Uncertainty-quantified Rollout Policy Adaptation for Unlabelled Cross-domain Temporal Grounding

Jian Hu[1], Zixu Cheng[1] Shaogang Gong[1], Isabel Guan[2], Jianye Hao[3] , Jun Wang[4] , Kun Shao[3]*
[1]Queen Mary University of London, [2]Hong Kong University of Science and Technology,
[3]Huawei Noah's Ark Lab, [4]University College London

Abstract

Video Temporal Grounding (TG) aims to temporally locate video segments matching a natural language description (a query) in a long video. While Vision-Language Models (VLMs) are effective at holistic semantic matching, they often struggle with fine-grained temporal localisation. Recently, Group Relative Policy Optimisation (GRPO) reformulates the inference process as a reinforcement learning task, enabling fine-grained grounding and achieving strong in-domain performance. However, GRPO relies on labelled data, making it unsuitable in unlabelled domains. Moreover, because videos are large and expensive to store and process, performing full-scale adaptation introduces prohibitive latency and computational overhead, making it impractical for real-time deployment. To overcome both problems, we introduce a Data-Efficient Unlabelled Cross-domain Temporal Grounding method, from which a model is first trained on a labelled source domain, then adapted to a target domain using only a small number of *unlabelled videos from the target domain*. This approach eliminates the need for target annotation and keeps both computational and storage overhead low enough to run in real time. Specifically, we introduce **U**ncertainty-quantified **R**ollout **P**olicy **A**daptation (**URPA**) for cross-domain knowledge transfer in learning video temporal grounding without target labels. URPA generates multiple candidate predictions using GRPO rollouts, averages them to form a pseudo label, and estimates confidence from the variance across these rollouts. This confidence then weights the training rewards, guiding the model to focus on reliable supervision. Experiments on three datasets across six cross-domain settings show that URPA generalises well using only a few unlabelled target videos. Codes will be released once published.

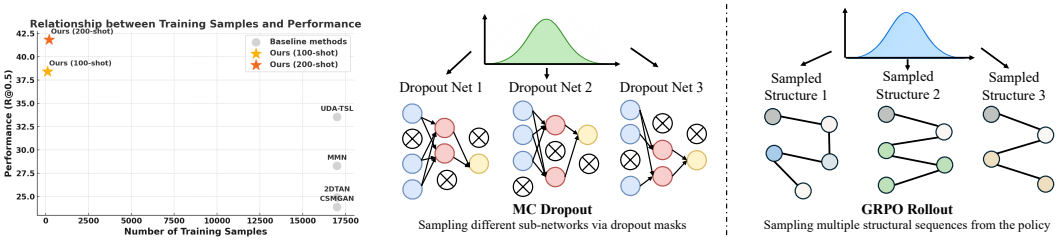
1 Introduction

Temporal Grounding (TG) localises the exact temporal segment in an untrimmed video that semantically corresponds to a given natural-language query [1, 12]. Accurate TG is fundamental to high-level applications such as activity detection [6] and embodied human–computer interaction [7].

While VLMs are effective at capturing holistic video semantics [22, 41, 18], they often struggle with fine-grained localisation, leading to suboptimal performance in temporal grounding. Some recent works apply Supervised Fine-Tuning (SFT) to better align video-query pairs [4, 39, 58]. However, since relevant segments typically cover only a small portion of the video, the model often overfits to redundant context, limiting its ability to perform precise grounding [60]. To improve temporal reasoning, Chain-of-Thought (CoT) post-training introduces explicit intermediate reasoning steps before prediction [37, 50]. While effective, this approach depends on manually annotated prompts, which are expensive to collect and hard to scale across domains and tasks. More recently, GRPO [14]

*Corresponding author: shaokun2@huawei.com

Preprint. Under review.



(a) Motivation of data-efficient CTG setting. (b) Comparison between MC Dropout and GRPO rollout.

Figure 1: (a) A comparison between full-data and data-efficient adaptation in Cross-domain Temporal Grounding (CTG). Existing methods adapt on thousands of unlabelled target videos (grey), which is slow and resource-heavy. We propose a data-efficient CTG setting using only 100 or 200 randomly selected target videos. Despite the limited data, our method matches or exceeds performance on the TaCoS $\rightarrow$ ActivityNet task. (b) Conceptual comparison between MC Dropout and GRPO rollout. MC Dropout samples subnetworks via stochastic neuron dropout and estimates uncertainty from output diversity. GRPO rollouts similarly sample diverse structural sequences from the policy. URPA leverages this property to generate averaged pseudo labels and estimate uncertainty via rollout standard deviation, enabling uncertainty-quantified adaptation without ground-truth labels.

formulates temporal grounding as a policy learning problem, where reinforcement learning is used to directly optimise segment predictions [47, 26]. This removes the need for hand-crafted prompts used in post-processing and achieves strong results on in-domain benchmarks. However, GRPO relies on ground-truth temporal boundaries to compute reward signals. requiring extensively labelled training videos. It limits its practical usefulness in real-world scenarios where annotations are unavailable. Moreover, GRPO-based temporal grounding performance drops significantly across domains due to distribution shift, revealing its poor generalisation to unseen data. In addition to annotation constraints, the sheer scale of video data poses practical challenges. A target domain often contains thousands of videos, which are expensive to store, and time-consuming to adapt. As shown in Fig. 1(a), existing adaptation methods require full retraining on the entire target set. Such latency and resource-hungry makes them unsuitable for time-sensitive applications like online or on-device deployment. We consider a more practical and generalisable approach to cross-domain temporal grounding where a model trained on a labelled source domain can be efficiently adapted to an unlabelled target domain using only sparse unlabelled target videos, with minimal latency and resource demands.

To address these challenges, we introduce a Data-Efficient Unlabelled Cross-domain Temporal Grounding approach. A model trained on a labelled source domain is adapted at test-time in deployment using only K unlabelled videos from the target domain ($K = 100 \sim 200$ in our experiments), allowing for real-time adaptation on limited compute resources. The main challenge in this approach lies in effectively leveraging limited unlabelled target data. Without annotated temporal boundaries, pseudo-labels are noisy and highly uncertain, resulting in poor cross-domain performance. To mitigate this, we introduce explicit uncertainty quantification to assess the reliability of pseudo-labels. As illustrated in Fig.1(b), MC Dropout [11] estimates uncertainty by sampling multiple sub-networks via stochastic dropout and computing the variance across their outputs. Similarly, GRPO rollouts produce diverse predictions by sampling from a stochastic policy. Inspired by this parallel, we estimate pseudo label confidence by measuring the standard deviation across multiple rollouts per <video, query> pair, and use it to guide adaptation with uncertainty-quantification weighted model adaptation.

To this end, we propose Uncertainty-quantified Rollout Policy Adaptation (URPA). After supervised training of a GRPO backbone on the source domain, we perform test-time adaptation by generating G stochastic rollouts for each of the K (K=100/200 in our experiments) randomly selected unlabelled videos in the target domain. A pseudo label is constructed by averaging the predicted boundaries across rollouts, and a small margin relaxation is applied to reduce the impact of outliers. To estimate the reliability of each pseudo label, we follow the Bayesian view of MC Dropout and use the variance across rollouts as an estimate of uncertainty. Samples with lower variance, indicating higher confidence, are given larger weights during a lightweight gradient update. This enables fast and effective adaptation without introducing significant latency. Sec. 4 provides a theoretical analysis of how rollout-based variance quantifies uncertainty.

Our contributions are threefold: (i) We introduce an unlabelled cross-domain temporal grounding approach to test-time model adaptation using only a small number of unlabelled target-domain videos. (ii) We propose Uncertainty-quantified Rollout Policy Adaptation (URPA), the first rein-

forcement learning-based self-learning adaptation method that combines pseudo-label generation with uncertainty-weighted rewards. URPA enables effective GRPO-based temporal grounding across domains *without* requiring ground-truth labels to compute reinforcement signals. (iii) We theoretically show that rollout variance approximates Bayesian predictive variance, quantifying epistemic uncertainty, and empirically demonstrate that URPA consistently outperforms strong baselines across six cross-domain video temporal grounding benchmarks.

2 Related Works

Test-time Adaptation. Test-time domain adaptation (TTDA) adapts a pre-trained model to unlabelled test data exhibiting distribution shift, prior to prediction [42, 19, 35]. TTDA approaches can be broadly divided into backward-free and backward-based methods. Backward-free methods adapt the model on-the-fly during inference, typically by updating batch normalization (BN) statistics without backpropagation. Representative works include DUA [32], which applies a running average to BN layers, and DIGA [45], which aligns distributions for semantic segmentation. While efficient per sample, such methods require adaptation for every test input, which may increase inference latency and introduce instability across samples. In contrast, backward-based methods adapt the model once using target data before inference begins, enabling faster and more stable test-time prediction. These methods often rely on self-supervised objectives like entropy minimization [42, 20]. However, video data is large and difficult to store, and performing test-time adaptation on all videos is time-consuming. Hence, we leverage only a small number of videos from the target domain to perform data-efficient test-time adaptation via backpropagation, and then apply the adapted model to evaluate the full set test video without further updates. This design ensures both adaptability and real-time efficiency, making it well-suited for practical scenarios with limited target domain data.

Temporal Grounding. Temporal grounding [12, 24] is a vision–language task that aims to localise the snippet that corresponds to a given natural language query with start and end timestamps in a video. Existing methods fall into two main categories: proposal-based approaches [12, 1, 51, 61, 48, 17], which generate candidate segments before matching them with the query; and proposal-free approaches [51, 56, 59, 34, 25, 33, 52], which directly predict the temporal boundaries in an end-to-end manner. Recently, VLM-based methods [39, 21, 28, 16, 58, 31, 15] have shown competitive performance by leveraging generalised knowledge from VLMs pre-training, while also maintaining the conversational capabilities of language models. However, both traditional and VLM-based approaches rely heavily on large amounts of labelled data and struggle to generalise to unseen domains [54, 27, 43, 8, 3]. Although some work has explored cross-dataset generalisation [30, 47, 26], this setting remains challenging due to two key issues: (i) target domain videos are often large, incurring high storage and adaptation costs; and (ii) temporal boundary annotation is labour-intensive. Moreover, our experiments show that the recent works, e.g., TimeZero [47] and Temporal-R1 [26], still perform poorly in cross-domain generalisation despite utilising GPRO to improve the model's generalisability with reinforcement learning. To address this, we propose a data-efficient cross-domain temporal grounding setting and a method that achieves performance comparable to fully supervised training using only a small number of unlabelled target domain videos. This significantly reduces resource requirements while improving cross-domain generalisability.

Uncertainty Estimation. Uncertainty estimation aims to capture either data-inherent noise (aleatoric uncertainty) or model-driven ambiguity (epistemic uncertainty) [5, 23]. In vision tasks, epistemic uncertainty is commonly approximated via Bayesian neural networks such as MC Dropout [10, 20], or through approximate reasoning methods [46, 36]. For VLMs, existing approaches quantify uncertainty through logit entropy [13], verbalized confidence enhanced by CoT prompting [49], or consistency-based diagnostics [63]. Recent work also explores semantic-level uncertainty by clustering outputs and computing entropy in embedding space [9]. In contrast, our URPA estimates epistemic uncertainty by computing the standard deviation across multiple rollout predictions, offering a lightweight and task-aligned signal to guide self-learning in unlabelled target domains.

3 Uncertainty-quantified Rollout Policy Adaptation

3.1 Problem Definition

In training a model for a new target domain video temporal grounding, we consider a labelled source domain $D_s = \{V_s^i, I_s^i\}_{i=1}^N$ and an target domain with a small number of unlabelled videos $D_t = \{V_t^i\}_{i=1}^K$, where I_s^i is the annotated interval for V_s^i, the number of training videos in the source

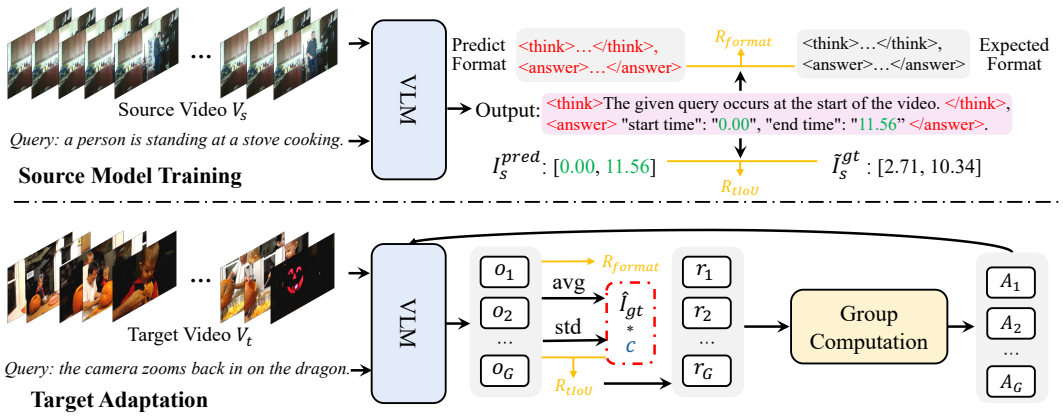


Figure 2: Uncertainty-quantified Rollout Policy Adaptation (URPA): During source model training, we perform supervised GRPO training using labelled videos V_s. Specifically, the format reward R_{format} encourages the model to "think first and then answer," while accuracy reward R_{tiou} aligns the predicted temporal grounding I_s^{pred} with the relaxed ground truth $\tilde{I}_s^{\text{gt}}$ for supervised learning. In target model knowledge adaptation, we adapt the model using K unlabelled target videos. For each video V_t, we first compute the average output over G rollouts to obtain a pseudo label $\hat{I}_t^{\text{gt}}$. Then calculate the standard deviation across these rollouts and transform it into a confidence score c to quantify uncertainty on pseudo labels, which is then used to weight different pseudo-labels when constructing a weighted reward function for test-time target model adaptation.

and target domains are N and K respectively, with $K \ll N$. A distribution shift exists between the source and target domains, i.e. $\mathcal{P}_s \neq \mathcal{P}_t$. The model is first pre-trained on the labelled source domain to learn some general knowledge of video temporal grounding. It then learns to adapt at test-time without labelled training in the target domain through a data-efficient unlabelled target domain videos. This real-time cross-domain unlabelled adaptation approach to video temporal grounding aims to optimise cross-domain knowledge transfer with only a few target videos without any annotations.

3.2 Remark on Group Relative Policy Optimization

In reinforcement learning, Group Relative Policy Optimization (GRPO) is a variant of Proximal Policy Optimization (PPO) that eliminates the need for a critic function. Instead, it directly evaluates the quality of predictions using a group of sampled responses. Given a question q, the model generates G candidate rollout responses $o = \{o_1, o_2, \ldots, o_G\}$ through policy sampling. A reward function then computes scores $r = \{r_1, \ldots, r_G\}$ by comparing each candidate with the ground truth. To normalize these rewards, GRPO computes their mean and standard deviation. The quality of each response is estimated by:

$$A_i = \frac{r_i - \text{mean}(r)}{\text{std}(r)}, \tag{1}$$

where A_i denotes the normalized advantage of the i-th response. GRPO optimizes the policy π_θ to maximize A_i, thereby encouraging beneficial deviation from the initial policy. A KL-divergence regularization term is further incorporated to constrain excessive deviation:

$$\max_{\pi_\theta} \mathbb{E}_{o\sim\pi_{\theta_{\text{old}}}(p)} \left[\left(\sum_{i=1}^{G} \frac{\pi_\theta(o_i)}{\pi_{\theta_{\text{old}}}(o_i)} \cdot A_i \right) - \beta\, D_{\text{KL}}\left(\pi_\theta \,\|\, \pi_{\text{ref}}\right) \right], \tag{2}$$

where β is a regularization coefficient controlling the strength of the penalty. Although GRPO has shown strong performance on math reasoning tasks, it struggles to generalise under domain shifts, especially for temporal grounding on unlabelled videos. This is due to both the difficulty of label-free optimisation and the high storage and adaptation cost associated with large-scale video data. To overcome these limitations, we introduce a data-efficient cross-domain temporal grounding setting, which consists of source domain model training and lightweight target domain adaptation.

3.3 Source Model Training

VLMs focus on coarse-grained understanding but lack fine-grained temporal localization capability. Therefore, post-training is necessary to equip the model with temporal reasoning and grounding abilities. To this end, we pursue two main objectives. First, we aim to stimulate explicit reasoning

chains within the model, enabling it to perform structured temporal inference. Second, we design task-specific reward functions to guide the model towards learning accurate temporal grounding, encouraging outputs that precisely align with relevant video segments.

Format Reward. To facilitate explicit reasoning, following [47], we introduce a format reward that encourages the model to structure its outputs according to a specified template. Specifically, we require the model to present its thought process within `<think>...</think>` tags, and its final answer within `<answer>...</answer>` tags. We use regular expression matching to determine whether the model's output adheres to the required format as follows:

$$R_{\text{format}} = \begin{cases} 1, & \text{if output matches format,} \\ 0, & \text{if output doesn't match format.} \end{cases} \tag{3}$$

Accuracy Reward. A core objective for source training is to design a well-crafted reward that guides the model to perform reasoning based on reward variations, thereby improving the quality of temporal grounding predictions. To guide the model towards this goal, we design an accuracy reward based on the overlap between the predicted and ground-truth temporal intervals.

However, event boundaries in videos are inherently ambiguous, and ground-truth annotations across datasets often exhibit labelling biases due to human subjectivity. Such biases can lead to performance degradation when models trained on the source domain are applied to the target domain. To mitigate this issue, we first relax the ground-truth interval $I_{\text{s}} = [\tau_s^{\text{start}}, \tau_s^{\text{end}}]$ by extending both boundaries by a fixed proportion α of the event duration, yielding the relaxed ground-truth $\tilde{I}_{\text{s}}^{gt}$:

$$\tilde{I}_{\text{s}}^{gt} = \left[\tilde{\tau}_s^{\text{start}},\ \tilde{\tau}_s^{\text{end}}\right] = \left[\max(0, \tau_s^{\text{start}} - \alpha(\tau_s^{\text{end}} - \tau_s^{\text{start}})),\ \min(1, \tau_s^{\text{end}} + \alpha(\tau_s^{\text{end}} - \tau_s^{\text{start}})\right], \tag{4}$$

where τ_s^{start} and τ_s^{end} are normalized timestamps within $[0, 1]$, and α is set to 0.1. Let $I_s^{\text{pred}} = [p_s^{\text{start}}, p_s^{\text{end}}]$ denote the predicted temporal interval. We then compute the relaxed temporal Intersection over Union (tIoU) between I_s^{pred} and $\tilde{I}_s^{\text{gt}}$ as:

$$R_{\text{tIoU}} = \frac{|\tilde{I}_s^{\text{gt}} \cap I_s^{\text{pred}}|}{|\tilde{I}_s^{\text{gt}} \cup I_s^{\text{pred}}|} = \frac{\max(0, \min(p_s^{\text{end}}, \tilde{\tau}_s^{\text{end}}) - \max(p_s^{\text{start}}, \tilde{\tau}_s^{\text{start}}))}{\max(p_s^{\text{end}}, \tilde{\tau}_s^{\text{end}}) - \min(p_s^{\text{start}}, \tilde{\tau}_s^{\text{start}})}, \tag{5}$$

where $\tilde{\tau}_s^{\text{start}}$ and $\tilde{\tau}_s^{\text{end}}$ are the relaxed ground-truth boundaries defined in Eq.(4). A higher R_{tIoU} indicates better alignment between the predicted and relaxed ground-truth segments. Thus, the overall supervised reward function for source domain training is defined as:

$$R_s = 0.5 \times R_{\text{format}} + 0.5 \times R_{\text{tIoU}}, \tag{6}$$

Supervised training on the source domain equips the model with basic spatio-temporal reasoning capability, but it still suffers from performance degradation on the target domain due to the domain shift between source and target domains, making target unsupervised adaptation essential.

3.4 Target Adaptation

GRPO requires ground-truth labels to compute rewards, which are unavailable in the target domain. Meanwhile, traditional domain adaptation methods typically assume full access to target domain data during training. Nevertheless, this assumption becomes impractical for large-scale video tasks due to substantial storage and computational costs. Thus we consider a more realistic setting where only a small number of K unlabelled target videos are available for adaptation, with $K \ll N$.

While pretraining on the source domain provides the model with general temporal grounding ability, performance still degrades under domain shift. To bridge this gap, we leverage the multiple candidate responses generated during GRPO optimization. For each target sample, the policy model samples G rollout response candidates. Although each individual response may be noisy, their aggregated statistics offer a more stable approximation. We therefore construct pseudo temporal labels by averaging predicted start and end timestamps:

$$\hat{\tau}_t^{\text{start}} = \frac{1}{G}\sum_{j=1}^{G}(p_t^{\text{start}})_{(j)}, \quad \hat{\tau}_t^{\text{end}} = \frac{1}{G}\sum_{j=1}^{G}(p_t^{\text{end}})_{(j)}, \tag{7}$$

where $(p_t^{\text{start}})_{(j)}$ and $(p_t^{\text{end}})_{(j)}$ denote the predicted timestamps from the j-th sampled response. These pseudo labels act as soft supervision after processed with Eq.(4) for test-time adaptation. However,

not all pseudo labels are equally reliable. To evaluate their quality, we estimate prediction uncertainty. Inspired by Bayesian deep learning, we treat the G rollout responses as approximate samples from a predictive distribution, analogous to the Monte Carlo Dropout approach. This allows us to use the standard deviation of the predicted timestamps as a proxy for uncertainty:

$$u = \sigma\left(\left\{(p_t^{\text{start}})_{(j)}\right\}_{j=1}^{G}\right) + \sigma\left(\left\{(p_t^{\text{end}})_{(j)}\right\}_{j=1}^{G}\right), \tag{8}$$

The uncertainty u is then converted into a confidence score c via an exponential decay function:

$$c = \exp(-\gamma u), \tag{9}$$

where γ is a hyperparameter that controls the influence of uncertainty on confidence. Finally, this confidence score is used to weight the contribution of the pseudo labels during reward computation. Specifically, we scale the temporal grounding reward (R_{tIoU}) by the confidence, while keeping the format reward unweighted:

$$R_s = 0.5 \times R_{\text{format}} + 0.5 \times R_{\text{tIoU}} \times c. \tag{10}$$

This confidence-aware reward formulation allows the model to better exploit informative pseudo labels while mitigating the effects of noisy or uncertain predictions, leading to more robust test-time adaptation in the absence of ground-truth labels.

4 Theoretical Analysis

This section proves that the empirical standard deviation obtained from multiple GRPO rollouts converges to the Bayesian predictive standard deviation, thereby quantifying epistemic uncertainty.

Theorem 4.1. *Fix an input x. Let $\pi_\theta(\tau \mid x)$ be the rollout policy learned by GRPO and let $p^\star(\tau \mid x)$ denote the Bayesian predictive distribution of the same model class trained on the source data. Assume **(i)** rollouts drawn from π_θ are i.i.d.; **(ii)** there exists a constant $M < \infty$ such that $\mathbb{E}_{\pi_\theta}[\tau^2] \le M$ and $\mathbb{E}_{p^\star}[\tau^2] \le M$; **(iii)** the Kullback–Leibler divergence is finite, $\varepsilon = \mathrm{KL}\big(\pi_\theta \,\|\, p^\star\big) < \infty$. Draw G independent rollouts $\{\tau^{(i)}\}_{i=1}^{G} \sim \pi_\theta(\cdot \mid x)$ and define*

$$\overline{\tau}_G = \frac{1}{G}\sum_{i=1}^{G} \tau^{(i)}, \qquad \widehat{\sigma}_G(x) = \sqrt{\frac{1}{G}\sum_{i=1}^{G}\big(\tau^{(i)} - \overline{\tau}_G\big)^2}.$$

As $G \to \infty$ and GRPO training drives $\varepsilon \to 0$, the estimator $\widehat{\sigma}_G(x)$ converges to the Bayesian predictive standard deviation $\mathrm{Std}_{p^\star}[\tau \mid x]$, providing a consistent measure of epistemic uncertainty.

Proof. For clarity we give the main steps; full details are in the supplementary material Sec. **??**.

Step 1: Consistency under the GRPO policy. Let $\mu_\pi = \mathbb{E}_{\pi_\theta}[\tau]$ and $\sigma_\pi(x) = \mathrm{Std}_{\pi_\theta}[\tau \mid x]$. Because the G rollouts $\{\tau^{(i)}\}$ are i.i.d. and $\mathbb{E}_{\pi_\theta}[\tau^2] \le M$ (Theorem assumption (ii)), the weak law of large numbers yields

$$\overline{\tau}_G \xrightarrow{p} \mu_\pi, \qquad \widehat{\sigma}_G(x) \xrightarrow{p} \sigma_\pi(x) \quad (G \to \infty). \tag{11}$$

Step 2: Relating the GRPO and Bayesian standard deviations. With $\varepsilon = \mathrm{KL}(\pi_\theta \| p^\star)$ (assumption (iii)), Pinsker's inequality gives a total-variation bound

$$\|\pi_\theta - p^\star\|_{\mathrm{TV}} \le \sqrt{\tfrac{1}{2}\,\varepsilon}. \tag{12}$$

Using (12), the bounded-moment assumption and $\mathrm{Var}[\tau] = \mathbb{E}[\tau^2] - (\mathbb{E}[\tau])^2$ one obtains the variance gap bound:

$$\big|\sigma_\pi^2(x) - \sigma_\star^2(x)\big| \le 4M\sqrt{\varepsilon}, \qquad \sigma_\star^2(x) = \mathrm{Var}_{p^\star}[\tau \mid x]. \tag{13}$$

Applying the identity $|a - b| \le \sqrt{|a^2 - b^2|}$ for non-negative a, b, we get:

$$\big|\sigma_\pi(x) - \sigma_\star(x)\big| \;\le\; \sqrt{4M\sqrt{\varepsilon}}. \tag{14}$$

Step 3: Convergence to Bayesian predictive standard deviation. GRPO optimisation decreases ε over training, so $\sqrt{\varepsilon} \to 0$. Combining (11) and the above yields

$$\big|\widehat{\sigma}_G(x) - \sigma_\star(x)\big| \;\le\; \underbrace{\big|\widehat{\sigma}_G(x) - \sigma_\pi(x)\big|}_{\text{vanishes by (11)}} + \underbrace{\big|\sigma_\pi(x) - \sigma_\star(x)\big|}_{\text{vanishes as } \varepsilon \to 0}, \tag{15}$$

Table 1: Performance comparisons on three cross-domain temporal grounding benchmarks. The best is in **bold**.

Method	Charades→ActivityNet		ActivityNet→TACoS		TACoS→Charades	
	R@0.5	R@0.7	R@0.3	R@0.5	R@0.5	R@0.7
Full-dataset Unsupervised Domain Adaptation						
CBP [44]	27.46	15.37	25.33	21.79	22.38	11.95
SCDM [55]	28.02	15.84	22.68	17.45	35.95	25.18
CMIN [62]	34.25	18.63	20.51	15.04	28.06	18.22
CSMGAN [29]	36.92	20.04	29.63	18.07	36.45	22.86
2DTAN [61]	39.17	21.76	33.72	21.16	25.81	17.37
DRN [57]	41.39	24.27	32.07	19.96	36.16	24.52
MMN [48]	44.06	24.98	36.94	22.08	33.73	20.04
UDA-TSL [30]	**49.48**	**32.15**	**42.40**	**29.83**	**41.39**	**28.63**
Domain Generalisation						
Qwen2.5-7B [53]	15.18	7.90	7.70	2.77	32.93	15.35
+ GRPO (Source Training only)	35.61	16.61	20.80	9.87	53.51	28.69
+ URPA (Source Training only)	36.46	18.78	21.58	10.26	54.40	29.81
Data-Efficient Unsupervised Domain Adaptation						
URPA (with 100-shot Target Adaptation)	40.13	20.88	21.82	10.16	54.78	30.08
URPA (with 200-shot Target Adaptation)	**42.57**	**21.25**	**21.97**	**10.38**	**55.54**	**32.04**

Method	TACoS→ActivityNet		Charades→TACoS		ActivityNet→Charades	
	R@0.5	R@0.7	R@0.3	R@0.5	R@0.5	R@0.7
Full-dataset Unsupervised Domain Adaptation						
CBP [44]	18.94	11.93	22.88	19.26	32.82	14.39
SCDM [55]	19.65	11.80	18.97	16.82	52.56	34.82
CMIN [62]	22.17	13.72	19.38	15.34	45.03	31.74
CSMGAN [29]	23.88	14.67	25.43	16.12	45.60	32.28
2DTAN [61]	24.90	16.38	30.12	19.81	36.34	22.61
DRN [57]	24.93	18.52	28.60	16.73	50.47	29.02
MMN [48]	28.29	20.86	34.09	19.17	50.78	23.17
UDA-TSL [30]	**33.54**	**26.16**	**36.42**	**25.48**	**60.26**	**41.03**
Domain Generalisation						
Qwen2.5-7B [53]	15.18	7.90	7.70	2.77	32.93	15.35
+ GRPO (Source Training only)	34.82	17.02	12.55	5.13	62.13	36.06
+ URPA (Source Training only)	36.23	18.13	13.91	6.27	63.36	37.47
Data-Efficient Unsupervised Domain Adaptation						
URPA (with 100-shot Target Adaptation)	38.41	19.80	14.49	7.23	64.35	38.76
URPA (with 200-shot Target Adaptation)	**41.83**	**21.84**	**16.62**	**8.25**	**65.12**	**39.57**

which tends to zero in probability as $G \to \infty$ and $\varepsilon \to 0$. Hence

$$\widehat{\sigma}_G(x) \xrightarrow{p} \sigma_\star(x), \tag{16}$$

showing that the rollout standard deviation is a consistent estimator of epistemic uncertainty. □

5 Experiments

To evaluate the effectiveness of the proposed method, we conduct experiments on three widely used temporal grounding datasets: TACoS [38], ActivityNet Captions [2], and Charades-STA [40].

5.1 Experimental Setup

Datasets. We evaluate our method on three benchmark datasets: TACoS, ActivityNet Captions, and Charades-STA. ActivityNet Captions contains approximately 20000 untrimmed YouTube videos annotated with 100000 natural language descriptions. Following the standard split, we use 37417 sentence-video pairs for training, and 17031 for testing. TACoS consists of 127 cooking-related videos. We adopt the public split, which includes 10146 and 4083 query-segment pairs for training and testing, respectively. Charades-STA is built on the Charades dataset, containing 12408 training and 3720 testing moment-query pairs. Codes are given in supplemental materials.

Baselines. To evaluate both cross-domain generalisation and adaptation, we follow the standard protocol for temporal grounding under domain shift. In each experiment, one dataset serves as the labelled source domain, while the remaining two act as unlabelled target domains, yielding six cross-domain configurations. We compare our method under the data-efficient unsupervised adaptation

Table 2: Ablation study on Charades → Activitynet dataset showing the impact of uncertainty estimation and soft labelling.

Variant	R@0.3	R@0.5	R@0.7	mIoU
Qwen2.5-7B	25.11	15.18	7.90	18.47
Qwen2.5-7B + Our GRPA Target Adaptation (100-shot)	2.67	1.46	0.68	1.96
Qwen2.5-7B + Source Training with GRPO	53.91	35.61	16.61	35.81
Qwen2.5-7B + Source Training with our GRPA	55.23	36.46	18.78	37.64
URPA w/o uncertainty-quantification + w/o relaxed tIOU (200-shot)	58.58	31.15	20.75	39.86
URPA w/o uncertainty + w/ relaxed tIOU (200-shot)	63.72	41.48	20.47	43.10
URPA (200-shot)	**64.61**	**42.57**	**21.25**	**43.65**

Table 3: Ablation study of GRPO hyperparameters on the Charades → ActivityNet task: (a) effect of reward scaling factor γ, and (b) effect of rollout count G during adaptation.

γ	R@0.3	R@0.5	R@0.7	mIoU
2	59.01	40.04	19.98	40.09
5	59.06	**40.44**	20.13	40.25
10	**59.12**	40.13	**20.88**	**40.46**
25	58.86	39.89	19.97	40.11

(a) Varying γ with 100-shot unsupervised adaptation.

G (rollouts)	R@0.3	R@0.5	R@0.7	mIoU
4	63.41	42.27	21.39	42.55
8	**64.61**	**42.57**	21.25	**43.65**
16	63.19	41.78	**21.73**	42.40
32	62.56	41.65	21.54	42.15
64	60.08	40.98	20.86	41.37

(b) Varying G with 200-shot unsupervised adaptation.

setting, where the source-pretrained model is adapted using only K = 100 or 200 unlabelled target videos with our URPA. This setting is contrasted with the following baselines: (1) Full-set adaptation: The model is trained on the labelled source domain and adapted using the entire unlabelled target domain. We report results for several state-of-the-art unsupervised adaptation methods, including CBP [44], SCDM [55], CMIN [62], CSMGAN [29], 2DTAN [61], DRN [57], MMN [48], and UDA-TSL [30]. (2) Domain Generalisation: We evaluate the base Qwen2.5-7B model [53] (without any fine-tuning), a GRPO-trained model on the source domain, and a model trained using only the source training phase of our URPA framework. All three are trained exclusively on the source domain and directly evaluated on the target domain without any access to target data.

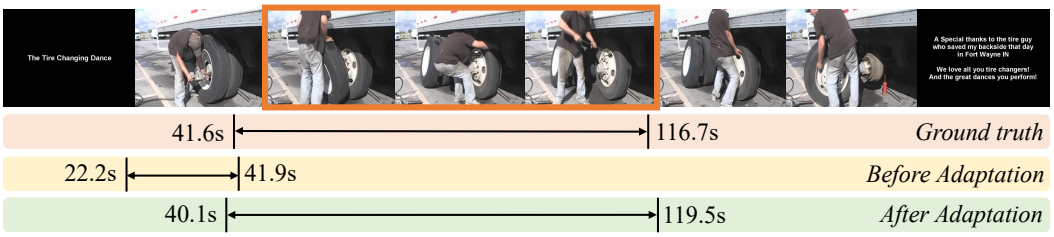
Implementation Details. We first fine-tune Qwen2.5-7B on the labelled source domain, and then perform data-efficient adaptation on 100 or 200 randomly selected samples from the unlabelled target domain. In all experiments, the maximum prompt length is set to 4096, the maximum response length to 2048, the number of rollouts to 8, batch size to 16, and we train for 1 epoch. All models are implemented in PyTorch on 32 NVIDIA V100 GPUs. Codes are in supplemental materials.

5.2 Results and Analysis

Experimental Results. Tab. 1 shows the performance of our method on the unsupervised cross-domain temporal grounding tasks. We report results using our data-efficient unsupervised adaptation approach with $K = 100$ and $K = 200$ unlabelled target samples. We compare against several baselines, including: (1) traditional UDA methods trained on full labelled source and unlabelled target domain data; (2) zero-shot domain generalisation methods trained only on the source domain and the base Qwen2.5-7B model. Our results demonstrate that data-efficient adaptation with just a small number of unlabelled target samples can significantly improve the performance of the source-pretrained model. Compared to full-set UDA methods, our method achieves competitive or even better performance in several settings. In particular, for TACoS → ActivityNet, TACoS → Charades, and ActivityNet → Charades tasks, our method outperforms state-of-the-art UDA methods in R@0.5. These results highlight the effectiveness and efficiency of our approach in realistic cross-domain scenarios with limited adaptation data. We also compare our approach with data-efficient target-supervised learning baselines in Appendix **??**, further validating its effectiveness.

Ablation Study. In Tab. 2, we present an ablation study on the Charades→ActivityNet cross-domain temporal grounding task to assess the impact of different components in our framework. The first row reports the performance of the base model, Qwen2.5-7B [53], directly applied to the temporal grounding task on ActivityNet without any adaptation. The poor performance indicates that the base model lacks inherent temporal localization capabilities. The second row presents the performance of Qwen2.5-7B adapted to the target domain using our method without any source domain pre-training. The subpar results suggest that source domain training provides essential task-specific knowledge

Query: He then pulls off the tire and begins unscrewing the back one.



Query: The person bunches up dough and lays it out on a board.

Figure 3: Qualitative Analysis on Charades → ActivityNet.

necessary for effective adaptation. Rows three and four show the results of models pre-trained on the source domain using the original GRPO algorithm and our improved method, respectively, and then directly evaluated on the target domain. Both outperform the base model significantly, with our method achieving better results than the original GRPO. This demonstrates that supervised training on the source domain aids the model in learning temporal grounding, and our soft accuracy reward approach mitigates biases introduced by manual annotations during pre-training. The last three rows analyse different modules in our target domain adaptation strategy, starting from a source-trained model. In the fifth row, pseudo labels are used directly for adaptation without modification in Eq.(4) and uncertainty quantification in Eq.(10), resulting in the lowest performance. In the sixth row, we apply Eq.(4) to the pseudo labels, which significantly improves results, indicating that soft label learning helps mitigate the effects of label noise. Finally, the last row introduces our uncertainty-quantified reward weighting mechanism (Eq.10) with soft labelling, yielding further improvements. This confirms the effectiveness of our GRPO-based uncertainty estimation strategy in guiding test-time adaptation.

Parameter Analysis. We analyse the effect of key hyperparameters on Charades → ActivityNet task performance. The reward scaling factor γ (used in Eq.(9)) controls the sensitivity to uncertainty: a larger γ increases the influence of predicted uncertainty on reward weighting. Tab. 3(a) reports the results on the 100-shot transfer setting with varying values of γ. We observe that $\gamma = 10$ yields the best performance, but the results are relatively stable across different γ values, indicating that our method is robust to this hyperparameter. Tab. 3(b) investigates the impact of the number of rollouts G in GRPO during 200-shot adaptation. The performance peaks when $G = 8$, and further increasing the number of rollouts does not lead to consistent gains, while incurring higher computational and memory costs. They suggest that our method is both effective and efficient under practical settings.

Qualitative Analysis. Fig. 3 illustrates several qualitative examples from the Charades → ActivityNet transfer setting. Despite not using any labelled data during target adaptation, our method produces significantly more accurate temporal grounding than the baseline. Notably, in these cases where the before adaptation model completely fails to localise the correct segment, our approach is able to capture meaningful behavioural patterns in the target domain and make precise predictions. These results highlight the effectiveness of our method in adapting to unseen domains without supervision.

6 Conclusion

In this paper, we propose URPA, a data-efficient method for cross-domain temporal grounding that leverages the rollout mechanism of GRPO to estimate predictive uncertainty and assess the reliability of predictions on unlabelled target-domain videos. This allows effective test-time adaptation using only a small number of target samples. We theoretically demonstrate that the standard deviation across rollouts approximates epistemic uncertainty. Extensive experiments across six cross-domain benchmarks validate the effectiveness of URPA in improving temporal grounding under domain shift.

References

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

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 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, 2015.

[3] Zixu Cheng, Yujiang Pu, Shaogang Gong, Parisa Kordjamshidi, and Yu Kong. Shine: Saliency-aware hierarchical negative ranking for compositional temporal grounding. In *European Conference on Computer Vision*, pages 398–416. Springer, 2024.

[4] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

[5] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009.

[6] Jianfeng Dong, Xianke Chen, Minsong Zhang, Xun Yang, Shujie Chen, Xirong Li, and Xun Wang. Partially relevant video retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 246–257, 2022.

[7] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4065–4080, 2021.

[8] Xiang Fang, Wanlong Fang, Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Renfu Li, Zichuan Xu, Lixing Chen, Panpan Zheng, et al. Not all inputs are valid: Towards open-set video moment retrieval using language. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 28–37, 2024.

[9] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

[10] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016.

[11] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[12] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017.

[13] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

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

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

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

[17] Zhijian Hou, Wanjun Zhong, Lei Ji, Difei Gao, Kun Yan, Wk Chan, Chong-Wah Ngo, Mike Zheng Shou, and Nan Duan. Cone: An efficient coarse-to-fine alignment framework for long video temporal grounding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8013–8028, 2023.

[18] Jian Hu, Zixu Cheng, Chenyang Si, Wei Li, and Shaogang Gong. Cos: Chain-of-shot prompting for long video understanding. *arXiv preprint arXiv:2502.06428*, 2025.

[19] Jian Hu, Hongya Tuo, Chao Wang, Lingfeng Qiao, Haowen Zhong, Junchi Yan, Zhongliang Jing, and Henry Leung. Discriminative partial domain adversarial network. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pages 632–648. Springer, 2020.

[20] Jian Hu, Haowen Zhong, Fei Yang, Shaogang Gong, Guile Wu, and Junchi Yan. Learning unbiased transferability for domain adaptation by uncertainty modeling. In *European Conference on Computer Vision*, pages 223–241. Springer, 2022.

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

[22] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710, 2024.

[23] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *arXiv preprint arXiv:1703.04977*, 2017.

[24] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017.

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

[26] Hongyu Li, Songhao Han, Yue Liao, Jialin Gao, and Si Liu. Envolving temporal reasoning capability into lmms via temporal consistent reward. `https://github.com/appletea233/Temporal-R1`, 2025.

[27] Juncheng Li, Junlin Xie, Long Qian, Linchao Zhu, Siliang Tang, Fei Wu, Yi Yang, Yueting Zhuang, and Xin Eric Wang. Compositional temporal grounding with structured variational cross-graph correspondence learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3032–3041, 2022.

[28] Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, YiQing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Vu Tu, et al. Groundinggpt: Language enhanced multi-modal grounding model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6657–6678, 2024.

[29] Zhijie Lin, Zhou Zhao, Zhu Zhang, Qi Wang, and Huasheng Liu. Weakly-supervised video moment retrieval via semantic completion network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11539–11546, 2020.

[30] Daizong Liu, Xiang Fang, Xiaoye Qu, Jianfeng Dong, He Yan, Yang Yang, Pan Zhou, and Yu Cheng. Unsupervised domain adaptative temporal sentence localization with mutual information maximization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 3567–3575, 2024.

[31] Ye Liu, Zongyang Ma, Zhongang Qi, Yang Wu, Ying Shan, and Chang W Chen. Et bench: Towards open-ended event-level video-language understanding. *Advances in Neural Information Processing Systems*, 37:32076–32110, 2024.

[32] M Jehanzeb Mirza, Jakub Micorek, Horst Possegger, and Horst Bischof. The norm must go on: Dynamic unsupervised domain adaptation by normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14765–14775, 2022.

[33] WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23023–23033, 2023.

[34] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10810–10819, 2020.

[35] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*, pages 16888–16905. PMLR, 2022.

[36] N. Pawlowski, A. Brock, Matthew C. H Lee, M. Rajchl, and B. Glocker. Implicit weight uncertainty in neural networks. *arXiv preprint arXiv:1711.01297*, 2017.

[37] You Qin, Wei Ji, Xinze Lan, Hao Fei, Xun Yang, Dan Guo, Roger Zimmermann, and Lizi Liao. Grounding is all you need? dual temporal grounding for video dialog. *arXiv preprint arXiv:2410.05767*, 2024.

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

[39] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14313–14323, 2024.

[40] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 510–526. Springer, 2016.

[41] Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[42] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

[43] Guolong Wang, Xun Wu, Zheng Qin, and Liangliang Shi. Routing evidence for unseen actions in video moment retrieval. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 3024–3035, 2024.

[44] Jingwen Wang, Lin Ma, and Wenhao Jiang. Temporally grounding language queries in videos by contextual boundary-aware prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12168–12175, 2020.

[45] Wei Wang, Zhun Zhong, Weijie Wang, Xi Chen, Charles Ling, Boyu Wang, and Nicu Sebe. Dynamically instance-guided adaptation: A backward-free approach for test-time domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24090–24099, 2023.

[46] Ximei Wang, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Transferable calibration with lower bias and variance in domain adaptation. *arXiv preprint arXiv:2007.08259*, 2020.

[47] Ye Wang, Boshen Xu, Zihao Yue, Zihan Xiao, Ziheng Wang, Liang Zhang, Dingyi Yang, Wenxuan Wang, and Qin Jin. Timezero: Temporal video grounding with reasoning-guided lvlm. *arXiv preprint arXiv:2503.13377*, 2025.

[48] Zhenzhi Wang, Limin Wang, Tao Wu, Tianhao Li, and Gangshan Wu. Negative sample matters: A renaissance of metric learning for temporal grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2613–2623, 2022.

[49] Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*, 2023.

[50] Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. Large language models can learn temporal reasoning. *arXiv preprint arXiv:2401.06853*, 2024.

[51] Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. Multilevel language and vision integration for text-to-clip retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9062–9069, 2019.

[52] Shen Yan, Xuehan Xiong, Arsha Nagrani, Anurag Arnab, Zhonghao Wang, Weina Ge, David Ross, and Cordelia Schmid. Unloc: A unified framework for video localization tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13623–13633, 2023.

[53] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[54] Yitian Yuan, Xiaohan Lan, Xin Wang, Long Chen, Zhi Wang, and Wenwu Zhu. A closer look at temporal sentence grounding in videos: Dataset and metric. In *Proceedings of the 2nd international workshop on human-centric multimedia analysis*, pages 13–21, 2021.

[55] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. *Advances in Neural Information Processing Systems*, 32, 2019.

[56] Yitian Yuan, Tao Mei, and Wenwu Zhu. To find where you talk: Temporal sentence localization in video with attention based location regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9159–9166, 2019.

[57] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10287–10296, 2020.

[58] Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, et al. Timesuite: Improving mllms for long video understanding via grounded tuning. *arXiv preprint arXiv:2410.19702*, 2024.

[59] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6543–6554, 2020.

[60] Qi Zhang, Sipeng Zheng, and Qin Jin. No-frills temporal video grounding: Multi-scale neighboring attention and zoom-in boundary detection. *arXiv preprint arXiv:2307.10567*, 2023.

[61] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12870–12877, 2020.

[62] Zhu Zhang, Zhijie Lin, Zhou Zhao, and Zhenxin Xiao. Cross-modal interaction networks for query-based moment retrieval in videos. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 655–664, 2019.

[63] Ruochen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. *arXiv preprint arXiv:2305.03268*, 2023.