Test-time Counterfactual Calibration for Hallucination-Resistant Temporal Grounding

Chufan Yi[1⋆], Hongyu Qu[1⋆], Shiyu Xuan[1], Rui Yan[1], Xiangbo Shu[1✉], Fang Zhao[2], and Guo-Sen Xie[1✉]

[1] Nanjing University of Science and Technology, China
{chufanyi,quhongyu,shiyu_xuan,ruiyan,shuxb}@njust.edu.cn
gsxiehm@gmail.com
[2] Nanjing University, China
fzhao@nju.edu.cn

Abstract. Existing video temporal grounding methods typically operate under a closed-world assumption, relying on the implicit premise that every natural language query must have a corresponding event somewhere in the video. This always-answer paradigm severely compresses the decision boundary during training, as models are seldom exposed to the alternative hypothesis that the queried event may be absent. Consequently, when users ask about specific events that cannot be matched in the video within an open-world environment, these models exhibit systematic temporal hallucinations. This may stem from incorrect attribution of textual, visual, or multimodal information, leading them to confidently generate plausible time frames for events that do not actually exist. To address this limitation, we propose HRVTG, a test-time adaptation framework that dynamically calibrates the decision boundary of the model during inference. Built upon a frozen grounding backbone, the proposed framework constructs three types of counterfactual hallucination probes in a self-supervised manner, targeting the aforementioned hallucinations. We dynamically reshape the decision boundary through abstention signals from counterfactual probes and consistency rewards from genuine queries. This is executed by introducing the GRPO algorithm, effectively reformulating the task into an online policy optimization problem in reinforcement learning. Furthermore, we introduce decoupled metrics to independently evaluate hallucination resistance and real-event grounding. Experiments indicate that our method suppresses hallucinated responses to fabricated queries. It achieves accurate selective prediction while maintaining strong grounding performance on real events. The code will be released at https://github.com/CVL-hub/HRVTG.

Keywords: Video Temporal Grounding · Reinforcement Learning · Test-time Adaptation · Multimodal Large Language Models

⋆ Equal contribution. ✉ Corresponding author.

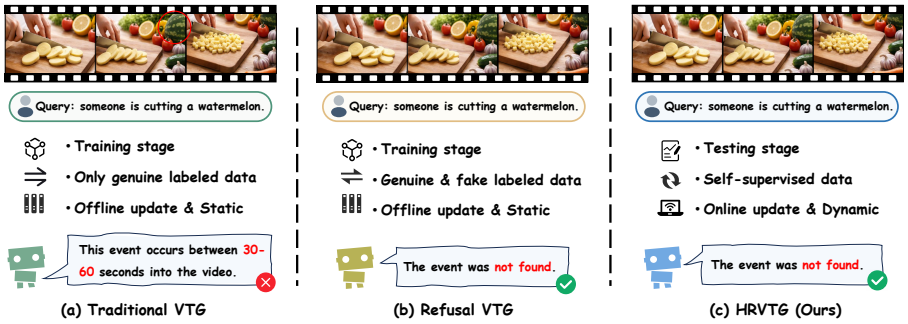

Fig. 1: Comparison of three VTG paradigms under a counterfactual query. For example, while the video depicts chopping a potato, the counterfactual query specifies "cutting a watermelon." (a) Traditional VTG, trained only on positive samples, exploits semantic priors around the action "cut" to hallucinate a spurious interval. (b) Static Relevance VTG learns to output a null-grounding signal via offline training on predefined negative data, but its decision boundary remains static. (c) Our HRVTG calibrates the decision boundary dynamically during inference through self-supervised counterfactual probes, achieving selective prediction without annotations.

1 Introduction

Video temporal grounding (VTG) is a core capability in modern video understanding systems. Given an untrimmed video and a natural language query, a VTG model is required to return the precise start and end timestamps of the event described by the query. This capability underpins a wide range of applications, including video retrieval [9,25,39,53,64], highlight extraction [6,26,38,51], interactive video editing [3–5,35], and agentic systems [24,46,59] that must reason about *when* something happens rather than only *what* happens.

Early VTG methods [2,22,50] predominantly follow an encode-interact-decode paradigm, encoding video and text separately, computing segment relevance through cross-modal interaction, and regressing temporal boundaries. More recently, VTG methods [15,32,58] have increasingly adopted multimodal large language models (MLLMs) as backbones and moved toward a generative paradigm, where timestamps are produced via structured text generation (*e.g.*, the model first outputs a rationale and then reports "start to end" within an `<answer>` field). While this paradigm improves expressivity and reasoning, it also amplifies a long-standing yet overlooked assumption in VTG, where both training and evaluation presume that the queried event *almost always exists* in the video.

This *always-answer* bias learned under a closed-world setting can induce temporal hallucinations and yield erroneous outputs in open-world deployment. Rather than verifying visual evidence, models often exploit semantic language priors or visual similarities to forcibly align counterfactual queries with mismatched video segments, producing confident but spurious intervals (Fig. 1(a)). The absence of selective temporal grounding capability in such scenarios severely

hinders VTG deployment in high-reliability settings such as surveillance [27,41], medical investigation [63,65], and robot navigation [7,23]. Recent studies [10,17, 45] attempt to calibrate this forced-alignment bias by incorporating predefined negative samples during offline training (Fig. 1(b)). However, these methods inherently learn a static set of fabrication patterns, and this offline calibration often fails when deployed against video distributions or unseen hallucination triggers. We argue that the core issue is not merely a lack of a designated rejection class. Rather, it reflects a miscalibrated decision boundary at inference time that prevents the model from effectively distinguishing genuine evidence-supported grounding from null-attribution states under insufficient evidence.

To systematically mitigate this forced-alignment bias, we first introduce a taxonomy of VTG hallucinations, including text-driven, vision-driven, and multimodal joint hallucinations. Guided by this taxonomy, we propose a self-supervised paradigm to construct counterfactual hallucination probes at test time. These probes automatically synthesize missing-event scenarios and attach verifiable pseudo-labels directly from standard positive-only benchmarks, eliminating the need for additional human annotation. Building on these probes, we propose HRVTG, a test-time adaptation framework that supports online streaming temporal grounding. We adopt Group Relative Policy Optimization (GRPO) with verifiable rewards [14] and reformulate grounding as an online policy optimization problem, enabling dynamic decision-boundary adjustment during inference. We introduce multiple rewards: (i) a consistency anchor reward aligns the model with a frozen strong grounder on genuine queries to preserve grounding ability; (ii) a counterfactual repulsion reward suppresses temporal hallucinations triggered by missing events; (iii) an asymmetric KL elastic constraint enables a controllable trade-off between stability and plasticity. Our implementation reuses and caches video features to support streaming updates, making test-time calibration practical for deployment. In addition, we introduce two decoupled evaluation metrics, Conditional IoU and Balanced Decision F1-score, to alleviate the coupling issues of prior metrics. Experimental results show that our model improves resistance to fabricated events or hallucinations while preserving, or slightly enhancing, the original grounding ability.

Overall, the main contributions are threefold:

- We propose a taxonomy of temporal hallucinations in VTG, characterizing triggers from multiple perspectives, and use it to construct self-supervised counterfactual probes that yield high-quality missing-event samples.
- We present a GRPO-based test-time adaptation framework that dynamically reshapes the decision boundary through abstention signals from counterfactual probes and consistency rewards from genuine queries, improving hallucination resistance while preserving real-event grounding.
- We introduce decoupled evaluation metrics that separately assess grounding precision and selective prediction quality.

2 Related Work

2.1 MLLMs for Video Temporal Grounding

Early video temporal grounding methods [11, 18, 26, 54, 57, 61] predominantly follow an encode-interact-decode paradigm. They encode video and text separately, compute segment relevance through cross-modal interaction, and regress temporal boundaries. While models like Moment-DETR [18] and UMT [26] establish a solid foundation, they are often constrained by closed vocabularies and limited reasoning. To address this, subsequent work incorporates vision-language pretraining [19, 30] and LLMs [40] for semantic enrichment, as exemplified by query-rewriting frameworks such as ED-VTG [29]. Recent approaches directly adopt video MLLMs as backbones, reformulating temporal grounding as structured text generation. VTimeLLM [15] and Grounded-VideoLLM [43] enable precise grounding through temporal-aware training and discretized time tokens. To bridge the gap between cross-entropy loss and grounding metrics (IoU), reinforcement learning with verifiable rewards has been introduced [20, 28, 49]. VideoChat-R1 [21] and TimeZero [48] apply temporal-aware rule-based rewards to enhance video reasoning, while Time-R1 [47] further incorporates curriculum learning and cold-start CoT for more comprehensive performance gains.

However, the generative paradigm solidifies an always-answer bias into a forced-alignment pattern at inference. Whether trained via SFT or RL, models persistently return intervals even for non-existent events, inducing severe temporal hallucinations in open-world deployment. To mitigate this, we propose a GRPO-based framework that introduces verifiable rewards and counterfactual hallucination probes at test time. By performing online policy calibration on a frozen strong grounder, our method learns an adaptive decision boundary resistant to forced-alignment hallucinations.

2.2 Test-time Adaptation

Test-time Adaptation (TTA) leverages unlabeled test data to update a model online, mitigating performance degradation caused by train–test distribution shifts. Classic paradigms include entropy-minimization updates such as Tent [42] and continual, streaming-oriented methods such as CoTTA [44]. Early TTA primarily focuses on image tasks [36, 37, 60], where TPT [37] treats learnable prompts as lightweight adaptable components. This idea has also been extended to handle temporal distribution shifts [12, 13], and DTS-TPT [55] further introduces multi-step temporal sampling and dual semantic consistency for zero-shot action recognition.

More recent work [52, 56, 66] reframes TTA as reinforcement learning with learnable signals, using self-consistency as a reward. TTRL [66] observes that consensus mechanisms (*e.g.*, majority voting) can provide unlabeled yet useful reward cues for inference-time self-evolution. MM-UPT [52] extends this to multimodal settings via pseudo rewards and synthetic questions for continual self-learning. Our work focuses on calibrating the semantic decision boundary by applying online continual updates to VTG during test. Specifically, self-supervised

counterfactual probes provide verifiable repulsion signals, pushing TTA beyond distribution adaptation toward debiasing and boundary calibration.

2.3 Selective Prediction in VTG

In real-world applications, user queries often contain counterfactual statements, mismatched details, or events that do not exist in the video at all. Particularly under the MLLM for video understanding paradigm, models are trained to generate structured textual outputs, which makes them prone to producing high-confidence temporal hallucinations and misleading downstream decision-making. Recent work [10, 17, 45] has framed this behavior as relevance judgment and query refusal [62]. Conventional VTG implicitly assumes all samples are positive. RaTSG [10] extends this into a setting with relevance feedback, enabling end-to-end prediction that localizes events when present and abstains when absent. Building on this, CausalVTG [45] explicitly models dataset bias from a causal perspective and learns less biased cross-modal representations via causal adjustment, thereby supporting the detection of missing query content. More recently, RA-RFT [17] uses reinforcement learning to unify grounding and refusal under a single policy-optimization objective, teaching the model to abstain on hard-irrelevant queries while providing explanations.

Most existing refusal/relevance methods rely on offline-constructed negative samples or explicit refusal annotations, and they solidify refusal capabilities during training. When deployed under new query distributions, domain shifts in videos, or more deceptive pseudo-evidence segments, models may still produce temporal hallucinations because they lack dynamic calibration at inference time. Motivated by this, we propose a taxonomy of temporal hallucinations in video-text generation for VTG and use it to construct three types of counterfactual probes in a self-supervised manner. Based on these probes, HRVTG dynamically examines the model's decision boundary during inference and performs on-the-fly calibration via reinforcement learning updates, endowing the model with immunity to temporal hallucination triggers encountered during inference.

3 VTG Hallucination Taxonomy

Existing video temporal grounding (VTG) methods typically operate under a closed-world assumption, implicitly presuming that every natural language query must have a corresponding event interval in the video. This "always-answer" training bias compresses the model's decision boundary at inference. When deployed in open-world settings and faced with counterfactual or unrealizable queries, models constrained by forced-alignment bias cannot output a null-grounding signal, leading to systematic temporal hallucinations. To systematically mitigate this issue, we first define a taxonomy of temporal hallucinations in VTG from the perspective of cognitive mechanisms (Sec. 3.1), and further construct counterfactual probes at test time in a self-supervised manner (Sec. 3.2).

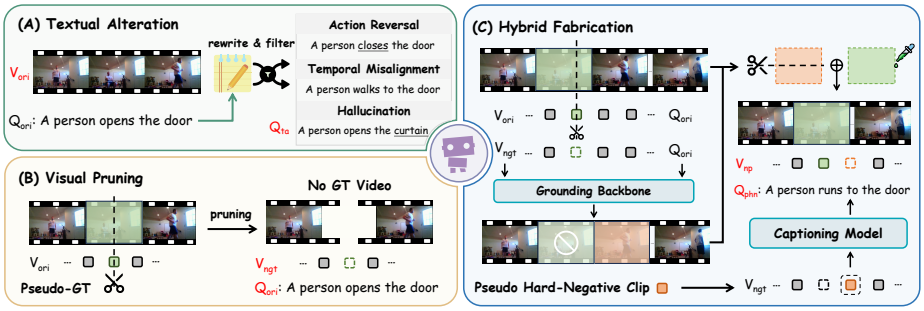

Fig. 2: Overview of our counterfactual probes construction pipeline.

3.1 Temporal Hallucinations in VTG

According to the sources of misattribution produced by multimodal models when handling video-text alignment, we categorize temporal hallucinations in VTG into the following three core patterns:

Text-driven Hallucinations. This type of hallucination arises from the model's excessive reliance on language priors and surface-level word associations, similar to language bias observed in VideoQA [33]. When the query reverses key actions, disrupts temporal order, or tampers with core attributes, the model may still latch onto shallow semantic cues (*e.g.*, specific nouns or verbs) and forcibly align them with mismatched video segments. Typical triggers include:

- (i) Fine-grained mismatch: the query reverses action direction or polarity (*e.g.*, open → close, enter → exit), but the model localizes a similar action segment in the same scene.
- (ii) Disrupted temporal relations: the query imposes an ordering constraint (A before B) absent or inverted in the video, but the model still returns the most similar time window.
- (iii) Attribute/entity substitution: key attributes or entities are replaced with ones absent from the video (*e.g.*, watermelon → potato), but the model treats semantically similar segments as valid evidence.

Vision-driven Hallucinations. This type of hallucination is mainly triggered by global context and redundant visual background. When the target event described by the query is removed, but the video retains similar scene backgrounds or person entities, the model can be misled by coarse-grained visual similarity [1, 8]. This results in seemingly reasonable temporal grounding outputs in the absence of core action evidence. Typical triggers include:

- (i) Strong context but missing evidence: prerequisite actions (*e.g.*, picking up a tool, walking toward a location) are present, yet the key action itself is unseen or does not occur in the video.
- (ii) Repetitive scene consistency: in long videos, recurring person-scene combinations cause the model to mistake scene-related segments for evidence.

– (iii) Inertial alignment under evidence removal: even after the true event segment is excised, the model searches for the most similar window among adjacent clips, exposing its insensitivity to missing evidence.

Multimodal Joint Hallucinations. This is the most deceptive pattern, jointly induced by text and vision. Pseudo-evidence segments in open-world videos can exhibit high overlap with the queried event in both semantic and visual feature spaces, yielding abnormally high cross-modal alignment scores. Combined with naturally phrased queries, the model is led to produce high-confidence intervals under the illusion that both modalities provide support.

The significance of the above VTG hallucination taxonomy is that different hallucination types correspond to different sources of pseudo-evidence and distortion paths, which in turn determines where the most effective calibration signals should be constructed.

3.2 Counterfactual Hallucination Probe Construction

To enable stable hallucination resistance without sacrificing real-event grounding, we argue that the key lies not in injecting a large number of manually created negative samples during offline training, but in ensuring the model continuously encounters missing-event samples that (i) conform to the current video distribution and (ii) carry verifiable pseudo-labels during inference, to calibrate the decision boundary on the fly. Therefore, we propose a self-supervised counterfactual probe construction strategy built upon a frozen strong grounding model. For each real test-stream sample $x^{pos} = (V_{\text{ori}}, Q_{\text{ori}})$, we automatically generate a paired counterfactual sample $x^{cf} = (V_{\text{cf}}, Q_{\text{cf}})$ with a pseudo-label indicating event absence. Guided by the VTG hallucination taxonomy introduced above, counterfactual samples are constructed along three complementary paths.

Textual Alteration. This probe targets text-driven hallucinations by modifying only the query while keeping the video unchanged, thereby exposing forced alignment induced by language priors. The core objective is to construct a counterfactual query Q_{ta} that remains locally similar to the original query Q_{ori} but is semantically invalid, ensuring that the corresponding event is strictly absent in V. This probe checks whether semantic similarity induces the model to produce temporal hallucinations. Given Q_{ori}, we generate minimally shifted counterfactuals via three operation modes:

– (i) Action Reversal: change the action direction/polarity while keeping the subject and scene words as unchanged as possible;
– (ii) Temporal Misalignment: change the event order or temporal logical constraints (precedence relations, simultaneous occurrence, *etc.*);
– (iii) Attribute Hallucination: replace key attributes/entities (object, color, quantity, body part, tool) to make them inconsistent with the video facts.

To ensure high-precision event absence, we apply a retrieval consistency check using the frozen strong grounding model to filter generated candidates. This

probe thereby compels the model to distinguish "superficial semantic similarity" from "genuine evidence support", providing a reliable repulsion signal for decision boundary calibration.

Visual Pruning. This probe specifically targets vision-driven hallucinations by maintaining Q_{ori} while modifying the video to simulate missing-evidence scenarios. We obtain the evidence interval as Pseudo-GT from high-confidence predictions of the frozen strong grounding model [47]. This interval is then excised from the video, and the remaining clips are concatenated in temporal order to yield a video V_{ngt} without the target event.

This probe preserves contextual elements consistent with the original video, including people, scenes, and preceding/succeeding states, while removing the decisive evidence interval. It precisely exposes the erroneous tendency of the model to treat contextual cues as grounding evidence.

Hybrid Fabrication. This probe constructs the most deceptive counterfactual scenario: a pseudo-evidence clip that closely matches the query is present during construction but withheld from the final input, and the query itself is generated from that clip in a self-supervised manner rather than arbitrarily fabricated. The resulting sample thus simultaneously presents a naturally phrased query and a video rich in semantically similar background segments, yet devoid of genuine supporting evidence. Specifically, we first temporarily excise the Pseudo-GT τ_{pgt} from V_{ori} to obtain V_{ngt}. We then apply the frozen strong grounding model with the Q_{ori} on V_{ngt} to retrieve the most relevant interval τ_{phn} as pseudo hard-negative clip. Next, we use a captioning model to generate a textual description Q_{phn} from τ_{phn}, matching the style of Q_{ori}. Finally, we construct the query video V_{phn} by removing τ_{phn} from the original V_{ori} while keeping τ_{pgt} intact.

The three probes construct "event-absent" samples from the textual, visual, and cross-modal perspectives, covering the most prevalent and high-risk hallucination triggers in open-world VTG [31]. Unlike offline negative-sample training with fixed fabrication patterns, these probes are generated on-the-fly for each input video, enabling continuous calibration of the decision boundary toward genuine evidence auditing throughout inference.

4 Methodology

4.1 Preliminary

Problem Formulation. Given an untrimmed video V and a text query q, an MLLM-based generative VTG model takes the prompt $x = \Phi(V, q)$, composed of system and user instructions along with video-modality tokens. The model generates an autoregressive completion $y = (a_1, \ldots, a_T)$, which defines a policy:

$$\pi_\theta(y \mid x) = \prod_{t=1}^{T} \pi_\theta(a_t \mid x, a_{<t}), \tag{1}$$

where θ denotes the policy parameters updated online via a LoRA adapter while the backbone remains frozen. Here, a_t denotes the t-th generated token.

The model outputs either a time interval within the `<answer></answer>` tags, or an evidence-absence signal (*i.e.*, "not found") when the queried event is absent. A deterministic parsing function $g(\cdot)$ maps the output to:

$$\hat{\tau} = g(y) \in \mathbb{R}^2 \cup \{\emptyset\}, \tag{2}$$

where $\hat{\tau} = (\hat{s}, \hat{e})$ denotes predicted start and end times, and $\emptyset$ denotes evidence absence. The absence decision is triggered only within `<answer>` to avoid misclassifying negative sentences in the reasoning chain.

Group Relative Policy Optimization (GRPO). To efficiently calibrate model behavior in the TTA setting, we introduce GRPO [14, 34], which estimates the advantage baseline by sampling a group of outputs $\{o_1, o_2, ..., o_G\}$ for the same input x, avoiding the need for a separate value critic network. For each input x, the policy π_θ generates G outputs scored by the reward $\mathcal{R}(o, x)$. The relative advantage $\hat{A}_i$ of each output is:

$$\hat{A}_i = \frac{r_i - \text{mean}(r_1, ..., r_G)}{\text{std}(r_1, ..., r_G) + \epsilon}. \tag{3}$$

The core objective function of GRPO is as follows:

$$\mathcal{J}(\theta) = \mathbb{E}_{x \sim \mathcal{S}} \left[\frac{1}{G} \sum_{i=1}^{G} \left(\frac{\pi_\theta(o_i|x)}{\pi_{\theta_{old}}(o_i|x)} \hat{A}_i - \beta \cdot D_{KL}(\pi_\theta || \pi_{ref}) \right) \right], \tag{4}$$

where $\frac{\pi_\theta(o_i|x)}{\pi_{\theta_{old}}(o_i|x)}$ is the importance sampling ratio, D_{KL} is used to constrain the magnitude of policy updates, and β is the KL penalty coefficient.

Pipeline. At test time, we process a continuous data stream $\mathcal{S} = \{x_1, x_2, \ldots, x_T\}$. At each step t, given the real sample $x_t^{pos} = (V_{\text{ori}}, Q_{\text{ori}})_t$, we construct a paired counterfactual sample $x_t^{cf} = (V_{\text{cf}}, Q_{\text{cf}})_t$ in a self-supervised manner (see Sec. 3.2), where V_{cf} or Q_{cf} is altered to ensure the target event is absent. Using the frozen reference policy π_{ref} as an anchor on the mixed stream $\{x_t^{pos}, x_t^{cf}\}$, we perform online continual updates via GRPO at each step. We then evaluate the final adapted model on test data that include queries for non-existent events.

4.2 Verifiable Rewards

Since ground-truth labels are unavailable at test time, we design a rule-based verifiable reward mechanism that applies different guidance signals according to the sample type (positive or counterfactual). The total reward r consists of three components: the consistency anchor reward R_{ca}, the hallucination repulsion reward R_{hr}, and the format reward R_{fmt}.

Consistency Anchor Reward. To prevent catastrophic forgetting during adaptation, we anchor the model's behavior on positive samples using the frozen reference policy π_{ref} (*i.e.*, the initial only-positive grounding model). For a positive sample x^{pos}, π_{ref} greedily decodes a deterministic reference output y_{ref}. We then compare the current policy's prediction y with y_{ref} via interval IoU, and assign a positive reward if the overlap exceeds threshold δ. The definition is:

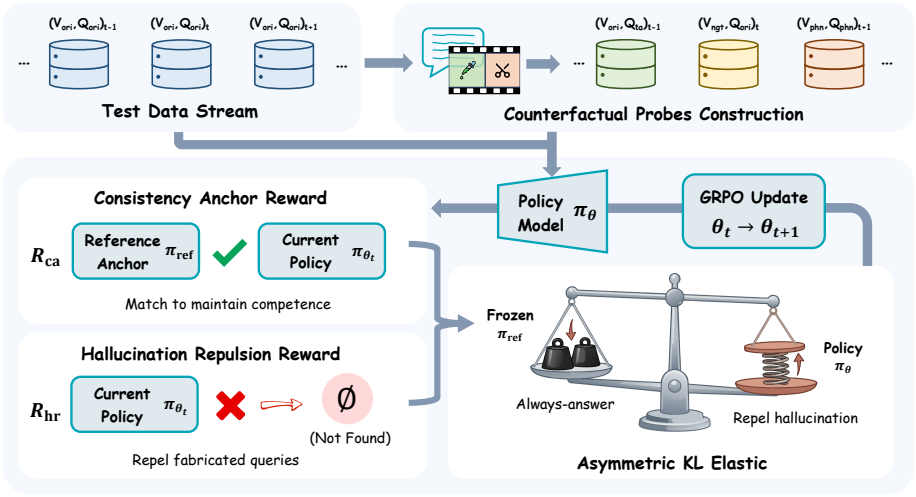

Fig. 3: Overview of HRVTG. At test time, genuine samples are paired with self-supervised counterfactual probes and fed to the policy model π_θ. Online GRPO updates are guided by a consistency anchor reward R_{ca} that preserves grounding competence against a frozen reference π_{ref}, and a hallucination repulsion reward R_{hr} that encourages evidence-absence signals for fabricated queries. An asymmetric KL elastic constraint balances stability on genuine samples with plasticity on counterfactual ones.

$$\text{IoU}(y, y_{\text{ref}}) = \frac{\max\left(0, \min(\hat{e}, \hat{e}_{\text{ref}}) - \max(\hat{s}, \hat{s}_{\text{ref}})\right)}{\max(\hat{e}, \hat{e}_{\text{ref}}) - \min(\hat{s}, \hat{s}_{\text{ref}})}, \tag{5}$$

$$R_{\text{ca}}(y, y_{\text{ref}}) = \begin{cases} r_{\text{ca}}, & \text{IoU}(y, y_{\text{ref}}) \geq \delta \ \wedge \ \hat{\tau} \neq \emptyset \ \wedge \ \hat{\tau}^{\text{ref}} \neq \emptyset, \\ 0, & \text{otherwise}, \end{cases} \tag{6}$$

where δ is the IoU threshold and $r_{\text{ca}} > 0$. If either side rejects or the interval cannot be parsed, the reward is 0. We do not apply an explicit negative reward, to avoid overly aggressive penalties on positive samples during the online stage that could degrade grounding ability.

Hallucination Repulsion Reward. For counterfactual samples x^{cf} constructed via text or visual tampering, they naturally come with a pseudo-label indicating non-existence. Therefore, the reward function directly drives the model to output a specific rejection keyword (*e.g.*, "not found"). The definition is:

$$R_{\text{hr}}(y) = \begin{cases} r_{\text{hr}}, & g(y) = \emptyset, \\ 0, & \text{otherwise}, \end{cases} \tag{7}$$

where $r_{\text{hr}} > 0$ is a constant. Note that we do not impose an explicit negative reward for "outputting an interval". Instead, we rely on within-group advantage normalization to make candidates with correct rejection relatively better,

thereby inducing an implicit contrastive learning effect and reducing the instability caused by overly strong penalties.

Format Reward. To ensure that the generative VTG output is structured and parsable, we add a lightweight format reward that encourages the output to strictly contain the `<answer>` tags:

$$R_{\text{fmt}}(y) = \mathbb{I}\left(y \text{ matches the required answer tag pattern}\right). \tag{8}$$

In summary, the final reward for each sample is piecewise defined as:

$$r = \begin{cases} R_{\text{ca}}(y, y_{\text{ref}}) + \alpha_{\text{fmt}} R_{\text{fmt}}(y), & x^{pos} \in \mathcal{X}^{pos}, \\ R_{\text{hr}}(y), & x^{cf} \in \mathcal{X}^{cf}, \end{cases} \tag{9}$$

where α_{fmt} is the format reward weight (set to the order of 0.2 in implementation). The format reward is only applied to positive samples to avoid encouraging tag-only outputs when evidence is absent.

4.3 Asymmetric KL Elastic Constraint

Standard GRPO uses a single KL penalty coefficient β for all samples. In our mixed data stream, this is suboptimal: for positive samples x^{pos}, π_θ should stay close to π_{ref} to preserve grounding; for counterfactual samples x^{cf}, π_{ref} often hallucinates spurious intervals, so π_θ should deviate and output an evidence-absence signal. We therefore adopt sample-type-aware asymmetric KL weighting:

$$\lambda(x) = \begin{cases} \rho, & x \in \mathcal{X}^{pos}, \\ 1, & x \in \mathcal{X}^{cf}, \end{cases} \tag{10}$$

where $\lambda(x)$ is used as a multiplicative factor on the base KL coefficient β (*i.e.*, the effective KL weight is $\beta \cdot \lambda(x)$), and $\rho > 1$ thus enforces stronger KL regularization on positive samples while allowing greater policy deviation on counterfactual samples to facilitate hallucination resistance.

5 Experiment

5.1 Experiment Setup

Benchmarks. To evaluate HRVTG on counterfactual hallucinations, we require samples in which the queried event is absent from the video. RaTSG [10] reconstructs two standard benchmarks, Charades-STA [11] and ActivityNet Captions [16], by pairing each genuine sample with a "no-answer" counterpart (1:1 positive-to-negative ratio), yielding Charades-RF and ActivityNet-RF with test sets of 3,720 and 34,062 samples, respectively. To further evaluate stability across all three hallucination types, we apply the probe construction method from Sec. 3.2 to TVGBench [47], constructing TVGBench-RF, which covers text-driven, vision-driven, and multimodal joint hallucinations. We split TVGBench into a TTA set and an evaluation set of 400 samples each.

Evaluation Metrics. Previous work based on offline-learned rejection [10, 17, 45] universally suffers from metric coupling. When the test set contains many irrelevant or fabricated queries, a model can inflate its overall score by being overly conservative (*i.e.*, triggering abstention too frequently), masking degraded grounding on truly relevant queries. To more accurately assess the ability to "resist hallucinations without sacrificing grounding", we redesign the evaluation metrics to decouple boundary calibration from temporal grounding performance:

- **(i) Conditional IoU ($\mathbf{mIoU_{Cond}}$).** In previous metric systems, whenever a model outputs an abstention decision for a negative sample (fabricated query), its IoU is defaulted to 1. This inflates the overall mIoU and makes it hard to tell whether a model is imprecise in grounding or simply grounding the wrong target. We therefore compute mIoU only over True Positives (TP), where both prediction and ground truth are positive, so the score reflects grounding precision on real events.
- **(ii) Balanced Decision F1-score ($\mathbf{F1_{Balanced}}$).** Prior work often relies on accuracy or a single-class F1 on fabricated samples to assess rejection, which overlooks the severe class imbalance between positive and negative queries in open-world settings. Following open-set recognition practice, we treat answer/refuse as a binary classification task and take the harmonic mean of F1-scores for both classes. This penalizes extreme strategies (always refuse or always answer) and better reflects decision-boundary health.

$$\mathrm{mIoU}_{\mathrm{Cond}} = \frac{1}{|TP|} \sum_{i \in TP} \mathrm{IoU}(\hat{\tau}_i, \tau_i), \tag{11}$$

$$\mathrm{F1}_{\mathrm{Balanced}} = \frac{2 \times \mathrm{F1}_{\mathrm{Refuse}} \times \mathrm{F1}_{\mathrm{Answer}}}{\mathrm{F1}_{\mathrm{Refuse}} + \mathrm{F1}_{\mathrm{Answer}}}. \tag{12}$$

Implementation Details. We adopt Qwen2.5-VL-Instruct at 3B and 7B scales as the base MLLM backbone. During test-time adaptation, we attach a LoRA adapter to the frozen backbone and perform online GRPO updates. For the 3B model, the learning rates are set to 5×10^{-6} on Charades-RF, 1×10^{-6} on ActivityNet-RF, and 5×10^{-5} on TVGBench-RF; for 7B model, they are 2×10^{-6}, 1×10^{-6}, and 2×10^{-5}, respectively. The IoU threshold is set to $\delta = 0.3$ on Charades-RF, and $\delta = 0.2$ on ActivityNet-RF and TVGBench-RF. Across all datasets, we fix $r_{\mathrm{ca}} = 0.8$, $r_{\mathrm{hr}} = 0.2$, $\alpha_{\mathrm{fmt}} = 0.2$, and $\rho = 4$.

5.2 Main Results

To evaluate our proposed HRVTG framework, we compare it with two baseline models on the Charades-RF, ActivityNet-RF, and TVGBench-RF datasets. The Base Model refers to Qwen2.5-VL without downstream VTG fine-tuning. The Only-positive Model refers to the one fine-tuned exclusively on genuine, positive VTG samples, representing the traditional always-answer paradigm. The results are summarized in Tab. 1, Tab. 2, and Tab. 3.

Table 1: Model Evaluation Results on Charades-RF based on Qwen2.5-VL.

Model	Size	Grounding				Hallucination		
		R@0.3	R@0.5	R@0.7	mIoU	F1_Ans	F1_Rfs	F1_Bal
Base Model	3B	50.36	25.90	13.67	30.27	13.64	67.42	22.69
Only-positive Model	3B	67.60	38.53	14.90	41.48	69.50	33.38	45.12
Our Model	3B	71.24	40.49	15.58	43.29	74.86	62.80	68.30
Base Model	7B	29.02	13.20	5.88	19.50	55.31	73.56	63.14
Only-positive Model	7B	72.59	53.43	25.55	48.33	81.36	75.40	78.27
Our Model	7B	73.97	54.07	27.01	48.94	83.00	80.91	81.94

Table 2: Model Evaluation Results on ActivityNet-RF based on Qwen2.5-VL.

Model	Size	Grounding				Hallucination		
		R@0.3	R@0.5	R@0.7	mIoU	F1_Ans	F1_Rfs	F1_Bal
Base Model	3B	17.70	12.60	5.41	17.70	52.31	73.17	61.00
Only-positive Model	3B	38.46	21.95	10.19	26.61	73.90	51.22	60.51
Our Model	3B	40.79	23.00	10.89	28.03	83.81	84.01	83.91
Base Model	7B	28.33	17.36	9.25	21.34	46.51	73.15	56.87
Only-positive Model	7B	44.15	27.22	14.04	30.85	85.15	85.43	85.29
Our Model	7B	47.50	28.03	13.03	31.87	89.37	87.71	88.53

The core challenge of adding hallucination resistance is potential degradation of real-event localization. On Charades-RF and ActivityNet-RF, HRVTG preserves and slightly improves grounding. According to our decoupled metric $\text{mIoU}_{\text{Cond}}$ (denoted as mIoU), the 3B model rises from 41.48 to 43.29 and from 26.61 to 28.03, respectively. This indicates that the consistency anchor reward retains grounding ability during test-time adaptation, preventing catastrophic forgetting. More importantly, gains on $\text{F1}_{\text{Balanced}}$ (denoted as F1_Bal) are much larger. Compared with the Only-positive Model, which shows severe forced-alignment bias and keeps producing evidence for absent events, HRVTG lifts 3B's F1_Bal from 45.12 to 68.30 on Charades-RF and from 60.51 to 83.91 on ActivityNet-RF. These results show that counterfactual repulsion signals teach the model to output null-attribution states for fabricated queries. TVGBench-RF is more challenging than the above two benchmarks because it contains deceptive text-driven, vision-driven, and multimodal joint fabrications. The very low 3B Only-positive score reflects how limited-reasoning models fail under such complex counterfactual triggers. In contrast, HRVTG increases F1_Bal to 58.83 while improving $\text{mIoU}_{\text{Cond}}$ from 23.34 to 26.09. This demonstrates that our three targeted counterfactual probes effectively simulate these extreme cases and provide strong robustness against complex multimodal misattributions.

5.3 Ablation Studies

We analyze three critical hyperparameters of our framework on the TVGBench-RF. All ablation experiments are conducted with Qwen2.5-VL-3B.

Table 3: Model Evaluation Results on TVGBench-RF based on Qwen2.5-VL.

Model	Size	Grounding				Hallucination		
		R@0.3	R@0.5	R@0.7	mIoU	F1_Ans	F1_Rfs	F1_Bal
Base Model	3B	26.76	14.08	9.86	18.30	49.05	58.86	53.51
Only-positive Model	3B	32.59	18.35	6.33	23.34	66.18	5.28	9.78
Our Model	3B	40.26	18.53	7.99	26.09	66.67	52.65	58.83
Base Model	7B	32.58	19.10	7.87	22.64	33.78	67.47	45.02
Only-positive Model	7B	36.70	24.47	10.37	26.31	70.74	42.09	52.78
Our Model	7B	37.65	24.10	12.35	26.93	72.17	62.35	66.90

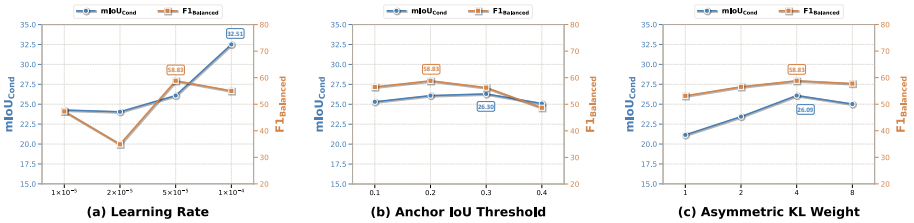

Fig. 4: Ablation study of key hyperparameters.

Learning Rate: Fig. 4(a) shows a nonlinear trade-off between conditional grounding ($\text{mIoU}_{\text{Cond}}$) and decision-boundary quality ($\text{F1}_{\text{Balanced}}$) as the TTA learning rate increases. The best $\text{F1}_{\text{Balanced}}$ (58.83) is achieved at 5×10^{-5}, while the highest $\text{mIoU}_{\text{Cond}}$ appears at a larger learning rate, likely due to survivor bias. With an overly large learning rate, the model becomes too conservative, answers only easy events, and rejects many valid ones, which collapses the decision boundary. By contrast, the best-$\text{F1}_{\text{Balanced}}$ setting better balances hallucination resistance and localization on hard, ambiguous cases. This mismatch suggests that evaluating open-world grounding only with temporal alignment metrics, without decision-boundary quality, can be misleading.

Anchor IoU Threshold (δ): determines the difficulty of obtaining anchor rewards during test-time adaptation. If the threshold is too low, the anchor condition becomes overly permissive and fails to distinguish high-quality evidence, causing drift in $\text{mIoU}_{\text{Cond}}$. If it is too high, the model struggles to receive positive rewards and tends to optimize for blanket rejection.

Asymmetric KL Weight (ρ): regulates the relative penalty between positive and counterfactual samples. As shown in Fig. 4(c), $\rho = 1$ (*i.e.*, standard symmetric GRPO) yields the lowest scores (21.15 and 53.16), proving that uniform constraints fail to decouple grounding from rejection. Increasing ρ to 4 yields the best results on both metrics, validating that strict regularization on genuine queries preserves grounding while flexible constraints on fabricated probes enhance selective prediction. When $\rho = 8$, both metrics drop slightly, suggesting that overly rigid anchoring on positive samples limits adaptation plasticity.

6 Conclusion

In this paper, we address the always-answer bias in video temporal grounding, which induces systematic temporal hallucinations in open-world deployments. We introduce a taxonomy of VTG hallucinations and construct self-supervised counterfactual probes via textual alteration, visual pruning, and hybrid fabrication. Building on these probes, we propose HRVTG, a test-time adaptation framework that dynamically calibrates the decision boundary via GRPO, balancing consistency anchor on genuine queries with counterfactual repulsion signals on fabricated ones. Evaluated through novel decoupled metrics, HRVTG significantly enhances hallucination resistance while preserving real-event grounding performance, paving the way for more reliable multimodal video understanding.

Acknowledgements

The work is supported by the National Natural Science Foundation of China (No. 62276134, U25A20442, 62427808, 62506167, and 62476124).

References

1. Bao, W., Yu, Q., Kong, Y.: Evidential deep learning for open set action recognition. In: ICCV. pp. 13349–13358 (2021)
2. Cao, Z., Zhang, B., Du, H., Yu, X., Li, X., Wang, S.: Flashvtg: Feature layering and adaptive score handling network for video temporal grounding. In: WACV. pp. 9226–9236. IEEE (2025)
3. Ceylan, D., Huang, C.H.P., Mitra, N.J.: Pix2video: Video editing using image diffusion. In: ICCV. pp. 23206–23217 (October 2023)
4. Chai, W., Guo, X., Wang, G., Lu, Y.: Stablevideo: Text-driven consistency-aware diffusion video editing. In: ICCV. pp. 23040–23050 (2023)
5. Chai, W., Guo, X., Wang, G., Lu, Y.: Stablevideo: Text-driven consistency-aware diffusion video editing. In: ICCV. pp. 23040–23050 (October 2023)
6. Chen, C.C., Lo, L.W., Lin, S.J.: Cohets: A highlight extraction method using textual streams of streaming videos. Knowledge-Based Systems **258**, 110000 (2022)
7. Chen, T., Shorinwa, O., Bruno, J., Swann, A., Yu, J., Zeng, W., Nagami, K., Dames, P., Schwager, M.: Splat-nav: Safe real-time robot navigation in gaussian splatting maps. IEEE Transactions on Robotics **41**, 2765–2784 (2025)
8. Choi, J., Gao, C., Messou, J.C., Huang, J.B.: Why can't i dance in the mall? learning to mitigate scene bias in action recognition. In: NeurIPS. vol. 32 (2019)
9. Dong, J., Chen, X., Zhang, M., Yang, X., Chen, S., Li, X., Wang, X.: Partially relevant video retrieval. In: ACM MM. pp. 246–257 (2022)
10. Dong, J., Peng, X., Liu, D., Qu, X., Yang, X., Bao, C., Wang, M.: Temporal sentence grounding with relevance feedback in videos. In: NeurIPS (2024)
11. Gao, J., Sun, C., Yang, Z., Nevatia, R.: Tall: Temporal activity localization via language query. In: ICCV. pp. 5267–5275 (2017)
12. Ge, W., Huang, P., Yan, R., Qu, H., Xie, G., Shu, X.: Reliable and diverse hierarchical adapter for zero-shot video classification. In: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence. pp. 1026–1034 (2025)

13. Ge, W., Qu, H., Yan, R., Xie, G.S., Yao, Y., Shu, X., Tang, J.: Condensed test-time adaptation of vlms for action recognition. In: CVPR. pp. 16977–16987 (June 2026)
14. Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
15. Huang, B., Wang, X., Chen, H., Song, Z., Zhu, W.: Vtimellm: Empower llm to grasp video moments. In: CVPR. pp. 14271–14280 (2024)
16. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Carlos Niebles, J.: Dense-captioning events in videos. In: ICCV. pp. 706–715 (2017)
17. Lee, J.S., Lee, S., Jung, S., Li, B., Lee, J.H.: Learning to refuse: Refusal-aware reinforcement fine-tuning for hard-irrelevant queries in video temporal grounding. arXiv preprint arXiv:2511.23151 (2025)
18. Lei, J., Berg, T.L., Bansal, M.: Detecting moments and highlights in videos via natural language queries. In: NeurIPS. vol. 34, pp. 11846–11858 (2021)
19. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML. pp. 19730–19742. PMLR (2023)
20. Li, W., Chu, G., Chen, J., Xie, G.S., Shan, C., Zhao, F.: Lad-reasoner: Tiny multimodal models are good reasoners for logical anomaly detection. arXiv preprint arXiv:2504.12749 (2025)
21. Li, X., Yan, Z., Meng, D., Dong, L., Zeng, X., He, Y., Wang, Y., Qiao, Y., Wang, Y., Wang, L.: Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. arXiv preprint arXiv:2504.06958 (2025)
22. Lin, K.Q., Zhang, P., Chen, J., Pramanick, S., Gao, D., Wang, A.J., Yan, R., Shou, M.Z.: Univtg: Towards unified video-language temporal grounding. In: ICCV. pp. 2794–2804 (2023)
23. Lin, K., Chen, P., Huang, D., Li, T.H., Tan, M., Gan, C.: Learning vision-and-language navigation from youtube videos. In: ICCV. pp. 8317–8326 (2023)
24. Liu, J., Yang, L., Luo, H., Wang, F., Li, H., Wang, M.: Preacher: Paper-to-video agentic system. In: ICCV. pp. 17129–17139 (October 2025)
25. Liu, Y., He, J., Li, W., Kim, J., Wei, D., Pfister, H., Chen, C.W.: R 2-tuning: Efficient image-to-video transfer learning for video temporal grounding. In: European conference on computer vision. pp. 421–438. Springer (2024)
26. Liu, Y., Li, S., Wu, Y., Chen, C.W., Shan, Y., Qie, X.: Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In: CVPR. pp. 3042–3051 (2022)
27. Lyu, Z., Zhang, Y.: A novel temporal moment retrieval model for apron surveillance video. Computers and Electrical Engineering **107**, 108616 (2023)
28. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. In: NeurIPS. vol. 35, pp. 27730–27744 (2022)
29. Pramanick, S., Mavroudi, E., Song, Y., Chellappa, R., Torresani, L., Afouras, T.: Enrich and detect: Video temporal grounding with multimodal llms. In: ICCV. pp. 24297–24308 (2025)
30. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PmLR (2021)
31. Rawal, R., Shirkavand, R., Huang, H., Somepalli, G., Goldstein, T.: Argus: Hallucination and omission evaluation in video-llms. In: ICCV. pp. 20280–20290 (2025)

32. Ren, S., Yao, L., Li, S., Sun, X., Hou, L.: Timechat: A time-sensitive multimodal large language model for long video understanding. In: CVPR. pp. 14313–14323 (2024)
33. Sahoo, P., Meharia, P., Ghosh, A., Saha, S., Jain, V., Chadha, A.: A comprehensive survey of hallucination in large language, image, video and audio foundation models. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 11709–11724 (2024)
34. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
35. Shi, Y., Xue, C., Liew, J.H., Pan, J., Yan, H., Zhang, W., Tan, V.Y.F., Bai, S.: Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. In: CVPR. pp. 8839–8849 (June 2024)
36. Shin, I., Tsai, Y.H., Zhuang, B., Schulter, S., Liu, B., Garg, S., Kweon, I.S., Yoon, K.J.: Mm-tta: multi-modal test-time adaptation for 3d semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16928–16937 (2022)
37. Shu, M., Nie, W., Huang, D.A., Yu, Z., Goldstein, T., Anandkumar, A., Xiao, C.: Test-time prompt tuning for zero-shot generalization in vision-language models. In: NeurIPS. vol. 35, pp. 14274–14289 (2022)
38. Sul, J., Han, J., Lee, J.: Mr. hisum: A large-scale dataset for video highlight detection and summarization. In: NeurIPS. vol. 36, pp. 40542–40555 (2023)
39. Tian, M., Li, G., Qi, Y., Wang, S., Sheng, Q.Z., Huang, Q.: Rethink video retrieval representation for video captioning. Pattern Recognition **156**, 110744 (2024)
40. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
41. Ullah, F.U.M., Obaidat, M.S., Ullah, A., Muhammad, K., Hijji, M., Baik, S.W.: A comprehensive review on vision-based violence detection in surveillance videos. ACM Computing Surveys **55**(10), 1–44 (2023)
42. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. In: ICLR (2021)
43. Wang, H., Xu, Z., Cheng, Y., Diao, S., Zhou, Y., Cao, Y., Wang, Q., Ge, W., Huang, L.: Grounded-videollm: Sharpening fine-grained temporal grounding in video large language models. arXiv preprint arXiv:2410.03290 (2024)
44. Wang, Q., Fink, O., Van Gool, L., Dai, D.: Continual test-time domain adaptation. In: CVPR. pp. 7201–7211 (2022)
45. Wang, Q., Chen, S., Shen, Y.: Causalvtg: Towards robust video temporal grounding via causal inference. In: NeurIPS (2025)
46. Wang, X., Zhang, Y., Zohar, O., Yeung-Levy, S.: Videoagent: Long-form video understanding with large language model as agent. In: European Conference on Computer Vision. pp. 58–76. Springer (2024)
47. Wang, Y., Wang, Z., Xu, B., Du, Y., Lin, K., Xiao, Z., Yue, Z., Ju, J., Zhang, L., Yang, D., et al.: Time-r1: Post-training large vision language model for temporal video grounding. arXiv preprint arXiv:2503.13377 (2025)
48. Wang, Y., Xu, B., Yue, Z., Xiao, Z., Wang, Z., Zhang, L., Yang, D., Wang, W., Jin, Q.: Timezero: Temporal video grounding with reasoning-guided lvlm. arXiv e-prints pp. arXiv–2503 (2025)
49. Wang, Y., Zhang, Z., Wang, R.: Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 8640–8665 (2023)

50. Wang, Z., Wang, L., Wu, T., Li, T., Wu, G.: Negative sample matters: A renaissance of metric learning for temporal grounding. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 2613–2623 (2022)
51. Wei, F., Wang, B., Ge, T., Jiang, Y., Li, W., Duan, L.: Learning pixel-level distinctions for video highlight detection. In: WACV. pp. 3073–3082 (2022)
52. Wei, L., Li, Y., Wang, C., Wang, Y., Kong, L., Huang, W., Sun, L.: Unsupervised post-training for multi-modal llm reasoning via grpo. arXiv preprint arXiv:2505.22453 (2025)
53. Wu, W., Luo, H., Fang, B., Wang, J., Ouyang, W.: Cap4video: What can auxiliary captions do for text-video retrieval? In: CVPR. pp. 10704–10713 (2023)
54. Xu, X., Wang, H., Xie, G.S., Shan, C., Zhao, F.: Hitea: Hierarchical temporal alignment for training-free long-video temporal grounding. In: ICLR (2026)
55. Yan, R., Qu, H., Shu, X., Li, W., Tang, J., Tan, T.: Dts-tpt: dual temporal-sync test-time prompt tuning for zero-shot activity recognition. In: IJCAI. pp. 1534–1542 (2024)
56. Yan, R., Wang, J., Qu, H., Du, X., Zhang, D., Tang, J., Tan, T.: Test-v: Test-time support-set tuning for zero-shot video classification. In: Kwok, J. (ed.) IJCAI. pp. 2143–2151. International Joint Conferences on Artificial Intelligence Organization (8 2025), main Track
57. Zhang, D., Dai, X., Wang, X., Wang, Y.F., Davis, L.S.: Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In: CVPR. pp. 1247–1257 (2019)
58. Zhang, J., Guo, Y., Potamias, R.A., Deng, J., Xu, H., Ma, C.: Vtimecot: Thinking by drawing for video temporal grounding and reasoning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 24203–24213 (2025)
59. Zhang, L., Zhao, T., Ying, H., Ma, Y., Lee, K.: Omagent: A multi-modal agent framework for complex video understanding with task divide-and-conquer. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 10031–10045 (2024)
60. Zhang, M., Levine, S., Finn, C.: Memo: Test time robustness via adaptation and augmentation. In: NeurIPS. vol. 35, pp. 38629–38642 (2022)
61. Zhang, S., Peng, H., Fu, J., Luo, J.: Learning 2d temporal adjacent networks for moment localization with natural language. In: AAAI. pp. 12870–12877 (2020)
62. Zhang, X.Y., Xie, G.S., Li, X., Mei, T., Liu, C.L.: A survey on learning to reject. Proceedings of the IEEE **111**(2), 185–215 (2023)
63. Zhang, Z., Tao, F., Xiang, T., Zhao, F., Xie, G.S.: Graph interaction prompt network for few-shot medical image anomaly detection. In: 2025 IEEE International Conference on Bioinformatics and Biomedicine. pp. 1435–1442. IEEE (2025)
64. Zheng, M., Cai, X., Chen, Q., Peng, Y., Liu, Y.: Training-free video temporal grounding using large-scale pre-trained models. In: European Conference on Computer Vision. pp. 20–37. Springer (2024)
65. Zhu, B., Fang, H., Sui, Y., Li, L.: Deepfakes for medical video de-identification: Privacy protection and diagnostic information preservation. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. pp. 414–420 (2020)
66. Zuo, Y., Zhang, K., Sheng, L., Qu, S., Cui, G., Zhu, X., Li, H., Zhang, Y., Long, X., Hua, E., et al.: Ttrl: Test-time reinforcement learning. arXiv preprint arXiv:2504.16084 (2025)