hannan, Md Mohaiminul Islam, Jindong Gu, Thomas Seidl, and Gedas Bertasius. Revisionllm: Recursive vision-language model for temporal grounding in hour-long videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19012–19022. IEEE, 2025.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, et al. spacy: Industrial-strength natural language processing in python. 2020.

Chengyang Hu, Guanshuo Wang, Fufu Yu, Qiong Jia, Shouhong Ding, and Lizhuang Ma. Sedetr: Explicit semantic exploration for generalizability and distinguishability in video temporal grounding. In *European Conference on Computer Vision*, pp. 152–170. Springer, 2026.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14271–14280. IEEE, 2024.

Wenxuan Huang, Bohan Jia, Shaosheng Cao, Zheyu Ye, Zhe Xu, Yao Hu, Shaohui Lin, et al. Vision-r1: Incentivizing reasoning capability in multimodal large language models. In *International Conference on Learning Representations*, volume 2026, pp. 63794–63812, 2026.

Mingyu Jeon, Sunjae Yoon, Jonghee Kim, and Junyeong Kim. Granalign: Granularity-aware alignment framework for zero-shot video moment retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5287–5295, 2026.

Minseok Kang, Minhyeok Lee, Minjung Kim, Donghyeong Kim, and Sangyoun Lee. Empower words: Dualground for structured phrase and sentence-level temporal grounding. *Advances in Neural Information Processing Systems*, 38:92472–92499, 2026.

Pilhyeon Lee and Hyeran Byun. Bam-detr: Boundary-aligned moment detection transformer for temporal sentence grounding in videos. In *European Conference on Computer Vision*, pp. 220–238. Springer, 2024.

Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34:11846–11858, 2021.

Pandeng Li, Chen-Wei Xie, Hongtao Xie, Liming Zhao, Lei Zhang, Yun Zheng, Deli Zhao, and Yongdong Zhang. Momentdiff: Generative video moment retrieval from random to real. *Advances in neural information processing systems*, 36:65948–65966, 2023.

Zongyao Li, Yongkang Wong, Satoshi Yamazaki, Jianquan Liu, and Mohan Kankanhalli. Object-centric framework for video moment retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6771–6779, 2026.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *Advances in neural information processing systems*, 38:40783–40818, 2026.

Ye Liu, Jixuan He, Wanhua Li, Junsik Kim, Donglai Wei, Hanspeter Pfister, and Chang Wen Chen. R 2-tuning: Efficient image-to-video transfer learning for video temporal grounding. In *European conference on computer vision*, pp. 421–438. Springer, 2024.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2034–2044. IEEE, 2025.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, et al. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

WonJun Moon, Sangeek Hyun, SuBeen Lee, and Jae-Pil Heo. Correlation-guided query-dependency calibration for video temporal grounding. *arXiv preprint arXiv:2311.08835*, 2023a.

WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23023–23033. IEEE, 2023b.

Seojeong Park, Jiho Choi, Kyungjune Baek, and Hyunjung Shim. Momentmix augmentation with length-aware detr for temporally robust moment retrieval. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1127–1136. IEEE, 2026.

Robin L Plackett. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2):193–202, 1975.

David Pujol-Perich, Sergio Escalera, and Albert Clapés. Sparse-dense side-tuner for efficient video temporal grounding. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21515–21524. IEEE, 2025.

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. *arXiv preprint arXiv:2402.11435*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Ran Ran, Jiwei Wei, Shiyuan He, Zeyu Ma, Chaoning Zhang, Ning Xie, and Yang Yang. Kda: Knowledge diffusion alignment with enhanced context for video temporal grounding. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23311–23320. IEEE, 2025.

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013.

Allen Ren, Justin Lidard, Lars Ankile, Anthony Simeonov, Pulkit Agrawal, Anirudha Majumdar, Benjamin Burchfiel, Hongkai Dai, and Max Simchowitz. Diffusion policy policy optimization. In *International Conference on Learning Representations*, volume 2025, pp. 77288–77329, 2025.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14313–14323. IEEE, 2024.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. Pmlr, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Hao Sun, Mingyao Zhou, Wenjing Chen, and Wei Xie. Tr-detr: Task-reciprocal transformer for joint moment retrieval and highlight detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4998–5007, 2024.

Xiaolong Sun, Liushuai Shi, Le Wang, Sanping Zhou, Kun Xia, Yabing Wang, and Gang Hua. Diversifying query: Region-guided transformer for temporal sentence grounding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 7131–7139, 2025a.

Xiaolong Sun, Le Wang, Sanping Zhou, Liushuai Shi, Kun Xia, Mengnan Liu, Yabing Wang, and Gang Hua. Moment quantization for video temporal grounding. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–10. IEEE, 2025b.

Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Xiansheng Chen, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning of vision language models. *Advances in neural information processing systems*, 38:5772–5822, 2026.

Jiajin Tang, Zhengxuan Wei, Yuchen Zhu, Cheng Shi, Guanbin Li, Liang Lin, and Sibei Yang. Simdetr: Unlock detr for temporal sentence grounding. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22760–22771. IEEE, 2025.

Xu Tang, Yao Hu, and Yongchao Xul. Decoupling moment from event for video temporal grounding. In *Computer Vision–ECCV 2026: 19th European Conference, Malmö, Sweden, September 8–12, 2026, Proceedings, Part IX*, pp. 156. Springer Nature, 2026.

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. *Advances in Neural Information Processing Systems*, 38:83330–83364, 2026.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European conference on computer vision*, pp. 396–416. Springer, 2024.

Zhengxuan Wei, Jiajin Tang, and Sibei Yang. Augmenting moment retrieval: Zero-dependency two-stage learning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3401–3412. IEEE, 2025.

Yongliang Wu, Xinting Hu, Yuyang Sun, Yizhou Zhou, Wenbo Zhu, Fengyun Rao, Bernt Schiele, and Xu Yang. Number it: Temporal grounding videos like flipping manga. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13754–13765. IEEE, 2025.

Yicheng Xiao, Zhuoyan Luo, Yong Liu, Yue Ma, Hengwei Bian, Yatai Ji, Yujiu Yang, and Xiu Li. Bridging the gap: A unified video comprehension framework for moment retrieval and highlight detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18709–18719. IEEE, 2024.

Jin Yang, Ping Wei, Huan Li, and Ziyang Ren. Task-driven exploration: Decoupling and inter-task feedback for joint moment retrieval and highlight detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18308–18318. IEEE, 2024.

Zuhao Yang, Yingchen Yu, Yunqing Zhao, Shijian Lu, and Song Bai. Timeexpert: An expert-guided video llm for video temporal grounding. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24286–24296. IEEE, 2025.

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. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

Shihao Yuan, Yahui Liu, Yang Yue, Jingyuan Zhang, Wangmeng Zuo, Qi Wang, Fuzheng Zhang, and Guorui Zhou. Ar-grpo: Training autoregressive image generation models via reinforcement learning. *arXiv preprint arXiv:2508.06924*, 2025.

Wenzheng Zeng, Difei Gao, Mike Zheng Shou, and Hwee Tou Ng. Factorized learning for temporally grounded video-language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20683–20693. IEEE, 2025.

Jinglei Zhang, Yuanfan Guo, Rolandos Alexandros Potamias, Jiankang Deng, Hang Xu, and Chao Ma. Vtimecot: Thinking by drawing for video temporal grounding and reasoning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24203–24213. IEEE, 2025.

Yifan Zhang, Chengxu Liu, Yujie Dun, and Xueming Qian. De 2 tr: Dual evidence detection transformer for video temporal grounding. In *European Conference on Computer Vision*, pp. 463–482. Springer, 2026a.

Zhengbo Zhang, Mark He Huang, Zhigang Tu, and Ming-Hsuan Yang. Dart: Difficulty-adaptive routing for zero-shot video temporal grounding. In *European Conference on Computer Vision*, pp. 73–91. Springer, 2026b.

Minghang Zheng, Zihao Yin, Yi Yang, Yuxin Peng, and Yang Liu. Omnivtg: A large-scale dataset and training paradigm for open-world video temporal grounding. *arXiv preprint arXiv:2604.25276*, 2026a.

Minghang Zheng, Zihao Yin, Yi Yang, Yuxin Peng, and Yang Liu. Temporal-aware reasoning optimization for video temporal grounding. *arXiv preprint arXiv:2606.09248*, 2026b.

Xinyang Zhou, Fanyue Wei, Lixin Duan, Angela Yao, and Wen Li. The devil is in the spurious correlations: boosting moment retrieval with dynamic learning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20981–20990. IEEE, 2025.

A Additional Analysis

A.1 Limitations

M-R1 operates on pre-extracted video representations and therefore depends on the semantic and temporal information encoded in the underlying features. In particular, the performance gains with SlowFast and CLIP features are smaller than those obtained with InternVideo2 features. This suggests that M-R1 may reach its performance limit earlier when the input representations lack sufficiently discriminative cues. Since our method primarily improves discrimination over the available representations rather than enriching the visual features themselves, improving robustness to weaker feature representations remains an important direction for future work.

A.2 Short-Moment Retrieval

We further examine retrieval performance on short moments, where small temporal deviations can substantially affect localization accuracy. Table 4 reports length-specific performance on QVHighlights using Avg. R1 and Avg. mAP over IoU thresholds from 0.5 to 0.95. M-R1 achieves 12.70 Avg. R1 and 18.82 Avg. mAP, outperforming both FlashVTG and LA-DETR. These results show that M-R1 remains effective for temporally compact moments that require more precise localization.

Table 4: Short-moment retrieval performance on QVHighlights.

Method	Avg. R1	Avg. mAP
FlashVTG	9.23	15.65
LA-DETR	11.58	16.35
M-R1	**12.70**	**18.82**

The gains on short moments are consistent with the discriminative design of M-R1. Short temporal intervals are particularly sensitive to feature mixing with their surrounding context, since even small temporal deviations can substantially affect localization accuracy. TA-MOD mitigates such cross-region mixing and preserves fine-grained temporal distinctions, while MM-GRPO and Rank-GRPO provide complementary semantic and preference discrimination among closely competing moments. This combination is particularly beneficial for temporally compact events.