ArrowGEV: Grounding Events in Video via Learning the Arrow of Time

Anonymous ACL submission

Abstract

Grounding events in videos serves as a fundamental capability in video analysis. While Vision-Language Models (VLMs) are increasingly employed for this task, existing approaches predominantly train models to associate events with timestamps in the forward video only. This paradigm hinders VLMs from capturing the inherent temporal structure and directionality of events, thereby limiting robustness and generalization. To address this limitation, inspired by the *arrow of time* in physics, which characterizes the intrinsic directionality of temporal processes, we propose ArrowGEV, a reinforcement learning framework that explicitly models temporal directionality in events to improve both event grounding and temporal directionality understanding in VLMs. Specifically, we categorize events into time-sensitive (e.g., putting down a bag) and time-insensitive (e.g., holding a towel in the left hand). The former denote events whose reversal substantially alters their meaning, while the latter remain semantically unchanged under reversal. For time-sensitive events, ArrowGEV introduces a reward that encourages VLMs to discriminate between forward and backward videos, whereas for time-insensitive events, it enforces consistent grounding across both directions. Extensive experiments demonstrate that ArrowGEV not only improves grounding precision and temporal directionality recognition, but also enhances general video understanding and reasoning ability.

1 Introduction

Grounding events in videos (GEV) is the task of localizing a specific timestamp in untrimmed videos described by natural language events (Anne Hendricks et al., 2017; Lin et al., 2023). As a fundamental ability in video analysis, GEV is crucial for fine-grained video analysis (Luo et al., 2025; Arefeen et al., 2024), video content retrieval (Aslandogan and Yu, 2002; Gabeur et al., 2020), dense video caption (Iashin and Rahtu, 2020; Seo et al., 2022), and video generation (Tan et al., 2024; Yang et al., 2025a; Menapace et al., 2024).

To tackle this challenge, early approaches relied on handcrafted architectures and video-query feature-matching strategies (Hou et al., 2022; Pan et al., 2023), but they suffer from constraints of predefined video snippets, suboptimal video-text features, and poor cross-task generalization. Recent work has shifted to end-to-end Vision Language Models (VLMs) (Wang et al., 2025a; Team et al., 2025; Xiaomi, 2025), which directly process videos and queries while retaining generalizability either via large-scale timestamp annotation training (Huang et al., 2024a; Li et al., 2024d; Zeng et al., 2025), integration of textual timestamp tokens/embeddings (Guo et al., 2025), or adaptation to event grounding via video segmentation (Guo et al., 2024b; Wang et al., 2024). Despite this progress, existing methods only align events with forward videos, failing to capture the intrinsic temporal structure and directionality of events.

To investigate this limitation, we present a pilot study (Section 3.1) and case analysis (Figure 1). As shown in Figure 1, VLMs often struggle to recognize that reversing a video can fundamentally change the semantics of the event, mistakenly associating reversed events with their forward counterparts. In contrast, for events unaffected by temporal reversal, models struggle to consistently locate timestamps in both directions.

To address this, we turn to the *Arrow of Time* (Eddington, 2019; Layzer, 1975), a foundational concept in physics that characterizes the intrinsic directionality of temporal processes. This perspective highlights that the semantics of real-world events are inherently tied to their temporal progression: reversing time may either yield a semantically distinct event or preserve the core meaning, depending on the nature of the process. Generally, the impact of this temporal directionality on event un-

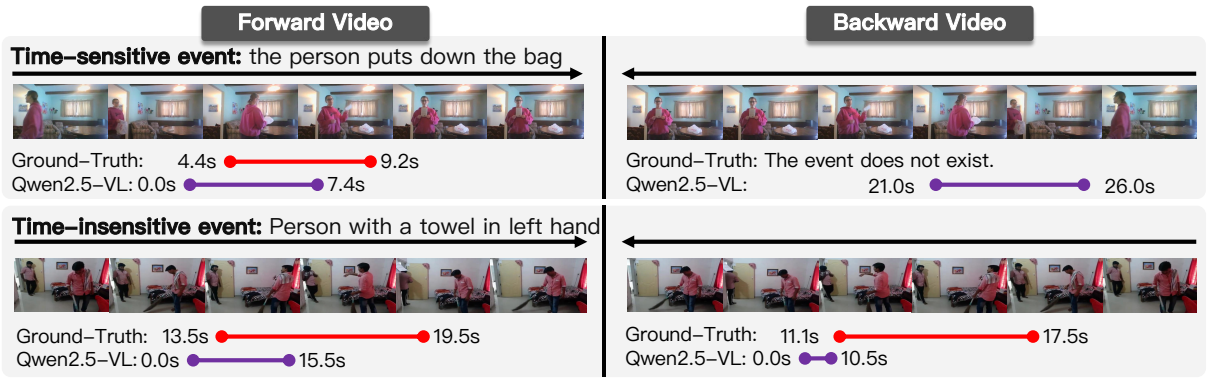

Figure 1: Two examples of Qwen2.5-VL-7B predicting event timestamps in forward and backward videos. In the top row, reversing the video changes the event semantics, while in the bottom row the event remains invariant. The model partially localizes events in forward videos but fails to recognize event absence in the first reversed case and cannot robustly localize the event in the second reversed video.

derstanding manifests in two distinct event types: time-sensitive events, where reversal fundamentally alters meaning (e.g., "a man picks up a glass" becomes "a man puts down a glass"), and time-insensitive events, whose meaning remains invariant under reversal (e.g., "a ball is on the table" retains the same semantic essence when time is reversed).

Inspired by *the Arrow of Time*, we propose ARROWGEV, a reinforcement learning (RL) framework, aiming to improve event grounding and the understanding of temporal directionality via learning the arrow of time. Unlike specialized architectures (Chen et al., 2021; Woo et al., 2024), our approach leverages RL to optimize temporal precision through a tailored reward signal directly, mitigating the tendency of VLMs to overfit textual timestamps rather than video semantics. At its core, ARROWGEV enables the VLM to learn temporal structures by discriminating between time-sensitive and time-insensitive events. We introduce a reward function that encourages distinct localizations for time-sensitive events and their reversed counterparts while enforcing consistent grounding for time-insensitive ones. To further enhance training efficiency, we propose a difficulty-aware strategy that dynamically emphasizes challenging samples through weighted adjustments and a curriculum-based filtering of well-solved examples.

We conduct extensive experiments on three GEV benchmark datasets, results indicate that VLMs trained with ARROWGEV significantly improve the event grounding performance. In addition, ARROWGEV Substantially improved the VLM's ability to understand temporal structures in event grounding. Finally, ARROWGEV also improves the out-of-distribution (OOD) performance on general video understanding and reasoning tasks.

2 Related Work

Grounding Events in Videos with VLMs. This task localizes specific events within untrimmed videos (Nan et al., 2021; Wang et al., 2018; Li et al., 2020; Zhao et al., 2017; Kulkarni and Fazli, 2025; Chen et al., 2025b; Yang et al., 2025b; Tian et al., 2025; Hannan et al., 2024; Mu et al., 2024a). While traditional methods rely on task-specific heads, recent Vision-Language Models (VLMs) leverage the reasoning capabilities of LLMs for unified temporal understanding (Huang et al., 2024a; Li et al., 2024d). To bridge the modality gap, existing research typically focuses on large-scale supervised fine-tuning with timestamp-based data (Zeng et al., 2025), introducing specialized temporal tokens (Hong et al., 2024), or utilizing video segmentation to align structural granularity (Huang et al., 2024b; Wang et al., 2024). Despite these advances, current VLM-based approaches largely overlook the intrinsic temporal directionality of events. In contrast, ARROWGEV introduces a principled framework that explicitly models temporal direction, facilitating more robust and physically consistent grounding.

Post-Training for VLMs. Post-training techniques are essential for adapting pre-trained VLMs to complex downstream tasks. While large-scale instruction tuning has significantly boosted performance in models like LLaVA-OV and MAmmoTH-VL (Li et al., 2024a; Guo et al., 2024a), recent research has pivoted toward RL to refine multimodal reasoning (Su et al., 2025; Meng et al., 2025). This

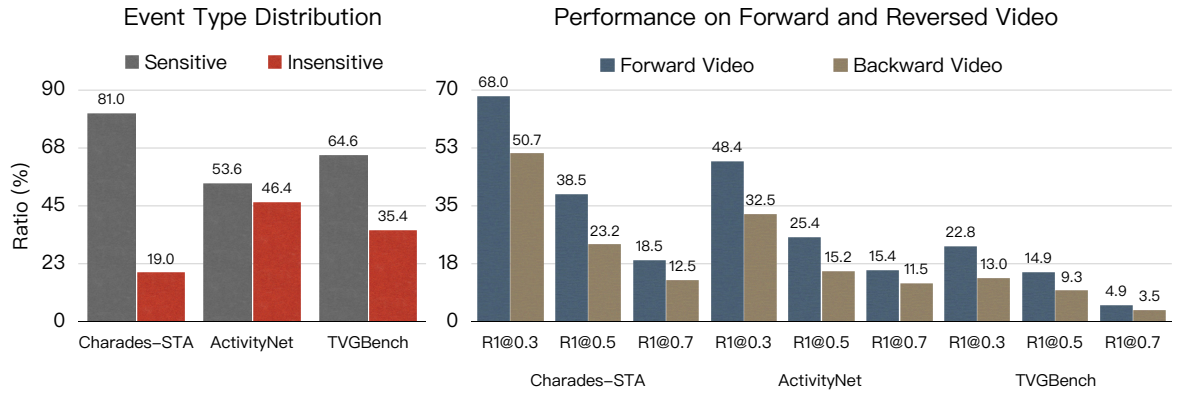

Figure 2: Quatitative Analysis of Qwen2.5-VL-7B on GEV Benchmarks. **Left:** statistics of the portion of time-sensitive and time-insensitive events across three benchmarks. **Right:** R1@m metrics on the time-sensitive subsets of three benchmarks.

shift is particularly evident in video understanding, where RL is increasingly employed to enhance event grounding and long-form reasoning (Feng et al., 2025; Wang et al., 2025b). However, existing RL approaches for videos often overlook the inherent temporal structure of events. Our work addresses this gap by explicitly instilling temporal directionality, guiding VLMs toward a more robust and physically grounded understanding of videos.

Time in Video. Temporal directionality is a foundational self-supervised signal for video representation learning, typically utilized through shuffle-and-learn or order-prediction tasks (Misra et al., 2016; Wei et al., 2018; Dwibedi et al., 2019). VLM frameworks often treat reversed videos merely as negative samples for contrastive alignment or as binary classification targets (Xu et al., 2021; Price and Damen, 2019). However, recent evidence suggests that VLMs remain alarmingly insensitive to temporal directionality in complex reasoning tasks (Xue et al., 2025; Du et al., 2024). Unlike prior works that focus on events where reversal fundamentally alters semantics, we address the nuances of temporally invariant events and move beyond coarse classification. Specifically, we investigate temporal directionality within the high-precision requirements of event grounding, demanding that models not only detect reversal but accurately localize events across the temporal axis.

3 Method

3.1 Pilot Study

To assess the ability of current Vision-Language Models (VLMs) to perceive temporal structure, we conduct a pilot study using the Qwen2.5-7B-VL-Instruct model. Our investigation focuses on time-sensitive events whose semantic meaning is fundamentally altered upon time reversal (e.g., "opening a door" becomes "closing a door"). As illustrated in Figure 2, such events constitute a significant portion of common benchmarks, particularly Charades-STA (Sigurdsson et al., 2016).

We evaluate the VLM by prompting it to localize time-sensitive events in both forward and backward videos. We then measure localization accuracy using Intersection over Union (IoU). For the forward video, we use the original ground truth, while for the backward video, we use a corresponding pseudo-ground truth created by reversing the original event's timestamps. Ideally, a model with a robust grasp of temporal structure and directionality should recognize that the described event no longer exists in the reversed sequence, resulting in an Intersection-over-Union (IoU) score close to zero. However, as shown in the right panel of Figure 2, the VLM localizes these time-sensitive events even in backward videos, producing a high IoU. This reveals a critical failure: the model fails to capture temporal structure, struggling to distinguish the semantic change of events in forward and backward video.

3.2 Background of GRPO: RL for LLM

As a pioneer among open-sourced R1-style LLMs, Deepseek-R1 (dee, 2025) leverages Group Relative Policy Optimization (GRPO) to train the policy model π_θ (i.e., the LLM) to think before answering, making it particularly well-suited for tasks with well-defined answers, such as mathematical reasoning. In the GRPO framework, given an input question p, the LLM samples G candidate responses $o = \{o_1, \ldots, o_G\}$, and a reward function $r(\cdot)$ assigns a reward score to each response, yielding

$\{r(o_1), \ldots, r(o_G)\}$. GRPO encourages the LLM to generate responses that maximize a weighted sum reward $R(o)$, defined by:

$$R(o) = \sum_{i=1}^{G} \frac{\pi_\theta(o_i)}{\pi_{\theta_{\text{old}}}(o_i)} \cdot \underbrace{\frac{r(o_i) - \text{mean}(\{r(o_j)\}_{j=1}^{G})}{\text{std}(\{r(o_j)\}_{j=1}^{G})}}_{\text{Advantage } A_i} \tag{1}$$

where $\pi_\theta(o)$ denotes the probability of LLM generating the response o, and $\pi_{\theta_{\text{old}}}$ represents the LLM parameters from a recently optimized state. And the latter term is the Advantage A_i of i-th candidate. To ensure training stability and avoid large deviations from the original language model behavior, the final training objective incorporates a KL-divergence regularization term (dee, 2025), penalizing divergence between π_θ and π_{old}:

$$\max_{\pi_\theta} \mathbb{E}_{o \sim \pi_{\theta_{\text{old}}}(p)}[R(o) - \beta \mathrm{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})] \tag{2}$$

where β is a scaling coefficient. We omit the clipping operation for simplicity.

3.3 ArrowGEV

Formulation. The task of Grounding events in videos requires a model $\mathcal{M}$ to map an untrimmed video V and a natural language event Π to a specific temporal timestamp $\mathcal{T}^{\text{fwd}} = \mathcal{M}(\mathcal{V}, \Pi)$ that accurately aligns with a ground-truth annotation $\mathcal{T}^{\text{gt}}$. As discussed in Section 3.1, the VLM struggles to capture the temporal structure of events in the video. To overcome this limitation, we leverage the backward video, denoted as $\mathcal{V}'$, to enhance the understanding of temporal structure. Our key insight is that the effect of temporal reversal is not uniform across all events; it depends on the event's intrinsic properties. Therefore, to construct a meaningful learning signal from the reversed video V', we first categorize events according to their temporal nature.

Event Categorization. We categorize events into two types:

1) *Time-sensitive events*: Events whose semantics are transformed into a distinct, often opposite, action when time is reversed (e.g., "opening a door" becomes "closing a door"). We expect a model to recognize the changed semantics and reduce the prediction on the same video segment.

2) *Time-insensitive events*: Events whose semantics are preserved under temporal reversal (e.g. "a car is parked"). We expect a model's prediction to be consistent with respect to time reversal.

Then, we design an RL framework to encourage the VLM to not only accurately localize the timestamps of the events in the forward video, but also to distinguish the existence of time-sensitive events and recognize time-insensitive events in the backward video. After training, ARROWGEV enables the VLM achieve above and learn the temporal structure more robustly to enhance the localization.

3.4 Temporal Directionality Reward Modeling

A reward function for GEV should ideally incentivize a model to localize the events while understanding an event's temporal structure.

To determine the type of each event, we use an LLM to perform reasoning to classify each event q as time-sensitive or time-insensitive. The model first generates the reasoning process $r \sim P_{\text{LLM}}(\cdot|p, q)$, and then determines the type of event $c(q) \sim P_{\text{LLM}}(\cdot|r, p, q)$. The analysis of event categorization verifies the reliability in the supplemental material Section 1, and see the prompt in the supplemental material Section 3.

Let $c(q) \in \{\text{insensitive}, \text{sensitive}\}$ denote the event's category. We formulate the grounding reward r as a linear combination of two components: a localization accuracy reward r_{acc}, which promotes precise event localization in the forward video, and a temporal directionality reward r_{temp}, which enforces temporal understanding. These are balanced by a weighting factor λ:

$$r_{\text{grounding}} = r_{\text{acc}} + \lambda r_{\text{temp}} \tag{3}$$

The accuracy reward r_{acc} is defined as timestamp-aware IoU (Wang et al., 2025c) with the ground-truth $\mathcal{T}^{\text{gt}}$, which encourages the alignment between start and end time with the ground-truth based on the standard IoU:

$$\text{tIoU}(\mathcal{T}^{\text{fwd}}, \mathcal{T}^{\text{gt}}) = \text{IoU} \cdot (1 - \frac{|t_s - t'_s|}{t}) \cdot (1 - \frac{|t_e - t'_e|}{t}) \tag{4}$$

$$r_{\text{acc}} \triangleq \text{tIoU}(\mathcal{T}^{\text{fwd}}, \mathcal{T}^{\text{gt}}) \tag{5}$$

To define the temporal directionality reward, let $\mathcal{T}^{\text{rev}} = \mathcal{M}(V', q)$ be the VLM's predicted timestamp on the backward video. We define a temporal reversal operator $\mathcal{R}(\mathcal{T}) \triangleq [d - t_e, d - t_s]$ for a timestamp $\mathcal{T} = [t_s, t_e]$ in a video of duration d. This allows us to compute a **directionality score**, S_c, measuring the alignment between the reversed prediction and the mirrored forward prediction:

$$S_c \triangleq \text{tIoU}(\mathcal{T}^{\text{rev}}, \mathcal{R}(\mathcal{T}^{\text{fwd}})) \tag{6}$$

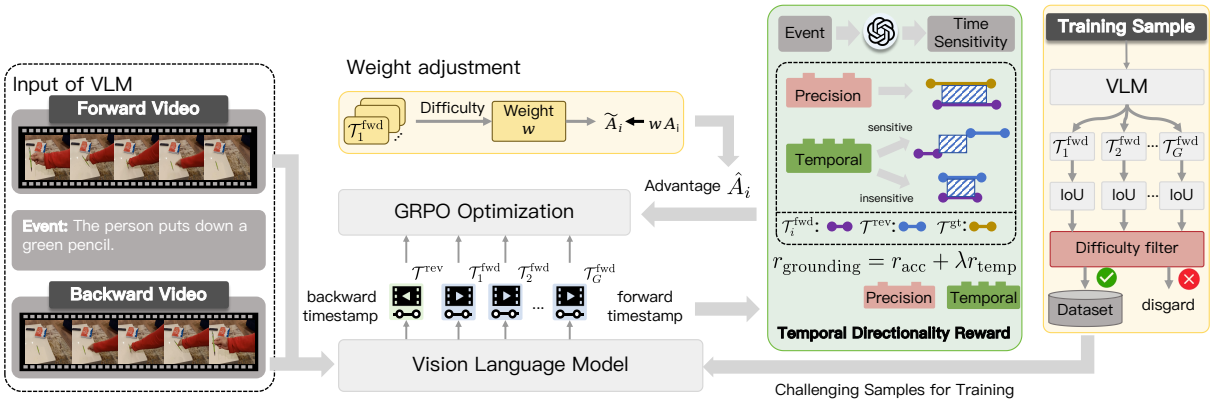

Figure 3: Overview of ARROWGEV. First, we input the event and both forward and backward videos into the VLM to obtain the predicted timestamps in both directions. Then we calculate the reward based on the category of the event. Having the reward for samples, we use GRPO with difficulty-aware training strategies to optimize the VLM for improved localization accuracy and directionality understanding.

The temporal directionality reward r_{temp} leverages this score to enforce the expected temporal properties of events. For time-insensitive events, a high reward ($r_{\text{temp}} = S_c$) reflects strong overlap between the predicted timestamps in the forward and backward videos, demonstrating consistency in the model's localization across both directions. In contrast, for time-sensitive events, a low reward is desirable, as it indicates that the model successfully distinguishes the semantic shifts introduced by temporal reversal. Specifically, the reward function ($r_{\text{temp}} = 1 - S_c$) penalizes the incorrect localization of time-sensitive events in backward videos, encouraging the generation to indicate the event's nonexistence, rather than an incorrect timestamp. Together, these cases motivate our final unified reward function:

$$r_{\text{grounding}} = r_{\text{acc}} + \lambda \cdot \begin{cases} S_c & \text{if } c(q) = \text{insensitive} \\ 1 - S_c & \text{if } c(q) = \text{sensitive} \end{cases} \tag{7}$$

By optimizing Equation 7, our model is explicitly incentivized not only to be accurate but also to develop an internal representation that is consistent with the fundamental temporal properties of events.

Reasoning Template Reward. To facilitate complex temporal reasoning, we adopt a *think-before-act* paradigm, requiring the model to generate intermediate rationales before predicting timestamps. We enforce this via a binary format reward $r_{\text{form}}(o) \in \{0, 1\}$, which validates if the output o strictly adheres to the template: `<think> ... </think> <answer> <`t_s `to` t_e`> </answer>`. This structural component is then integrated into the final objective, which combines our format, accuracy, and temporal directionality rewards:

$$r_{\text{final}} = r_{\text{grounding}} + r_{\text{form}}(o) \tag{8}$$

3.5 Training

To train the model, we propose two strategies to address the difficulty-bias issue and improve training efficiency.

Difficulty-Aware Weight Adjustment. During the training, samples progressively become easier for the model. This will lead to a difficulty-bias issue. To alleviate this, we propose a weight adjustment that enables the model to focus more on the hard samples. Specifically, we introduce a difficulty coefficient $w_i \propto -\text{tIoU}(\mathcal{T}^{\text{fwd}}, \mathcal{T}^{\text{gt}})$ for i-th sample to quantify the difficulty level. In this case, we sample G predictions by the VLM and calculate the difficulty weight:

$$w_i = \exp\left(\frac{1 - \frac{1}{G}\sum_{j=1}^{G} \text{tIoU}(\mathcal{T}_j^{\text{fwd}}, \mathcal{T}^{\text{gt}})}{\tau}\right), \tag{9}$$

This coefficient dynamically adjusts sample weights by computing the average tIoU of different responses for i-th sample: $\hat{A}_i = w_i A_i$.

Dynamic Curriculum via Difficulty Filtering. The efficacy of the above detailed policy optimization is highly dependent on the quality and challenge of the training data distribution, $\mathcal{D}$. As the policy VLM π_θ improves, a static data set can be dominated by samples that no longer provide a sufficient learning signal. To counteract this and maintain a challenging training environment, we implement a dynamic curriculum strategy that adaptively refines the data distribution throughout the training. Specifically, we initialize our training set, $\mathcal{D}_0$,

with the dataset from (Wang et al., 2025c), which is pre-filtered to emphasize samples of moderate difficulty. To ensure the model is persistently challenged as it learns, we introduce a dynamic difficulty filter at the conclusion of each training epoch e. To construct the dataset for the next epoch, $\mathcal{D}_{e+1}$, we evaluate each sample $(\mathcal{V}, \text{II}) \in \mathcal{D}_e$ against the current policy π_{θ_e}. A sample is deemed "mastered" and is subsequently removed if the policy consistently solves it with high accuracy. Formally, for each $(\mathcal{V}, \text{II})$, we generate a group of G rollout outputs $\{\mathcal{T}_i^{fwd}\}_{i=1}^{G}$. The sample is filtered out if its worst-case performance in the group exceeds a high-performance threshold η:

$$\mathcal{D}_{e+1} = \mathcal{D}_e \setminus \left\{ (\mathcal{V}, \text{II}) \in \mathcal{D}_e \mid \min_{i=1\ldots G} \text{IoU}(\mathcal{T}_i^{fwd}, \mathcal{T}^{\text{gt}}) > \eta \right\} \tag{10}$$

We set $\eta = 0.7$ in our experiments, which is the most strict metric for evaluating the quality of prediction in most of the previous work (Ye et al., 2025; Li et al., 2025; Nguyen et al., 2025). This adaptive curriculum ensures that the model continually focuses its capacity on unsolved or challenging problems, thereby maintaining a strong gradient signal and promoting the development of a more robust and generalizable policy.

4 Experiments

4.1 Experimental Settings.

Benchmarks. We evaluate our model on three GEV benchmarks: Charades-STA (Sigurdsson et al., 2016), ActivityNet (Caba Heilbron et al., 2015), and TVGBench (Wang et al., 2025c). To further evaluate the generalization ability, we further compare ARROWGEV on the video understanding and reasoning benchmarks, including TempCompass (Liu et al., 2024), MVBench (Li et al., 2024b), VSI-Bench (Yang et al., 2025c), VideoMMMU (Hu et al., 2025), MMVU (Zhao et al., 2025), and VideoMME (Fu et al., 2024).

Implementation Details. Our methodology is built upon the Qwen2.5-VL-7B-Instruct model (Bai et al., 2025). For computational efficiency, we process videos by sampling frames at 2 FPS. See more training details in Appendix B.

Evaluation Metrics. Following established protocols (Ren et al., 2024; Huang et al., 2024a), we report R1@m at various IoU thresholds. This metric calculates the percentage of test samples where the IoU between the top-ranked predicted temporal segment and the ground truth exceeds a given threshold $m \in \{0.3, 0.5, 0.7\}$. As a complementary metric, we also report the mean IoU (mIoU) averaged across the entire test set. For the video understanding and reasoning tasks, we evaluate performance using standard accuracy. To further quantitatively assess the model's comprehension of temporal directionality, we introduce the **Temporal Directionality Discrepancy (TDD)** metric. The core idea behind TDD is that a model that truly understands temporal direction should behave differently based on an event's intrinsic time sensitivity. It is formally defined as:

$$\text{TDD(m)} = \frac{\text{R1@m(fwd)} - \text{R1@m(rev)}}{\text{R1@m(fwd)}}, \tag{11}$$

where R1@m(fwd) denotes R1@m when predictions align with the ground truth $\mathcal{T}_{gt}$ on the forward video $\mathcal{V}$, and R1@m(rev) denotes R1@m on the reversed video $\mathcal{V}'$ with respect to the mirrored ground truth $\mathcal{R}(\mathcal{T}_{gt})$. The interpretation of the TDD depends on the event category. For Time-sensitive events, ideal models should accurately localize the forward event (R1@m(fwd) $\rightarrow$ 1) but recognize its absence upon reversal due to the nonexistence of the event (R1@m(rev) $\rightarrow$ 0), yielding a TDD approaching 1. For Time-insensitive events, models should demonstrate temporal invariance by consistently localizing the event in both directions (R1@m(fwd) $\approx$ R1@m(rev)), yielding a TDD approaching 0. This demonstrates that the model correctly recognizes the event's invariance to temporal direction and exhibits strong consistency.

4.2 Main Results

Table 1 compares the performance of ARROWGEV with state-of-the-art methods. As expected, models trained directly on the target benchmarks generally outperform zero-shot approaches. We further observe that RL-based methods consistently surpass those trained with SFT, likely because they optimize directly on temporal signals grounded in videos and the strong generalization ability of RL (Chu et al., 2025; Peng et al., 2025). In particular, ARROWGEV achieves higher accuracy than the strongest SFT-based baseline on all R1@m metrics. Moreover, compared to other RL-based methods, ARROWGEV yields average improvements in R1@m of 2.6% on Charades-STA, 1.9% on ActivityNet, and 2.5% on TVGBench. We attribute these gains to explicit training to understand the temporal structure, which enhances the robustness of the model in localizing events within videos.

Method	Charades-STA				ActivityNet				TVGBench			
	R1@0.3	R1@0.5	R1@0.7	mIOU	R1@0.3	R1@0.5	R1@0.7	mIOU	R1@0.3	R1@0.5	R1@0.7	mIOU
2D-TAN* (Zhang et al., 2020)	57.3	45.8	27.9	-	60.4	43.4	25.0	-	-	-	-	-
UniVTG* (Lin et al., 2023)	72.6	60.2	38.6	-	56.1	43.4	24.3	-	-	-	-	-
SSRN* (Zhu et al., 2022)	-	65.5	42.6	-	-	54.5	33.2	-	-	-	-	-
SnAG* (Mu et al., 2024b)	-	64.6	46.2	-	-	48.6	30.6	-	-	-	-	-
EaTR* (Jang et al., 2023)	-	68.4	44.9	-	-	58.2	37.6	-	-	-	-	-
HawkEye* (Wang et al., 2024)	72.5	58.3	28.8	-	55.9	34.7	17.9	-	-	-	-	-
TimeSuite* (Zeng et al., 2025)	79.4	67.1	43.0	-	-	-	-	-	-	-	-	-
ChatVTG (Qu et al., 2024)	52.7	33.0	15.9	-	40.7	22.5	9.4	-	-	-	-	-
TimeChat (Ren et al., 2024)	46.7	32.2	15.7	32.2	30.2	16.9	8.2	21.8	22.4	11.9	5.3	-
HawkEye (Wang et al., 2024)	50.6	31.4	14.5	-	49.1	29.3	10.7	-	-	-	-	-
VTimeLLM (Huang et al., 2024a)	51.0	27.5	11.4	31.2	44.0	27.8	14.3	30.4	-	-	-	-
TimeSuite (Zeng et al., 2025)	69.9	48.7	24.0	-	-	-	-	-	31.1	18.0	8.9	-
Momentor (Qian et al., 2024)	42.9	23.0	12.4	29.3	42.6	26.6	11.6	28.5	-	-	-	
VTG-LLM (Guo et al., 2025)	52.0	33.8	15.7	-	–	8.3	3.7	12.0	-	-	-	-
Time-R1† (Wang et al., 2025b)	77.6	59.0	32.4	52.4	55.2	36.4	19.7	38.1	40.4	27.0	12.6	27.5
TVG-R1† (Chen et al., 2025a)	60.7	36.1	13.8	39.3	53.9	33.7	17.5	37.6	32.1	18.1	9.4	23.0
VideoChat-Flash (Li et al., 2024c)	74.5	53.1	27.6	-	-	-	-	-	32.8	19.8	10.4	-
TRACE (Guo et al., 2024b)	-	40.3	19.4	-	-	37.7	24.0	-	37.0	25.5	14.6	-
Qwen-2.5-VL-7B †	59.6	38.7	16.5	38.3	33.9	21.4	13.1	25.2	25.4	16.4	8.5	17.8
ARROWGEV-7B	**78.0**	**61.6**	**37.2**	**54.1**	**58.5**	**38.2**	**20.3**	**39.9**	**41.9**	**29.5**	**16.0**	**29.2**

Table 1: Results on GEV benchmarks. The methods marked in gray* represent fine-tuning on corresponding benchmarks, while those in black indicate zero-shot settings. † denotes that the results are reproduced with the official weights for fair comparison.

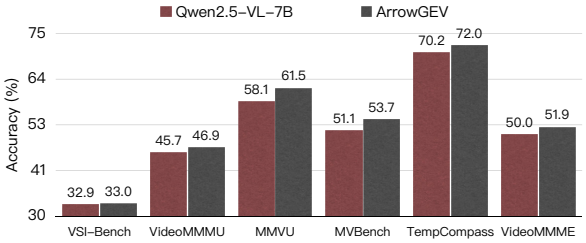

Figure 4: OOD results on six general video understanding and reasoning benchmarks.

4.3 OOD Generalization

Beyond GEV tasks, we further evaluate ARROWGEV on general video understanding and reasoning benchmarks. As shown in Figure 4, ARROWGEV yields consistent and significant improvements over the base Qwen2.5-VL-7B model across all benchmarks. These results highlight the versatility of our approach and demonstrate that explicitly modeling the arrow of time enhances OOD generalization across diverse video understanding and reasoning scenarios. We attribute this improvement to ARROWGEV's ability to deeply understand time structure, which is a core element of general video understanding and reasoning.

4.4 Improvement on Temporal directionality Understanding

To quantify temporal directionality, we evaluate ARROWGEV on both the time-sensitive subset and the insensitive subset. For the time-sensitive subset, a larger TDD indicates better temporal directionality understanding, as the model is less likely to associate meaning-changing timestamps in the backward video with the event. Conversely, for the insensitive subset, a smaller TDD score reflects better temporal directionality understanding, as it suggests the model consistently recognizes events in both forward and backward videos.

As illustrated in Figure 5, ARROWGEV achieves a dramatic improvement over the base model in the time-sensitive and time-insensitive subsets. For time-sensitive subsets (upper row), the TDD score increases markedly, demonstrating that ARROWGEV successfully discriminates semantics that diverge under temporal reversal. Conversely, on the time-insensitive subset (bottom row), ARROWGEV achieves a significant reduction in TDD. This reduction confirms that the model maintains consistent localization across temporal flips, ensuring robust performance regardless of event directionality. These results indicate that ARROWGEV instills a deeper and more robust understanding of temporal directionality in VLMs. We further provide a case study in Appendix D.

4.5 Ablation Study

We conduct ablation studies to analyze the impact of individual components in ARROWGEV. Table 2 summarizes component-wise ablations: 1) Incorpo-

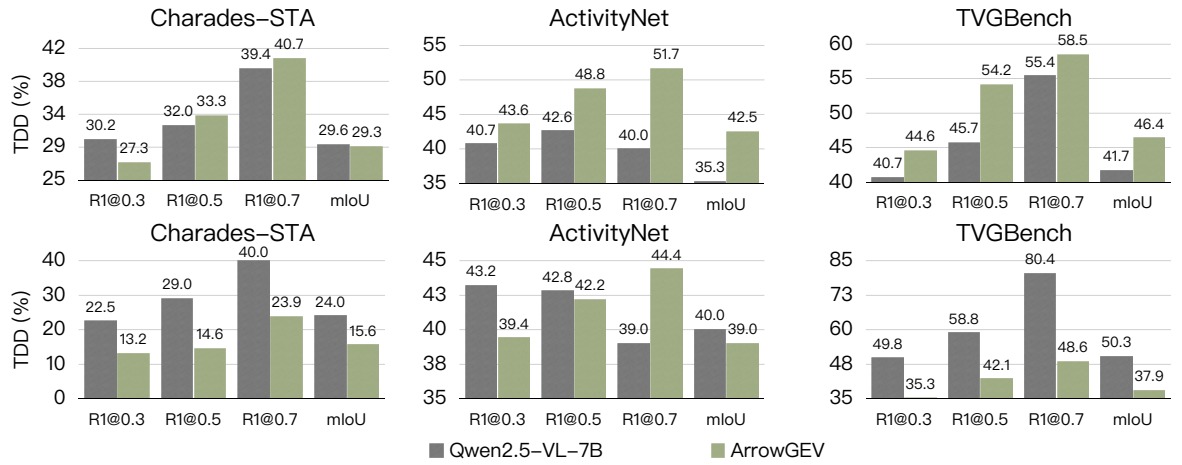

Figure 5: Results of the temporal directionality discrepancy (TDD) metric on three benchmarks. The upper row reports results on the time-sensitive subset, where higher values indicate better temporal directionality understanding, while the bottom row shows results on the time-insensitive subset, where lower values are preferable.

Method	Charades-STA				ActivityNet				TVGBench			
	R1@0.3	R1@0.5	R1@0.7	mIOU	R1@0.3	R1@0.5	R1@0.7	mIOU	R1@0.3	R1@0.5	R1@0.7	mIOU
Qwen-2.5-VL-7B	59.6	38.7	16.5	38.3	33.9	21.4	13.1	25.2	25.4	16.4	8.5	17.8
+ GRPO	74.0	56.1	30.8	50.3	56.1	36.0	19.1	38.3	38.7	26.2	15.2	27.3
+ Weight	76.6	58.9	32.5	52.1	55.4	36.2	19.6	37.9	40.3	26.7	14.7	27.6
+ Filtering	75.2	58.6	35.1	51.9	56.1	37.1	19.7	38.7	41.0	28.5	14.1	28.0
+ Temporal Reward	**78.0**	**61.6**	**37.2**	**54.1**	**58.5**	**38.2**	**20.3**	**39.9**	**41.9**	**29.5**	**16.0**	**29.2**

Table 2: Ablation results on three GEV benchmarks.

rating the temporal reward substantially enhances robustness in event grounding. Removing this term leads to a 4–6% drop in R1@m across benchmarks relative to vanilla GRPO, underscoring the effectiveness of capturing temporal directionality for grounding performance. 2) Both weight adjustment and difficulty filtering contribute to consistent gains. Weight adjustment guides the model to focus on harder samples, while filtering removes already-solved examples, preventing wasted capacity. Together, these strategies deliver measurable improvements in precision and stability. 3) Beyond component-level refinements, adopting GRPO itself provides a strong performance boost: the base VLM achieves a 10–30% improvement on GEV benchmarks, confirming the necessity of reinforcement learning for event grounding.

4.6 Performance on Different VLMs

To evaluate the versatility of our learning framework, we conducted experiments using different VLMs. Specifically, we examined the performance of our framework on Qwen-2.5-VL-Instruct-3B, and compared it with the best-performing RL baseline. Table 3 presents the mIoU metric on the benchmarks. The results indicate that ARROWGEV consistently improve the performance of the base model and outperform the best-performing baseline, indicating the effectiveness of our framework across different pre-trained VLMs. We report the performance on all metrics in Appendix C.

Method	Charades-STA	ActivityNet	TVGBench
Time-R1-3B	40.7	23.7	19.8
Qwen-2.5-VL-3B	28.7	15.7	12.6
ARROWGEV-3B	**42.3**	**24.3**	**20.3**

Table 3: mIoU metric across three GE benchmarks.

5 Conclusion

We introduce ARROWGEV, which trains VLMs to enable both event grounding and video understanding through learning the temporal directionality. In addition to learning event localization in forward videos, ARROWGEV aims to distinguish the absence of time-sensitive events and recognize time-insensitive events in backward videos. Extensive experiments demonstrate that ARROWGEV outperforms various baselines on the event grounding task and improves general video understanding and reasoning performance compared to the base model.

Limitation

While our method demonstrates strong performance, the computational resources required in this model are expensive. Additionally, although our work primarily focuses on grounding events in videos, we believe that learning the *Arrow of Time* can be extended to more video reasoning tasks. We leave the exploration of this direction as future work.

References

2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812.

Md Adnan Arefeen, Biplob Debnath, Md Yusuf Sarwar Uddin, and Srimat Chakradhar. 2024. Vita: An efficient video-to-text algorithm using vlm for rag-based video analysis system. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2266–2274.

Y Alp Aslandogan and Clement T. Yu. 2002. Techniques and systems for image and video retrieval. *IEEE transactions on Knowledge and Data Engineering*, 11(1):56–63.

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, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970.

Ruizhe Chen, Zhiting Fan, Tianze Luo, Heqing Zou, Zhaopeng Feng, Guiyang Xie, Hansheng Zhang, Zhuochen Wang, Zuozhu Liu, and Huaijian Zhang. 2025a. Datasets and recipes for video temporal grounding via reinforcement learning. *arXiv preprint arXiv:2507.18100*.

Yi-Wen Chen, Yi-Hsuan Tsai, and Ming-Hsuan Yang. 2021. End-to-end multi-modal video temporal grounding. *Advances in Neural Information Processing Systems*, 34:28442–28453.

Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, and 1 others. 2025b. Scaling rl to long videos. *arXiv preprint arXiv:2507.07966*.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. 2025. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*.

Yang Du, Yuqi Liu, and Qin Jin. 2024. Reversed in time: A novel temporal-emphasized benchmark for cross-modal video-text retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5260–5269.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. 2019. Temporal cycle-consistency learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1801–1810.

Arthur Eddington. 2019. *The nature of the physical world: THE GIFFORD LECTURES 1927*, volume 23. BoD–Books on Demand.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. 2025. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. 2020. Multi-modal transformer for video retrieval. In *European Conference on Computer Vision*, pages 214–229. Springer.

Jarvis Guo, Tuney Zheng, Yuelin Bai, Bo Li, Yubo Wang, King Zhu, Yizhi Li, Graham Neubig, Wenhu Chen, and Xiang Yue. 2024a. Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale. *arXiv preprint arXiv:2412.05237*.

Yongxin Guo, Jingyu Liu, Mingda Li, Dingxin Cheng, Xiaoying Tang, Dianbo Sui, Qingbin Liu, Xi Chen, and Kevin Zhao. 2025. Vtg-llm: Integrating timestamp knowledge into video llms for enhanced video temporal grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3302–3310.

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

Tanveer Hannan, Md Mohaiminul Islam, Thomas Seidl, and Gedas Bertasius. 2024. Rgnet: A unified clip retrieval and grounding network for long videos. In *European Conference on Computer Vision*, pages 352–369. Springer.

Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, and 1 others. 2024. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*.

Zhijian Hou, Wanjun Zhong, Lei Ji, Difei Gao, Kun Yan, Wing-Kwong Chan, Chong-Wah Ngo, Zheng Shou, and Nan Duan. 2022. Cone: An efficient coarse-to-fine alignment framework for long video temporal grounding. *arXiv preprint arXiv:2209.10918*.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. 2025. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*.

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

De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. 2024b. Lita: Language instructed temporal-localization assistant. In *European Conference on Computer Vision*, pages 202–218. Springer.

Vladimir Iashin and Esa Rahtu. 2020. Multi-modal dense video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 958–959.

Jinhyun Jang, Jungin Park, Jin Kim, Hyeongjun Kwon, and Kwanghoon Sohn. 2023. Knowing where to focus: Event-aware transformer for video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13846–13856.

Yogesh Kulkarni and Pooyan Fazli. 2025. Avatar: Reinforcement learning to see, hear, and reason over video. *arXiv preprint arXiv:2508.03100*.

David Layzer. 1975. The arrow of time. *Scientific American*, 233(6):56–69.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024b. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, Yu Qiao, Yali Wang, and Limin Wang. 2024c. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*.

Yan Li, Bin Ji, Xintian Shi, Jianguo Zhang, Bin Kang, and Limin Wang. 2020. Tea: Temporal excitation and aggregation for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 909–918.

Zeqian Li, Shangzhe Di, Zhonghua Zhai, Weilin Huang, Yanfeng Wang, and Weidi Xie. 2025. Universal video temporal grounding with generative multi-modal large language models. *arXiv preprint arXiv:2506.18883*.

Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Van Tu Vu, and 1 others. 2024d. Groundinggpt: Language enhanced multi-modal grounding model. *arXiv preprint arXiv:2401.06071*.

Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. 2023. Univtg: Towards unified video-language temporal grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2794–2804.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*.

Ziyang Luo, Haoning Wu, Dongxu Li, Jing Ma, Mohan Kankanhalli, and Junnan Li. 2025. Videoautoarena: An automated arena for evaluating large multimodal models in video analysis through user simulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8461–8474.

Willi Menapace, Aliaksandr Siarohin, Ivan Skorokhodov, Ekaterina Deyneka, Tsai-Shien Chen, Anil Kag, Yuwei Fang, Aleksei Stoliar, Elisa Ricci, Jian Ren, and 1 others. 2024. Snap video: Scaled spatiotemporal transformers for text-to-video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7038–7048.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, and 1 others. 2025. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*.

Ishan Misra, C Lawrence Zitnick, and Martial Hebert. 2016. Shuffle and learn: unsupervised learning using temporal order verification. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 527–544. Springer.

Fangzhou Mu, Sicheng Mo, and Yin Li. 2024a. Snag: Scalable and accurate video grounding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18930–18940.

Fangzhou Mu, Sicheng Mo, and Yin Li. 2024b. Snag: Scalable and accurate video grounding. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18930–18940.

Guoshun Nan, Rui Qiao, Yao Xiao, Jun Liu, Sicong Leng, Hao Zhang, and Wei Lu. 2021. Interventional video grounding with dual contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2765–2775.

Thong Thanh Nguyen, Yi Bin, Xiaobao Wu, Zhiyuan Hu, Cong-Duy T Nguyen, See-Kiong Ng, and Anh Tuan Luu. 2025. Multi-scale contrastive learning for video temporal grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 6227–6235.

Yulin Pan, Xiangteng He, Biao Gong, Yiliang Lv, Yujun Shen, Yuxin Peng, and Deli Zhao. 2023. Scanning only once: An end-to-end framework for fast temporal grounding in long videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13767–13777.

Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. 2025. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*.

Will Price and Dima Damen. 2019. Retro-actions: Learning'close'by time-reversing'open'videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0.

Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat-Seng Chua, Yueting Zhuang, and Siliang Tang. 2024. Momentor: Advancing video large language model with fine-grained temporal reasoning. *arXiv preprint arXiv:2402.11435*.

Mengxue Qu, Xiaodong Chen, Wu Liu, Alicia Li, and Yao Zhao. 2024. Chatvtg: Video temporal grounding via chat with video dialogue large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1847–1856.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 2024. 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*, pages 14313–14323.

Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. 2022. End-to-end generative pretraining for multimodal video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17959–17968.

Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. 2025. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*.

Zhenxiong Tan, Xingyi Yang, Songhua Liu, and Xinchao Wang. 2024. Video-infinity: Distributed long video generation. *arXiv preprint arXiv:2406.16260*.

Kwai Keye Team, Biao Yang, Bin Wen, Changyi Liu, Chenglong Chu, Chengru Song, Chongling Rao, Chuan Yi, Da Li, Dunju Zang, and 1 others. 2025. Kwai keye-vl technical report. *arXiv preprint arXiv:2507.01949*.

Shulin Tian, Ruiqi Wang, Hongming Guo, Penghao Wu, Yuhao Dong, Xiuying Wang, Jingkang Yang, Hao Zhang, Hongyuan Zhu, and Ziwei Liu. 2025. Ego-r1: Chain-of-tool-thought for ultra-long egocentric video reasoning. *arXiv preprint arXiv:2506.13654*.

Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2018. Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2740–2755.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025a. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, Xiangnan Fang, Zewen He, Zhenbo Luo, Wenxuan Wang, Junqi Lin, Jian Luan, and Qin Jin. 2025b. Time-r1: Post-training large vision language model for temporal video grounding. *arXiv preprint arXiv:2503.13377*.

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, and 1 others. 2025c. Time-r1: Post-training large vision language model for temporal video grounding. *arXiv preprint arXiv:2503.13377*.

Yueqian Wang, Xiaojun Meng, Jianxin Liang, Yuxuan Wang, Qun Liu, and Dongyan Zhao. 2024. Hawkeye: Training video-text llms for grounding text in videos.

Donglai Wei, Joseph J Lim, Andrew Zisserman, and William T Freeman. 2018. Learning and using the arrow of time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8052–8060.

Jongbhin Woo, Hyeonggon Ryu, Youngjoon Jang, Jae Won Cho, and Joon Son Chung. 2024. Let me finish my sentence: Video temporal grounding with holistic text understanding. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8199–8208.

LLM-Core-Team Xiaomi. 2025. Mimo-vl technical report. *Preprint*, arXiv:2506.03569.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*.

Zihui Xue, Mi Luo, and Kristen Grauman. 2025. Seeing the arrow of time in large multimodal models. *ArXiv*, abs/2506.03340.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Haolin Yang, Feilong Tang, Ming Hu, Qingyu Yin, Yulong Li, Yexin Liu, Zelin Peng, Peng Gao, Junjun He, Zongyuan Ge, and 1 others. 2025a. Scalingnoise: Scaling inference-time search for generating infinite videos. *arXiv preprint arXiv:2503.16400*.

Haolin Yang, Feilong Tang, Linxiao Zhao, Xiang An, Ming Hu, Huifa Li, Xinlin Zhuang, Boqian Wang, Yifan Lu, Xiaofeng Zhang, and 1 others. 2025b. Streamagent: Towards anticipatory agents for streaming video understanding. *arXiv preprint arXiv:2508.01875*.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2025c. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643.

Weihong Ye, Yachen Wang, He Wang, Zhipeng Zhang, Wenhao Li, and Xin Wang. 2025. Timezero: A reasoning-driven lvlm for temporal video grounding via reinforcement learning. *arXiv preprint arXiv:2503.13377*.

Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, Yali Wang, Yu Qiao, and Limin Wang. 2025. Timesuite: Improving MLLMs for long video understanding via grounded tuning. In *The Thirteenth International Conference on Learning Representations*.

Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. 2020. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, and 1 others. 2025. Mmvu: Measuring expert-level multi-discipline video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8475–8489.

Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. 2017. Temporal action detection with structured segment networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2914–2923.

Jiahao Zhu, Daizong Liu, Pan Zhou, Xing Di, Yu Cheng, Song Yang, Wenzheng Xu, Zichuan Xu, Yao Wan, Lichao Sun, and Zeyu Xiong. 2022. Rethinking the video sampling and reasoning strategies for temporal sentence grounding. In *Findings of the Association for Computational Linguistics: EMNLP 2022*.

Method	Charades-STA				ActivityNet				TVGBench			
	R1@0.3	R1@0.5	R1@0.7	mIOU	R1@0.3	R1@0.5	R1@0.7	mIOU	R1@0.3	R1@0.5	R1@0.7	mIOU
Time-R1-3B	62.6	40.0	18.2	40.7	34.8	19.8	9.0	23.7	30.5	17.0	**8.0**	19.8
Qwen-2.5-VL-3B	42.9	26.9	12.8	28.7	22.0	11.5	4.8	15.7	18.0	9.8	5.2	12.6
ARROWGEV-3B	**64.9**	**42.5**	**19.7**	**42.3**	**36.1**	19.4	**9.9**	**24.3**	**31.0**	**18.6**	7.5	**20.3**

Table 4: Results on three GEV benchmarks with 3B model.

Appendix

A Event Categorization Analysis

We use Qwen2.5-72B-Instruct (Yang et al., 2024) to categorize the type of events. To verify the reliability of the categorization, we watch the video and label 100 events from each dataset and compute the accuracy of the model prediction. The results in Table 5 indicate that LLMs's prediction is highly aligned with humans, and can well recognize whether an event is time-sensitive or time-insensitive.

B Training Details

We leverage 7B and 3B models of Qwen2.5-VL (Bai et al., 2025) series as our base model. They are trained on large scale image and video data and show strong instruction following and reasoning abilities. During the post-training stage, we train the model for 5 epochs, and set a batch size of 128, learning rate 2e-5, number of candidate response $G = 8$, and coefficient $\lambda = 0.5$, KL term $\beta = 0$, temperature in weight adjust $\tau = 2$. We search these hyperparameters on the validation set. The checkpoint from the final epoch is used for all evaluations. All experiments were conducted on a single node equipped with 8 $\times$ H20 GPUs.

C Additional Results.

We report the performance of ARROWGEV and compare with the high-performing RL baseline and the base model in Table 4 on the 3B model. The results show that ARROWGEV outperforms the strong RL baseline, Time-R1-3B, on most of the metrics across all three benchmarks. These results serve as evidence for the robustness and effectiveness of our framework, demonstrating its ability to generalize and enhance the temporal reasoning capabilities of different pre-trained VLMs.

Datasets	**Charades-STA**	**Activitynet**	**TVGBench**
Accuracy (%)	94.0	92.0	96.0

Table 5: Event categorization performance on three benchmarks.

D Case Study

Figure 6 presents qualitative comparisons against high-performing RL baselines and Qwen2.5-VL-7B, evaluating both time-sensitive and time-insensitive events. For the time-sensitive event "person opens the door," ARROWGEV achieves precise localization in the forward video. Crucially, when the video is reversed, the event in the video changes to "person closes the door". ARROWGEV correctly identifies the absence of the queried event. In contrast, competing methods erroneously localize the reversed "close" action as "open", failing to comprehend temporal directionality. When analyzing the time-insensitive event "person smiling at the laptop," ARROWGEV identifies the time interval in both forward and backward videos. These results demonstrate ARROWGEV's superior understanding of temporal semantics and robustness compared to other approaches.

E Prompt

We list the prompt for event categorization in Figure 7.

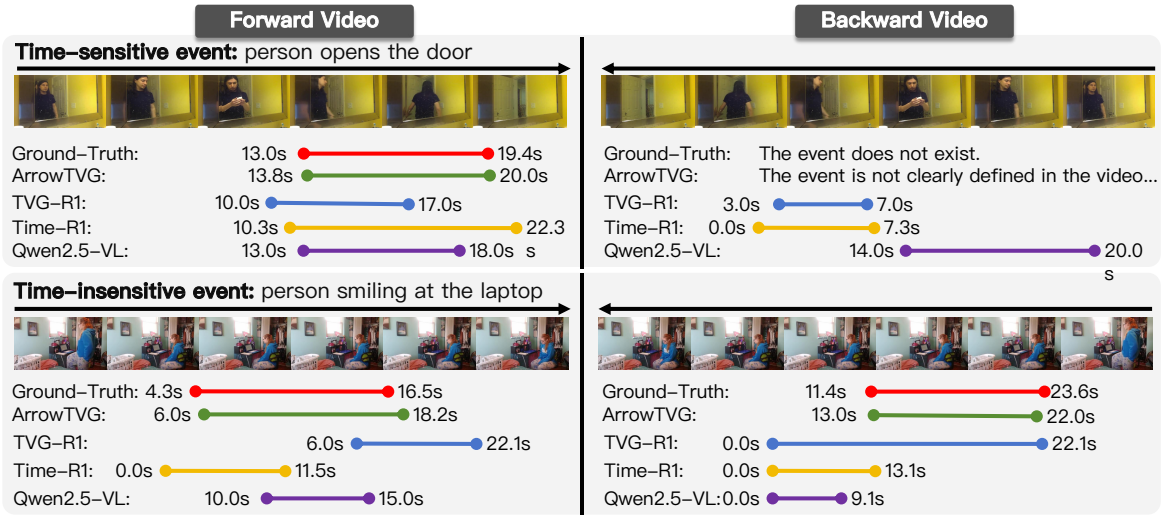

Figure 6: Case study on time-sensitive events (upper row) and time-insensitive events (bottom row). Although baseline methods can partially localize the correct timestamps in the forward video, they fail to distinguish the nonexistence of time-sensitive events in the backward videos and localize the time-insensitive events in the backward video.

Prompts for time sensitivity reasoning

You are an AI assistant specializing in the analysis of temporal properties of events.
You will be given a sentence describing an event.
Your task is to:
1. Analyze the event described in the sentence.
2. Determine if the event is temporally sensitive or insensitive.
3. Output the results in a strict JSON format without any additional text or explanations.
—
Input
Event Sentence: ${sentence}
—
Evaluation Criteria
Time-Sensitive (sensitive: yes): The event has a clear forward direction. If played in reverse, it describes a different, often nonsensical or opposite, event. This indicates temporal asymmetry.
Example: "A person puts a picture on the wall." (Reversed: "A person takes a picture off the wall.")
Example: "A glass shatters." (Reversed: "Shards of glass assemble into a whole glass.")
Time-Insensitive (sensitive: no): The event is a continuous state or a cyclical action. If played in reverse, the fundamental nature of the event does not change. This indicates temporal symmetry.
Example: "A person is playing with a light switch." (Reversed: Still looks like a person playing with a light switch.)
Example: "A ball is bouncing in place." (Reversed: Still a ball bouncing in place.)
—
Output Format
Now, please output your result below in a JSON format by filling in the placeholders in [] without any explanations:
"reason": "[Briefly explain why the event is time-sensitive or time-insensitive, describing the forward and reverse action.]",
"sensitive": "[yes/no]"
"""

Figure 7: The instruction for LLM to categorize events into different types.