DTP-OPSD: REINFORCEMENT LEARNING WITH DISTRACTOR-AWARE TEMPORAL-PRIVILEGE ON-POLICY SELF-DISTILLATION FOR TEMPORAL VIDEO GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Temporal Video Grounding (TVG) aims to localize the start and end times of an event described by a language query in a video. The task becomes particularly challenging when the same video contains multiple repeated, similar, or opposite events. On these hard examples, all responses sampled by Group Relative Policy Optimization (GRPO) may miss the target and receive zero temporal Intersection-over-Union (tIoU). Their group-relative advantages are then all zero, leaving no useful policy gradient signal. Further analysis shows that the model is often able to recognize the queried event but is distracted by other events in the same video and selects the wrong occurrence. To address this, we propose DTP-OPSD, a reinforcement learning framework with distractor-aware temporal-privilege on-policy self-distillation. The student observes the clean video and generates responses on-policy, while a shared-weight teacher observes a privileged video in which the target window is marked by red full-frame borders and same-video distractor windows by blue full-frame borders. For the distillation loss, the teacher evaluates student-generated prefixes and provides dense token-level Jensen–Shannon divergence supervision. This supervision remains informative even when GRPO provides no policy gradient signal. We suppress distillation when the teacher does not outperform the student and strengthen it for groups where the teacher performs better but all student rollouts receive zero tIoU. We also construct a compact hard-interference training set through VLM screening and human verification. With only 300 additional post-training examples, DTP-OPSD achieves the highest mIoU among the compared open-source 8B models on Charades-TimeLens, ActivityNet-TimeLens, and QVHighlights-TimeLens, reaching 55.93, 54.77, and 66.71, respectively. Compared with Pure GRPO trained on the same data, DTP-OPSD improves all 12 reported metrics, with R1@0.7 gains of 1.28, 1.62, and 1.78 points on the three benchmarks.

1 INTRODUCTION

Temporal Video Grounding (TVG) aims to identify the start and end times of an event described by a language query (Gao et al., 2017; Hendricks et al., 2017). It is a fundamental video-language task that requires joint understanding of the query semantics and the temporal structure of the video. Earlier methods localize target moments through proposal matching, boundary regression, or span prediction (Zhang et al., 2020; Lei et al., 2021). More recently, multimodal large language models have formulated TVG as timestamp generation through a language interface (Ren et al., 2024; Huang et al., 2024a; Wang et al., 2025a; Zhang et al., 2026). This formulation makes TVG more flexible, but accurate localization remains difficult when a video contains multiple related event occurrences that provide plausible matches to the query.

Reinforcement learning is now widely used to improve TVG with multimodal large language models (Wang et al., 2025b; Chen et al., 2025; Dong et al., 2026). A predicted temporal window can be scored directly against the annotated target using temporal Intersection-over-Union (tIoU). The tIoU score provides a direct reward, without requiring a learned reward model or extra reward la-

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



Figure 1: **Top.** Qwen3-VL-8B recognizes the relevant actions but selects an earlier occurrence from the clean video, resulting in zero tIoU. The privileged teacher view marks the target with red borders and the distractors with blue borders, helping the same model recover the correct interval. **Bottom left.** On QVHighlights-TimeLens, mIoU decreases as the number of valid distractors increases, while DTP-OPSD degrades less than Qwen3-VL-8B and Pure GRPO. **Bottom right.** On the full QVHighlights-TimeLens evaluation set, the privileged view improves localization over the clean view under identical model weights.

bels. Group Relative Policy Optimization (GRPO) samples several responses for the same input and trains the policy using their relative rewards within the group (Shao et al., 2024; DeepSeek-AI, 2025). However, GRPO cannot learn from a rollout group when all sampled responses receive the same reward, because mean centering assigns every response zero relative advantage. In TVG, a particularly consequential case is the all-zero group.

A major challenge arises when a video contains multiple events related to the query. The queried event may occur repeatedly, while other moments may contain visually or semantically similar actions or opposite actions involving the same actor and object. These distractors make it difficult to identify the specific occurrence requested by the query. We refer to this problem as same-video interference. Under strong interference, the model may recognize the queried event but still select the wrong occurrence.

Figure 1 illustrates this failure. In this example, Qwen3-VL-8B (Bai et al., 2025) recognizes the queried action but selects an earlier occurrence, resulting in zero tIoU. With the same model weights, the privileged view helps distinguish the target from distractors and recover the correct interval. Dataset-level diagnostics further show that localization becomes more difficult as the number of valid same-video distractors increases, while DTP-OPSD remains more robust. The privileged view also outperforms the clean view on the full QVHighlights-TimeLens evaluation set. These results suggest that some failures arise from incorrect occurrence selection under same-video interference rather than from a failure to recognize the queried action.

Under strong same-video interference, all sampled responses may select distractor occurrences and receive zero tIoU, leaving some of the hardest examples inactive in the current GRPO update.

To reduce wrong-occurrence predictions and provide a learning signal for all-zero rollout groups, we propose DTP-OPSD, a reinforcement learning framework with distractor-aware temporal-privilege

on-policy self-distillation. As shown in Figure 2, the student observes the clean video and samples responses on-policy. A teacher sharing the student's current weights observes a training-only privileged view, where the target region is marked in red and the distractors in blue. For distillation, the teacher evaluates student-generated prefixes and provides dense token-level Jensen–Shannon divergence supervision (Agarwal et al., 2024; Zhao et al., 2026; Yuan et al., 2026).

Since the privileged teacher is not reliable on every sample, we introduce an adaptive distillation controller. The controller compares the mean Student tIoU from the GRPO rollouts with the mean teacher tIoU from independently generated teacher rollouts under the same sampling settings. It suppresses distillation when the teacher does not outperform the student and strengthens it for groups where the teacher performs better but all student rollout tIoUs are zero. All remaining groups use the default distillation weight. This design prevents unreliable teacher supervision while providing stronger guidance when GRPO receives no reward-driven learning signal.

Effective self-distillation requires hard but teachable examples. We further construct HI-TVG-300 from the TimeLens-100K training set (Zhang et al., 2026), without introducing external videos. VLM screening and human verification identify examples with strong same-video interference. A teachability filter then retains examples where the privileged teacher clearly outperforms the clean-video student. The resulting dataset contains 300 high-quality, difficult but teachable training examples. Our contributions are:

- We study a key challenge in reinforcement learning-based TVG. Same-video distractors can cause wrong-occurrence predictions and all-zero tIoU groups that provide no relative learning signal to GRPO.
- We propose distractor-aware temporal privilege, using red borders to mark the target region and blue borders to mark distractor regions. This enables on-policy distillation from a shared-weight teacher to a clean-video student.
- We design an adaptive distillation controller that suppresses distillation when the teacher does not outperform the student and strengthens supervision for all-zero student rollout groups when the teacher performs better.
- We construct HI-TVG-300, a high-quality hard-interference training set, through VLM screening, human verification, and teacher–student teachability filtering.

2 Related Work

Temporal video grounding and reinforcement learning. Early TVG methods localize moments through proposal matching, boundary regression, or span prediction (Gao et al., 2017; Hendricks et al., 2017; Zhang et al., 2020; Lei et al., 2021). Video MLLMs later introduced timestamp-aware representations and grounding instruction tuning for generative temporal localization (Ren et al., 2024; Huang et al., 2024a;b; Wang et al., 2025a; Zeng et al., 2025; Guo et al., 2026; 2025). Recent work extends TVG to multi-moment annotations (Cao et al., 2025; Xu et al., 2026). Reinforcement learning with verifiable rewards has also been applied to multimodal and video reasoning (Huang et al., 2026; Feng et al., 2025; Li et al., 2025), and more recently to temporal video grounding (Wang et al., 2025b; Chen et al., 2025; 2026; Dong et al., 2026; Zhang et al., 2026). However, group-relative optimization loses its relative learning signal when all rollout rewards are identical. We study this failure under same-video interference, where competing occurrences can produce wrong-occurrence predictions and all-zero tIoU groups.

On-policy distillation and self-distillation. Knowledge distillation transfers a teacher distribution to a student (Hinton et al., 2015). On-policy distillation provides teacher supervision along student-generated trajectories, while On-Policy Self-Distillation creates a teacher–student capability gap by conditioning shared model weights on different inputs (Agarwal et al., 2024; Zhao et al., 2026). Recent multimodal self-distillation methods construct privileged views from localized visual evidence or higher-resolution images (Yuan et al., 2026; Zhu et al., 2026). Video-OPD applies on-policy distillation to temporal video grounding using a fixed frontier teacher to supervise student-generated trajectories (Li et al., 2026). In contrast, DTP-OPSD combines GRPO with shared-weight self-distillation and constructs query-dependent temporal privilege by marking the target occurrence

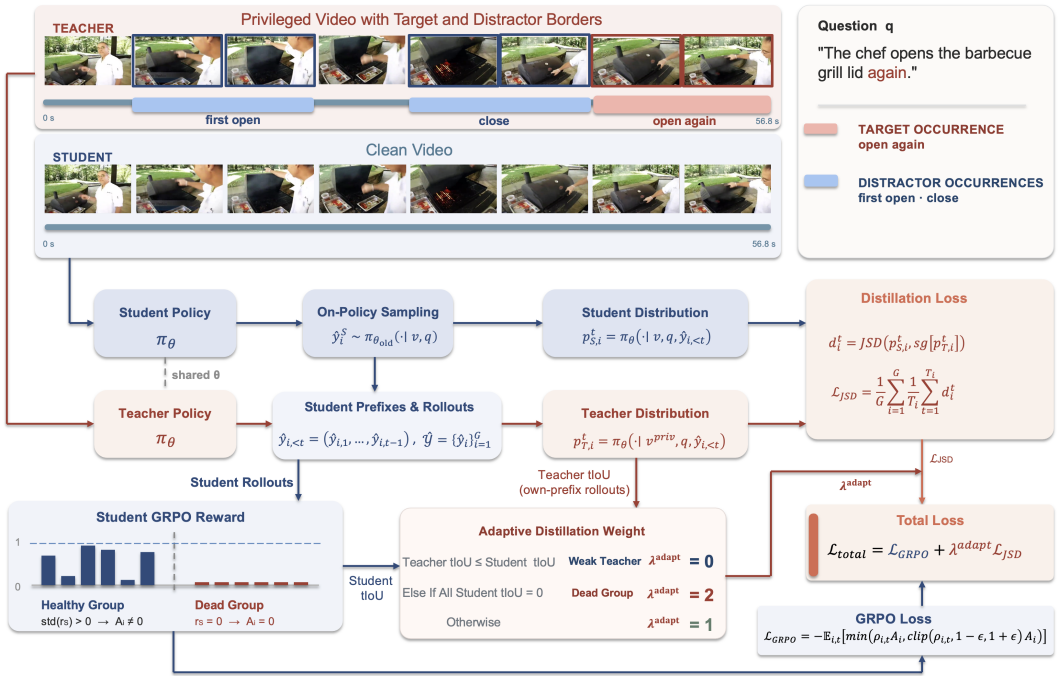

Figure 2: Overview of DTP-OPSD. The clean-video student samples G on-policy responses, whose tIoUs provide the GRPO rewards. The shared-weight teacher observes the privileged video and evaluates student-generated prefixes for token-level JSD. Separately, mean teacher tIoU is estimated from independent own-prefix rollouts under the same sampling settings. The controller uses the teacher and student tIoUs to adjust the group-level distillation weight. For clarity, the figure shows the group-level adaptive weighting, while Equation 5 gives the complete minibatch objective.

and its same-video distractors. Its adaptive controller suppresses unreliable teacher supervision and strengthens it for all-zero student groups.

3 METHOD

Given a video V and a language query q, the model predicts a temporal interval $\hat{\tau} = [\hat{t}_s, \hat{t}_e]$, with the ground-truth interval denoted by $\tau = [t_s, t_e]$. As shown in Figure 2, DTP-OPSD combines GRPO with distractor-aware temporal-privilege on-policy self-distillation. Section 3.1 describes tIoU-based GRPO and the all-zero rollout groups caused by same-video interference. Section 3.2 introduces temporal privilege and same-prefix JSD supervision. Section 3.3 presents teacher competence estimation and adaptive distillation control. Finally, Section 3.4 describes the construction of HI-TVG-300. In the notation below, g indexes a video–query rollout group (V_g, q_g), while i indexes one of the G student responses sampled for that group.

3.1 GRPO FOR TVG AND DEAD GROUPS

Reward and group-relative optimization. Following the thinking-free RLVR setting of TimeLens (Zhang et al., 2026), our only reward is temporal Intersection-over-Union:

$$\mathrm{tIoU}(\hat{\tau}, \tau) = \frac{|\hat{\tau} \cap \tau|}{|\hat{\tau} \cup \tau|}. \tag{1}$$

Here, $|\cdot|$ denotes the duration of an interval. A response that cannot be parsed into a valid interval receives zero reward. Unlike Time-R1 (Wang et al., 2025b), we use no format or temporal-distance reward, keeping the training reward consistent with the standard TVG evaluation metric.

For each video-query pair (V_g, q_g), the rollout policy $\pi_{\theta_{\text{old}}}$ samples G responses conditioned on the clean video and query. Each response is parsed into a predicted interval $\hat{\tau}_{g,i}$, and its tIoU is used as the student reward $r^S_{g,i}$. Following TimeLens, we compute the group-relative advantages by mean centering without standard-deviation scaling:

$$\bar{r}^S_g = \frac{1}{G}\sum_{j=1}^{G} r^S_{g,j}, \qquad A_{g,i} = r^S_{g,i} - \bar{r}^S_g. \tag{2}$$

We use these advantages in the group-relative GRPO objective (Shao et al., 2024; Zhang et al., 2026), denoted by $\mathcal{L}_{\text{GRPO}}$. The exact loss, completion-token normalization, and single-update optimization procedure are provided in Appendix A.1.

Dead groups under same-video interference. GRPO obtains a reward-driven learning signal from a rollout group only when its rewards differ. Under strong same-video interference, a critical failure mode occurs when every sampled response selects an incorrect occurrence that does not overlap with the target. In this case, $r^S_{g,i} = 0$ for all i, and Equation 2 yields $A_{g,i} = 0$ for all i. The group therefore contributes no reward-driven policy gradient to the current update. We refer to such an all-zero rollout group as a dead group. Although any constant-reward group has zero relative advantages, we use the all-zero condition to identify dead groups because it indicates complete failure to overlap the target. When the privileged teacher achieves a higher mean tIoU than the student, DTP-OPSD provides strengthened token-level supervision to these groups without relying on within-group reward differences.

3.2 Distractor-Aware Temporal-Privilege On-Policy Self-Distillation

Distractor-aware temporal privilege. For each training example, we construct an aligned clean video V and a privileged video V^{priv} with identical content and duration. In V^{priv}, frames inside the target window receive red full-frame borders, while frames inside the verified same-video distractor windows receive blue full-frame borders. Before rendering, we expand only the target window to 1.5 times its duration around its center. The red borders provide a coarse GT-derived temporal cue rather than displaying the annotated boundaries verbatim. Red takes priority wherever the target and distractor windows overlap. The clean video remains unchanged, and V^{priv} is used only during training.

Shared-weight teacher and student. Following On-Policy Self-Distillation (OPSD) (Zhao et al., 2026), the teacher and student are two conditional views of the same model. They share the current parameters θ and use the same text prompt, which does not mention the border colors. Their conditioning differs only in the video input. The resulting capability gap therefore comes from the privileged temporal evidence rather than additional parameters or textual hints.

On-policy same-prefix distillation. The student responses sampled for GRPO are reused for distillation. At each token position, the student and teacher compute their next-token distributions from the clean and privileged videos, respectively, while conditioning on the same student-generated prefix $\hat{y}_{g,i,<t}$. We denote these distributions by $p^t_{S,g,i}$ and $p^t_{T,g,i}$. The teacher therefore evaluates the generation states actually visited by the student, keeping the supervision on-policy.

We align the two distributions using Jensen–Shannon divergence (JSD) (Lin, 1991):

$$\begin{aligned} d^t_{g,i} &= \text{JSD}\big(p^t_{S,g,i}, \text{sg}\big[p^t_{T,g,i}\big]\big)\,, \\ \ell^{\text{JSD}}_g &= \frac{1}{G}\sum_{i=1}^{G}\frac{1}{T_{g,i}}\sum_{t=1}^{T_{g,i}} d^t_{g,i}. \end{aligned} \tag{3}$$

Here, $\text{sg}[\cdot]$ denotes stop-gradient, and $T_{g,i}$ is the number of valid completion tokens. Gradients propagate only through the clean-video student branch. This loss does not vanish merely because all student rollout tIoUs are zero.

3.3 Adaptive Distillation Control

Teacher competence from independent rollouts. The privileged view does not guarantee a better teacher prediction for every input. The teacher may ignore the border information or still select a distractor, so unreliable teacher supervision should be suppressed. To estimate teacher competence fairly, the teacher uses the same model snapshot, text prompt, sampling temperature, and number of rollouts G as the student. The only difference is the video input. The teacher independently generates G responses from the privileged video using its own autoregressive prefixes. Each response is parsed using the same rule as the student responses, and an invalid response receives zero reward. We denote the mean tIoU of these responses by $\bar{r}_g^T$ and use it as the teacher competence estimate. These teacher rollouts are used only by the adaptive controller and do not enter the JSD loss, which remains defined on student-generated prefixes in Equation 3.

Adaptive distillation weight. Let $\bar{r}_g^S$ denote the mean student tIoU defined in Equation 2. The group-level distillation multiplier is

$$\lambda_g^{\text{adapt}} = \begin{cases} 0, & \bar{r}_g^T \le \bar{r}_g^S, \\ 2, & \bar{r}_g^T > \bar{r}_g^S \text{ and } r_{g,i}^S = 0 \ \ \forall i, \\ 1, & \text{otherwise.} \end{cases} \tag{4}$$

The first condition removes distillation when the privileged teacher does not achieve a higher mean tIoU than the student. This includes groups where both the student and teacher receive only zero tIoU. The second condition strengthens distillation only when the teacher performs better and all student rollout tIoUs are zero. When the teacher performs better, all remaining groups, including student groups with identical nonzero rewards, retain the default weight.

Final objective. The overall objective combines the standard GRPO loss with group-adaptive token-level distillation:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{GRPO}} + \lambda_{\text{JSD}} \frac{1}{|\mathcal{B}|} \sum_{g \in \mathcal{B}} \lambda_g^{\text{adapt}} \ell_g^{\text{JSD}}, \tag{5}$$

where $\mathcal{B}$ is a minibatch of rollout groups and λ_{JSD} is the global distillation coefficient. The GRPO coefficient and λ_{JSD} are both set to 1 in our final configuration. The group-level multiplier $\lambda_g^{\text{adapt}} \in \{0, 1, 2\}$ controls the contribution of JSD for each rollout group. It removes unreliable teacher supervision and strengthens supervision for all-zero student groups when the teacher demonstrates better localization.

3.4 Construction of HI-TVG-300

Large-scale TVG datasets provide broad training coverage but are not designed to contain strong same-video interference or reliable privileged supervision. We therefore derive HI-TVG-300 entirely from the TimeLens-100K training set (Zhang et al., 2026). It reuses selected videos and queries while refining target intervals and adding verified distractor intervals, without introducing external videos or queries. As summarized in Figure 3, we first balance candidate videos across duration ranges and source datasets. GPT-5.6 then examines each complete video and its annotations to identify repeated, similar, or opposite event occurrences and propose refined target intervals. Human annotators verify the target boundaries, distractor intervals, video quality, and queries. We retain examples containing at least two valid same-video distractors. Finally, using the same Qwen3-VL-8B checkpoint and sampling settings, the clean-video student and privileged teacher independently generate eight responses. We retain an example when the mean teacher tIoU exceeds 0.7 and the mean student tIoU is below 0.3. This process produces 300 difficult but teachable examples with verified same-video interference. Further construction details, annotation protocols, and dataset statistics are provided in Appendix A.5.

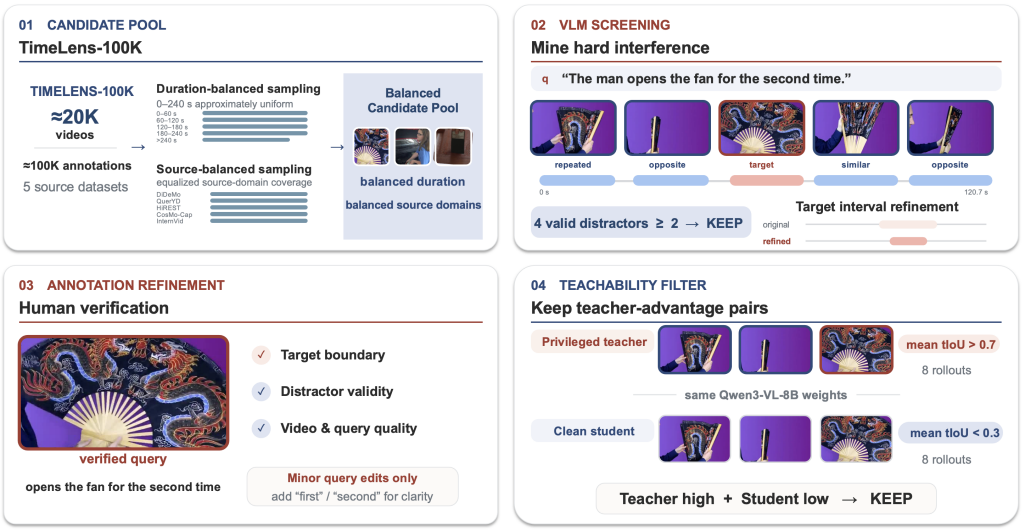

Figure 3: Construction pipeline of HI-TVG-300, derived entirely from the TimeLens-100K training set. Candidate videos are balanced by duration and source, mined for same-video interference, verified by human annotators, and filtered according to the privileged teacher advantage.

4 Experiments

We evaluate DTP-OPSD through comparisons with representative TVG methods and controlled ablations, with additional qualitative examples provided in Appendix C.

4.1 Experimental Setup

Benchmarks and evaluation metrics. We evaluate DTP-OPSD on Charades-STA (Gao et al., 2017), ActivityNet Captions (Krishna et al., 2017), and QVHighlights (Lei et al., 2021). Following TimeLens (Zhang et al., 2026), we use its corrected benchmarks for analysis and final reporting, referred to as Charades-TimeLens, ActivityNet-TimeLens, and QVHighlights-TimeLens. We further evaluate on OMTG-Bench (Xu et al., 2026) and QV-M^2 (Cao et al., 2025), where a query may correspond to multiple valid temporal intervals. We report mIoU as the primary metric, together with R1@0.3, R1@0.5, and R1@0.7. mIoU averages the per-example tIoU, while R1@m measures the percentage of predictions with a tIoU of at least m.

Implementation details. We initialize Qwen3-VL-8B from the publicly released TimeLens supervised checkpoint, denoted by TimeLens-SFT, and post-train it on HI-TVG-300, which is derived entirely from the TimeLens-100K training set. Videos are sampled at 2 FPS with a maximum of 448 frames and a visual-token budget of 14,336. For each input, the student samples $G = 8$ responses for GRPO. The teacher independently samples the same number of responses under identical sampling settings to estimate teacher competence. These teacher rollouts are not used to compute the JSD loss. We use a global batch of eight prompt-level rollout groups and a learning rate of 1×10^{-6}. Unless otherwise specified, results from our post-training experiments and training ablations are reported as the mean over three independent training runs. The global JSD coefficient is set to $\lambda_{\text{JSD}} = 1$, and the reference-policy KL coefficient is set to $\beta = 0$. Additional training and implementation details are provided in Appendix A.

4.2 Comparison with State-of-the-Art Methods

Main results. Table 1 compares DTP-OPSD with representative open-source TVG models on three TimeLens benchmarks. Pure GRPO trained on HI-TVG-300 already improves the TimeLens-SFT initialization on all three benchmarks. Building on this strong baseline, DTP-OPSD achieves 55.93, 54.77, and 66.71 mIoU on Charades-TimeLens, ActivityNet-TimeLens, and QVHighlights-

Table 1: Comparison with representative open-source TVG models on three TimeLens benchmarks. We evaluate all model checkpoints ourselves using the corrected TimeLens annotations and the same TimeLens evaluation protocol. TimeLens-SFT provides the common initialization for TimeLens-8B, Pure GRPO, and DTP-OPSD. Both Pure GRPO and DTP-OPSD are post-trained on our HI-TVG-300 set. The best and second-best results are shown in bold and underlined, respectively.

	Charades-TimeLens				ActivityNet-TimeLens				QVHighlights-TimeLens			
Method	R1@0.3	R1@0.5	R1@0.7	mIoU	R1@0.3	R1@0.5	R1@0.7	mIoU	R1@0.3	R1@0.5	R1@0.7	mIoU
VideoChat-R1-7B	51.93	30.86	11.72	33.78	35.03	23.94	11.36	25.01	29.37	19.13	9.42	21.56
Time-R1-7B	57.95	32.07	16.96	36.62	44.84	31.08	19.02	33.19	65.87	51.59	36.14	49.26
MiMo-VL-7B	57.91	42.63	20.58	39.65	49.34	38.76	22.48	35.52	57.12	42.61	28.46	41.57
Qwen3-VL-8B	69.24	53.41	27.57	48.32	62.18	51.29	34.43	46.86	74.27	64.69	49.35	59.48
Video-OPD-8B	75.62	60.57	32.71	53.16	66.19	57.13	39.47	51.82	76.94	68.72	54.45	64.39
TimeLens-8B	76.61	63.03	35.26	55.28	68.97	58.45	40.68	53.23	80.21	**71.62**	55.57	65.59
TimeLens-SFT	74.19	59.35	31.12	51.85	63.44	54.09	36.61	49.26	69.57	59.31	45.38	57.34
Pure GRPO	77.38	64.17	35.15	55.43	68.26	59.58	41.85	53.62	79.75	70.93	55.61	65.95
DTP-OPSD	**77.94**	**64.67**	**36.43**	**55.93**	**69.67**	**60.71**	**43.47**	**54.77**	**80.53**	71.45	**57.39**	**66.71**

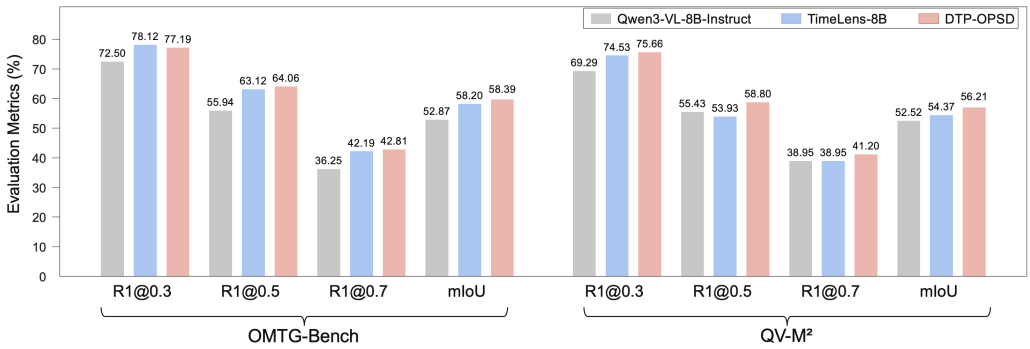

Figure 4: Single-interval grounding on OMTG-Bench and QV-M^2, both with multi-moment annotations. Each model predicts one interval and is evaluated against its best-matching ground-truth interval. Results are reported on all 320 OMTG-Bench queries and all 323 QV-M^2 test queries.

TimeLens, respectively. These are the highest mIoU scores among the compared open-source 8B models. Compared with Pure GRPO, DTP-OPSD improves all 12 reported metrics, including R1@0.7 gains of 1.28, 1.62, and 1.78 points on the three benchmarks. Among all compared models, DTP-OPSD ranks first on 11 of the 12 metrics and second on the remaining metric, demonstrating consistent improvements across different overlap thresholds.

Single-interval grounding under multi-moment annotations. We further evaluate DTP-OPSD on OMTG-Bench (Xu et al., 2026) and QV-M^2 (Cao et al., 2025). All models retain the standard single-interval prompt and output format, without benchmark-specific training or multi-interval decoding. For both benchmarks, we adopt the mIoU@1 protocol used by QV-M^2. Each prediction is evaluated against its best-matching ground-truth interval, and the resulting tIoU scores are used to compute mIoU and R1. As shown in Figure 4, DTP-OPSD achieves 58.39 mIoU on OMTG-Bench and 56.21 mIoU on QV-M^2. It performs comparably with TimeLens-8B on OMTG-Bench and exceeds it by 1.84 points on QV-M^2. These results show that DTP-OPSD remains effective when a query has multiple valid temporal occurrences. This protocol evaluates whether the model identifies one valid occurrence rather than whether it retrieves all annotated intervals.

4.3 Ablation Study

We ablate the adaptive controller, privileged video design, and data curation, with additional results in Appendix B. All variants use Qwen3-VL-8B initialized from TimeLens-SFT, with the same training settings and evaluation step. Except for the data ablation, all variants use HI-TVG-300. These analyses determine one shared configuration, which is applied unchanged to all other benchmarks.

Table 2: Effect of the adaptive distillation controller.

	Charades-TimeLens				ActivityNet-TimeLens				QVHighlights-TimeLens			
Variant	R1@0.3	R1@0.5	R1@0.7	mIoU	R1@0.3	R1@0.5	R1@0.7	mIoU	R1@0.3	R1@0.5	R1@0.7	mIoU
Fixed JSD	77.18	63.98	35.61	55.48	68.51	59.91	42.33	54.02	79.59	70.12	54.99	65.78
Rescue only	77.61	64.45	36.11	55.77	69.25	60.45	43.01	54.50	80.14	70.99	56.47	66.33
Suppression only	77.45	64.31	35.95	55.62	69.10	60.33	42.82	54.37	80.20	70.80	55.88	66.24
Full controller	**77.94**	**64.67**	**36.43**	**55.93**	**69.67**	**60.71**	**43.47**	**54.77**	**80.53**	**71.45**	**57.39**	**66.71**

Table 3: Privileged-view ablation on QVHighlights-TimeLens.

Privileged view	R1@0.3	R1@0.5	R1@0.7	mIoU
Red-shuffled target	77.78	68.23	51.97	64.44
Target only	80.18	71.18	56.10	66.12
Target + random distractors	80.02	70.96	55.68	65.94
Target + verified distractors	**80.53**	**71.45**	**57.39**	**66.71**

Table 4: Data-curation ablation on QVHighlights-TimeLens.

Training subset	R1@0.3	R1@0.5	R1@0.7	mIoU
Random subset	79.75	70.62	54.90	65.53
Human-verified subset	80.14	71.26	56.02	66.05
HI-TVG-300	**80.53**	**71.45**	**57.39**	**66.71**

Effect of the adaptive distillation controller. Table 2 evaluates adaptive JSD weighting. All variants retain GRPO and privileged-view JSD. Fixed JSD sets $\lambda_g^{\text{adapt}} = 1$ for every group, whereas Pure GRPO in Table 1 removes JSD entirely. Rescue only sets $\lambda_g^{\text{adapt}} = 2$ for eligible all-zero groups and otherwise retains the default weight, while Suppression only sets $\lambda_g^{\text{adapt}} = 0$ when the teacher does not outperform the student. Both variants improve over Fixed JSD, and their combination achieves the best mIoU on all three benchmarks, with gains of 0.45, 0.75, and 0.93 points. The full controller consistently outperforms both Pure GRPO and Fixed JSD, supporting adaptive JSD weighting based on teacher quality and all-zero groups.

Effect of the privileged video design. Table 3 shows that the gains depend on correct temporal cues rather than colored borders alone. Moving the red border to an incorrect interval gives 64.44 mIoU, whereas marking the true target reaches 66.12. Adding random distractor borders slightly reduces mIoU to 65.94. Replacing them with verified same-video distractors improves R1@0.7 from 55.68 to 57.39 and mIoU from 65.94 to 66.71. Thus, the improvement depends on both an aligned target cue and meaningful distractor cues.

Effect of training data curation. Table 4 compares three 300-example subsets from the same TimeLens-100K training pool. In the random subset, the original target is marked in red, while unverified intervals associated with other queries from the same video serve as candidate blue cues. Human screening and verification increase mIoU from 65.53 to 66.05, and teachability filtering further increases it to 66.71. These results show the cumulative benefits of example selection, distractor verification, and teacher–student teachability filtering.

5 Conclusion

We study same-video interference in TVG, where a model can recognize the queried event yet select the wrong occurrence. Under strong interference, all sampled intervals may miss the target, producing an all-zero tIoU group with no group-relative GRPO signal. DTP-OPSD addresses this failure by combining GRPO with distractor-aware temporal-privilege on-policy self-distillation. A shared-weight teacher uses training-only target and distractor cues to provide token-level JSD supervision on student-generated prefixes, while an adaptive controller suppresses supervision when the teacher does not outperform the student and strengthens it when the teacher performs better but all student rollout tIoUs are zero. Using only 300 curated examples derived from TimeLens-100K, DTP-OPSD achieves the highest mIoU among the compared open-source 8B models on all three TimeLens benchmarks. Compared with Pure GRPO trained on the same data, it improves all 12 reported metrics, with R1@0.7 gains of 1.28, 1.62, and 1.78 points. These gains across datasets and overlap thresholds show that training-only temporal privilege can complement relative-reward optimization without additional inputs or parameters during inference.

AI USE STATEMENT

This section is **required** and does not count toward the page limit. In this work, we used generative AI tools for initial candidate screening during dataset construction and for assistance with manuscript editing, formatting, and presentation. AI-generated screening outputs were treated only as candidates and were subsequently verified by human annotators. The authors reviewed all AI-assisted work, verified the experimental results and claims, and take full responsibility for the final content of this paper.

ETHICS STATEMENT

This study uses publicly available research datasets and model checkpoints; no private or newly collected personal data are involved, and no identity-related annotations are introduced. We follow the corresponding licenses and usage guidelines. Although this work poses no apparent ethical risks, video models may inherit privacy, demographic, and contextual biases from their source data. Their outputs may also be misapplied in unforeseen contexts. We therefore advocate responsible use and recommend additional privacy, fairness, and safety reviews before deployment in surveillance or other high-impact applications.

REPRODUCIBILITY STATEMENT

We are committed to ensuring the reproducibility of our work. The model initialization, training data, evaluation benchmarks, sampling settings, and hyperparameters are documented in Section 4.1. The method formulation, training algorithm, prompts, reward parser, temporal-privilege rendering procedure, and dataset statistics are provided in the main paper and supplementary material. GPT-5.6 is used only for initial candidate screening, while the resulting annotations are verified by human annotators. We plan to release the verified annotations, training configuration, and evaluation code to support reproduction of our experiments.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, pp. 21246–21263, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/5be69a584901a26c521c2b51e40a4c20-Paper-Conference.pdf.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Zhuo Cao, Heming Du, Bingqing Zhang, Xin Yu, Xue Li, and Sen Wang. When one moment isn't enough: Multi-moment retrieval with cross-moment interactions. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 171705–171723, 2025. doi: 10.52202/085713-5715. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/faaa187af979e46dc54f4d43f8b815b2-Paper-Conference.pdf.

Ruizhe Chen, Tianze Luo, Zhiting Fan, Heqing Zou, Zhaopeng Feng, Guiyang Xie, Hansheng Zhang, Zhuochen Wang, Zuozhu Liu, and Huaijian Zhang. Datasets and recipes for video temporal grounding via reinforcement learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 983–992, 2025. doi: 10.18653/v1/2025.emnlp-industry.66. URL https://aclanthology.org/2025.emnlp-industry.66/.

Zhaoyu Chen, Hongnan Lin, Yongwei Nie, Fei Ma, Xuemiao Xu, Fei Yu, and Chengjiang Long. Invert4TVG: A temporal video grounding framework with inversion tasks preserving action understanding ability. In *International Conference on Learning Representations*, pp. 125494–125513, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/cba6f4460a1f395f68a88598c86e79bd-Paper-Conference.pdf.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948.

Lu Dong, Haiyu Zhang, Han Lin, Ziang Yan, Xiangyu Zeng, Hongjie Zhang, Yifei Huang, Yi Wang, Zhen-Hua Ling, Limin Wang, and Yali Wang. VideoTG-R1: Boosting video temporal grounding via curriculum reinforcement learning on reflected boundary annotations. In *Proceedings of the 2026 International Conference on Multimedia Retrieval*, pp. 1432–1441, 2026. doi: 10.1145/3805622.3810709. URL https://doi.org/10.1145/3805622.3810709.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-R1: Reinforcing video reasoning in MLLMs. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 99114–99137, 2025. doi: 10.52202/085713-3310. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/8eb3976840e08b80dda9667562574246-Paper-Conference.pdf.

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*, pp. 5267–5275, 2017. doi: 10.1109/ICCV.2017.563.

Chaohong Guo, Xun Mo, Yongwei Nie, Fei Ma, Xuemiao Xu, and Chengjiang Long. TAR: Temporal anchor-constrained reasoning for video temporal grounding. *arXiv preprint arXiv:2508.07683*, 2025. URL https://arxiv.org/abs/2508.07683.

Chaohong Guo, Yihan He, Yongwei Nie, Fei Ma, Xuemiao Xu, and Chengjiang Long. T2SGrid: Temporal-to-spatial gridification for video temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3443–3454, 2026.

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*, pp. 5803–5812, 2017. doi: 10.1109/ICCV.2017.618.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

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*, pp. 14271–14280, 2024a.

De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. LITA: Language instructed temporal-localization assistant. In *Computer Vision – ECCV 2024*, pp. 202–218, 2024b. doi: 10.1007/978-3-031-73039-9_12. URL https://doi.org/10.1007/978-3-031-73039-9_12.

Wenxuan Huang, Bohan Jia, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-R1: Incentivizing reasoning capability in multimodal large language models. In *International Conference on Learning Representations*, pp. 63794–63812, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/680b2a8135b9c71278a09cafb605869e-Paper-Conference.pdf.

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*, pp. 706–715, 2017. doi: 10.1109/ICCV.2017.85.

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

Jiaze Li, Hao Yin, Haoran Xu, Boshen Xu, Wenhui Tan, Zewen He, Jianzhong Ju, Zhenbo Luo, and Jian Luan. Video-OPD: Efficient post-training of multimodal large language models for temporal video grounding via on-policy distillation. *arXiv preprint arXiv:2602.02994*, 2026. URL https://arxiv.org/abs/2602.02994.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. VideoChat-R1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025. URL https://arxiv.org/abs/2504.06958.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. ZeRO: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–16, 2020. doi: 10.1109/SC41405.2020.00024.

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*, pp. 14313–14323, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Haibo Wang, Zhiyang Xu, Yu Cheng, Shizhe Diao, Yufan Zhou, Yixin Cao, Qifan Wang, Weifeng Ge, and Lifu Huang. Grounded-VideoLLM: Sharpening fine-grained temporal grounding in video large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 959–975, 2025a. doi: 10.18653/v1/2025.findings-emnlp.50. URL https://aclanthology.org/2025.findings-emnlp.50/.

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, Xiangnan Fang, Zewen He, Zhenbo Luo, Wenxuan Wang, Junqi Lin, Jian Luan, and Qin Jin. Time-R1: Post-training large vision language model for temporal video grounding. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 83330–83364, 2025b. doi: 10.52202/085713-2793. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/7801b29c93b599b8d0c44138596bdeed-Paper-Conference.pdf.

Qi Xu, Yue Tan, Shihao Chen, Jiahao Meng, Anran Wang, Shunping Ji, Hao Fei, and Xiangtai Li. Towards one-to-many temporal grounding. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=hrZCqzmsPW.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-OPD: Learning to see fine details for multimodal LLMs via on-policy self-distillation. *arXiv preprint arXiv:2605.18740*, 2026. URL https://arxiv.org/abs/2605.18740.

Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, Yali Wang, Yu Qiao, and Limin Wang. TimeSuite: Improving MLLMs for long video understanding via grounded tuning. In *International Conference on Learning Representations*, pp. 38057–38081, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/5e8309c9ca683e11672e3dbcd4b87776-Paper-Conference.pdf.

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*, pp. 6543–6554, 2020. doi: 10.18653/v1/2020.acl-main.585. URL https://aclanthology.org/2020.acl-main.585/.

Jun Zhang, Teng Wang, Yuying Ge, Yixiao Ge, Xinhao Li, and Limin Wang. TimeLens: Rethinking video temporal grounding with multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10419–10429, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhang_TimeLens_Rethinking_Video_Temporal_Grounding_with_Multimodal_LLMs_CVPR_2026_paper.html.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026. URL https://arxiv.org/abs/2601.18734.

Qihui Zhu, Yuchen Wang, Zijian Wen, Tao Zhang, Mengjie Zhang, Yang Liu, Shuangwu Chen, Siying Wu, Jian Yang, and Xiaofeng Jiang. RP-OPSD: Resolution-privileged on-policy self-distillation for multimodal large language models. *arXiv preprint arXiv:2607.24447*, 2026. URL https://arxiv.org/abs/2607.24447.

A Implementation and Reproducibility Details

A.1 Training Objective and Update Procedure

For completion token $y_{g,i,t}$, the importance ratio between the current policy and the rollout snapshot is

$$\rho_{g,i,t}(\theta) = \frac{\pi_\theta(y_{g,i,t} \mid V_g, q_g, \hat{y}^S_{g,i,<t})}{\pi_{\theta_{\text{old}}}(y_{g,i,t} \mid V_g, q_g, \hat{y}^S_{g,i,<t})}. \tag{6}$$

Using the group-relative advantage $A_{g,i}$ from Equation 2, define

$$\widetilde{\rho}_{g,i,t}(\theta) = \text{clip}\big(\rho_{g,i,t}(\theta), 1-\epsilon, 1+\epsilon\big). \tag{7}$$

The GRPO loss is

$$\mathcal{L}_{\text{GRPO}} = -\frac{1}{|\mathcal{B}|}\sum_{g\in\mathcal{B}}\frac{1}{G}\sum_{i=1}^{G}\frac{1}{T_{g,i}}\sum_{t=1}^{T_{g,i}} \min\big(\rho_{g,i,t}(\theta)A_{g,i}, \widetilde{\rho}_{g,i,t}(\theta)A_{g,i}\big). \tag{8}$$

Here, $\mathcal{B}$ is a batch of prompt-level rollout groups, $G = 8$ is the number of student completions per prompt, and $T_{g,i}$ is the number of valid non-padding completion tokens. Prompt tokens are excluded, while an emitted end-of-sequence token is included. Each rollout batch is used for a single optimizer update. Therefore, $\theta = \theta_{\text{old}}$ when the gradient is evaluated and the importance ratio equals one at the update point, making clipping inactive in our setting. Equation 8 consequently produces the same update direction as the corresponding group-relative policy-gradient objective. We retain the standard GRPO ratio form to make its relation to the conventional implementation explicit. The reference-policy KL coefficient is $\beta = 0$, so no reference-KL term is added. The GRPO and JSD losses use the same completion-level and token-level normalization before they are combined in Equation 5. Algorithm 1 summarizes the complete training procedure.

Algorithm 1 DTP-OPSD Training with Adaptive Temporal-Privilege Distillation

Input: Training set $\mathcal{D}$, shared policy π_θ, rollout size G, clipping width ϵ, and global coefficient λ_{JSD}.
Output: Optimized clean-video policy π_θ.
Pre-render each privileged view V_g^{priv} by marking the expanded target interval in red and verified distractors in blue.
for each prompt batch $\mathcal{B}$ in one training epoch **do**
  Set the rollout snapshot $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$.
  for each rollout group $g \in \mathcal{B}$ **do**
    Sample G student responses from $\pi_{\theta_{\text{old}}}(\cdot \mid V_g, q_g)$.
    Parse their intervals, compute student tIoU rewards, and obtain the group-relative advantages $\{A_{g,i}\}_{i=1}^{G}$.
    Independently sample G teacher responses from $\pi_{\theta_{\text{old}}}(\cdot \mid V_g^{\text{priv}}, q_g)$ and compute $\bar{r}_g^T$ for the controller.
    for each student response $\hat{y}^S_{g,i}$ and valid token t **do**
      Evaluate the student distribution $p^t_{S,g,i}$ on $(V_g, q_g, \hat{y}^S_{g,i,<t})$ with gradients.
      Evaluate the teacher distribution $p^t_{T,g,i}$ on $(V_g^{\text{priv}}, q_g, \hat{y}^S_{g,i,<t})$ with stop-gradient.
    end for
    Compute the group-level JSD ℓ_g^{JSD} and select $\lambda_g^{\text{adapt}} \in \{0, 1, 2\}$ using Equation 4.
  end for
  Compute $\mathcal{L}_{\text{GRPO}}$ using Equation 8.
  Form $\mathcal{L}_{\text{total}}$ using Equation 5 and update θ once.
end for

For distillation, the teacher does not provide a separate target response. Its independently sampled responses are used only to estimate the teacher reward for the adaptive controller. The teacher distribution used by JSD is instead evaluated on each student-generated prefix.

Each rollout batch is used for one optimizer update and is not reused for inner optimization. Pure GRPO and DTP-OPSD follow the same single-update convention. Teacher distributions are detached, recomputed after every update, and never cached across updates.

A.2 Training Protocol

A global batch contains eight prompt groups, each with eight student completions, giving 64 student completions per update. DTP-OPSD also samples eight teacher completions per prompt for controller estimation. Teacher evaluation on student prefixes is a separate forward pass rather than an additional rollout. The dataloader drops the final incomplete batch. One epoch over the 300-example set therefore contains 37 optimizer updates, processes 296 examples, and omits four examples from the shuffled order. Within every paired run, Pure GRPO and DTP-OPSD use the same order and omit the same examples. Except for the compute-matched control in Section D.2, all methods use one epoch and report the final checkpoint after 37 updates. We do not select an intermediate checkpoint based on benchmark performance. Following the experimental practice of TimeLens (Zhang et al., 2026), we use one shared configuration across the TimeLens benchmarks. For compactness, detailed sensitivity results are shown on QVHighlights-TimeLens. The resulting configuration is applied unchanged to OMTG-Bench and QV-M^2. Unless otherwise specified, results from our post-training experiments and training ablations are reported as the mean over three independent training runs. The same three seeds are used for paired method comparisons. Corresponding standard deviations of the primary mIoU results are reported in Section D.1.

A.3 Memory-Efficient JSD

At a student-prefix position, let p_S and p_T be the clean-video student and privileged-view teacher distributions. We use symmetric JSD with equal mixture weights and temperature 1:

$$q = \mathrm{sg}[p_T], \qquad m = \frac{1}{2}(p_S + q), \tag{9}$$

$$\mathrm{JSD}(p_S, q) = \frac{1}{2}\,\mathrm{KL}(p_S \| m) + \frac{1}{2}\,\mathrm{KL}(q \| m). \tag{10}$$

The teacher distribution is detached, while the mixture retains its dependence on p_S. Thus, both terms contribute gradients to the student, but no gradient flows through the teacher branch.

To reduce memory, let $\mathcal{K}_t$ contain the 100 tokens with the highest teacher probability at position t. For either distribution p, we use the deterministic mapping

$$\Phi_t(p) = \left(\{p(v)\}_{v \in \mathcal{K}_t}, \sum_{v \notin \mathcal{K}_t} p(v) \right). \tag{11}$$

We compute $\mathrm{JSD}(\Phi_t(p_S), \mathrm{sg}[\Phi_t(p_T)])$ over the resulting normalized 101-category distributions. The final category keeps all probability mass outside the teacher top 100, so the retained tokens are not renormalized in isolation. By the data-processing inequality, this coarse-grained JSD is a lower bound on the full-vocabulary JSD and provides a memory-efficient approximation that preserves both the high-probability tokens and the remaining tail mass. Tokens merged into the tail category are constrained through their total probability mass rather than their individual probabilities.

Teacher forward passes use stop-gradient and BF16. Teacher rollout generation, prefix evaluation, and student optimization are performed sequentially, and compressed teacher logits are released immediately. Training uses gradient checkpointing and DeepSpeed ZeRO-2 (Rajbhandari et al., 2020) on eight NVIDIA H20 GPUs.

A.4 Prompt, Reward Parsing, and Video Rendering

The parser extracts the predicted start and end times and computes tIoU with the ground-truth interval. An unparseable interval, a negative timestamp, or an end time earlier than the start time receives zero reward. We use no separate format or temporal-distance reward.

The student and teacher use the same prompt:

```
Please find the visual event described by the sentence
'{query}', determining its starting and ending times.
The format should be:  'The event happens in
<start time> - <end time> seconds'.
```

An expected response is

```
The event happens in 12.0 - 18.5 seconds.
```

The prompt does not name the border colors and requires neither a `<think>` section nor an `<answer>` tag.

Privileged videos are pre-rendered and cached. Each target interval is expanded by $1.5\times$ about its center and clipped to the valid video range, so that the displayed cue does not directly coincide with the annotated boundaries. Frames in this region receive red full-frame borders, while frames in verified distractor intervals receive blue borders; red takes priority in overlap regions. Distractor intervals are not expanded, and the student receives the unchanged video.

A.5 Training Set and Evaluation Protocol

HI-TVG-300 contains 300 query–video pairs from the TimeLens-100K training split, without external videos or queries. Each example has at least two human-verified distractor occurrences from the same video. Candidates are sampled across source datasets and duration ranges. A VLM proposes repeated, similar, related, or opposite occurrences, after which human annotators verify the video, query, target boundaries, and all distractor intervals.

Using the same Qwen3-VL-8B checkpoint and sampling settings, the clean student and privileged teacher each generate eight independent completions during offline filtering. We retain an example when the teacher mean tIoU is above 0.7 and the student mean tIoU is below 0.3. Released annotations include the refined target and all verified distractor intervals.

The human-verified control in Table 4 is sampled after VLM screening and manual verification but before the teacher–student teachability filter. It therefore uses refined target boundaries and verified distractor intervals without filtering examples according to the privileged-teacher advantage.

All methods in the main comparison are evaluated using the corrected TimeLens annotations and the same TimeLens evaluation protocol. Results from our implementation use the same prompt, parser, and tIoU code. Videos are sampled at 2 FPS with at most 448 frames and a visual-token budget of 14,336. For QV-M^2, we evaluate all 323 queries in the official test split and apply no availability-based query filtering.

For the distractor-count analysis in Figure 1, target windows come from the corrected evaluation annotations. Candidate same-video distractors are screened and manually verified before examples are grouped by distractor count. These evaluation-side annotations are used only for analysis and do not affect training or model selection.

The random control is a fixed 300-example draw from the same TimeLens-100K pool. For each query, its original TimeLens target interval is expanded by $1.5\times$ and marked in red. Intervals annotated for other queries associated with the same video are used as candidate blue cues when available. This row uses neither manual target-boundary refinement nor human verification of whether the blue intervals are valid distractors for the current query. It obtains $79.75/70.62/54.90$ for R1@0.3/0.5/0.7 and 65.53 mIoU on QVHighlights-TimeLens. All rows use DTP-OPSD, which isolates the effect of successive stages in the data-curation pipeline under a fixed training method.

B Additional Ablations and Sensitivity

B.1 Contribution of the Training Objectives

All trained variants in Tables 5 and 6 use the same initialization, training data, one-epoch budget, and evaluation protocol. Rows containing privileged JSD use the adaptive controller in Equation 4; the No/No row is the TimeLens-SFT initialization without additional training.

Table 5: Objective components on QVHighlights-TimeLens.

GRPO	Privileged JSD	mIoU
No	No	57.34
Yes	No	65.95
No	Yes	15.04
Yes	Yes	**66.71**

Table 6: Comparison with ground-truth SFT on QVHighlights-TimeLens.

Method	mIoU
Pure GRPO	65.95
GRPO + GT-SFT	66.23
DTP-OPSD	**66.71**

For rows containing privileged JSD, student and teacher rollout rewards are retained for controller estimation even when the GRPO loss is disabled. Privileged JSD is a consistency objective rather than a sequence-level correctness objective. GRPO anchors optimization to the task-level tIoU reward, while JSD provides token-level guidance on student-generated prefixes. The JSD-only result shows that privileged JSD does not provide a sufficient standalone training objective. Combining it with GRPO increases mIoU from 65.95 to 66.71, suggesting that the two objectives are complementary in this setting.

B.2 Comparison with Ground-Truth SFT

We compare DTP-OPSD with a standard direct-supervision alternative. GRPO + GT-SFT jointly optimizes the GRPO objective and a token-level cross-entropy loss on the ground-truth timestamp response for every training example. The SFT coefficient is set to 1. All methods use the same initialization, training data, one-epoch budget, and evaluation protocol. GRPO + GT-SFT reaches 66.23 mIoU, compared with 65.95 for Pure GRPO. Under the same setting, DTP-OPSD reaches 66.71 mIoU, compared with 66.23 for GRPO + GT-SFT.

B.3 Global JSD Weight

We use $\lambda_{\text{JSD}} = 1$ as the shared value across all three TimeLens benchmarks. This choice is not based solely on QVHighlights-TimeLens. For compactness, Figure 5 reports only QVHighlights-TimeLens, where all three R1 metrics peak at 1 and decrease at larger weights.

B.4 Target Expansion and Dead-Group Recovery

Target expansion. We compare three variants that differ only in the duration of the displayed target region. The $1.0\times$ setting displays the annotated boundaries, whereas larger ratios provide progressively coarser temporal cues. Table 8 reports the mean results over the same three training runs. Performance varies by at most 0.35 mIoU across the tested ratios, and we use $1.5\times$ throughout the main experiments.

Dead-group recovery. Using a fixed diagnostic rollout configuration, 142 of the 300 training examples form all-zero groups at initialization: all eight student completions have zero tIoU. We evaluate this fixed subset at the three final checkpoints using the same diagnostic rollout configuration and mark an example as recovered when at least one completion has positive tIoU. The reported results are averaged over the three checkpoints, and all percentages use the same 142 examples as the denominator. The recovery results are reported in Table 9.

Destination of initial dead-group predictions. We further examine all $142 \times 8 = 1{,}136$ student rollouts from the 142 initial dead groups. For each prediction, we compute its maximum tIoU with

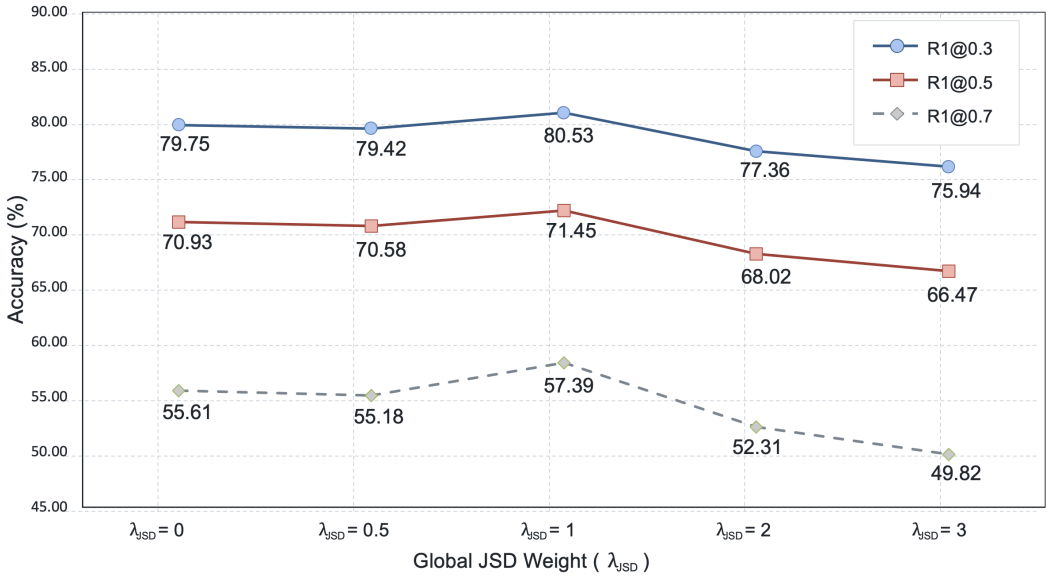

Figure 5: R1 on QVHighlights-TimeLens under different global JSD coefficients. The adaptive controller is enabled for every nonzero coefficient, and the final configuration uses $\lambda_{\mathrm{JSD}} = 1$. Values are means over three training runs.

all human-verified distractor intervals in the same video. A prediction is counted as distractor-overlapping when this maximum tIoU is greater than zero. All 1,136 predictions have zero tIoU with the target interval by construction. Table 7 summarizes the resulting overlap outcomes.

Table 7: Destination of predictions from the 142 initial dead groups. Percentages are computed over all 1,136 student rollouts.

Prediction outcome	Percentage
Overlaps at least one verified distractor	**79.8%**
No verified distractor overlap	20.2%

The high distractor-overlap rate indicates that many all-zero rollouts intersect verified competing occurrences rather than the annotated target. This pattern suggests that incorrect occurrence selection is a common source of all-zero groups in this diagnostic subset.

Table 8: Target-expansion sensitivity on QVHighlights-TimeLens.

Expansion ratio	mIoU
1.0×	66.51
1.5×	**66.71**
2.0×	66.36

Table 9: Recovery of the fixed initial all-zero subset at the final checkpoints.

Method	Initial	Recovered	At least one tIoU ≥ 0.3
Pure GRPO	142	55 (38.7%)	27.5%
DTP-OPSD	142	**91 (64.1%)**	**48.6%**

DTP-OPSD recovers more examples that initially provide no group-relative policy gradient, providing a targeted diagnostic of the failure mode addressed by the method.

C Additional Qualitative Example

Figure 6 shows a same-video interference case. Pure GRPO recognizes the relevant objects and action but predicts 17.0–18.0 seconds, which has no overlap with the ground-truth interval of 21.0–23.0 seconds. DTP-OPSD predicts 20.8–23.2 seconds and reaches a tIoU of 0.83. The improve-

ment comes from selecting the correct occurrence rather than adjusting the boundaries around the same incorrect occurrence.



Figure 6: Qualitative comparison on a video with related actions involving a laptop and a book. Pure GRPO selects an earlier occurrence, while DTP-OPSD recovers the queried occurrence and predicts close boundaries.

D Stability and Efficiency

D.1 Stability Across Seeds

Table 10 reports the mean and sample standard deviation of mIoU over the same three paired training runs. Within each seed pair, Pure GRPO and DTP-OPSD use the same initialization, training data order, and student-rollout seed. Both methods have standard deviations below 0.12 mIoU on all three benchmarks, while DTP-OPSD achieves the higher mean in each case.

Table 10: Mean mIoU and sample standard deviation over three paired training runs.

Method	Charades-TimeLens	ActivityNet-TimeLens	QVHighlights-TimeLens
Pure GRPO	55.43 ± 0.08	53.62 ± 0.10	65.95 ± 0.11
DTP-OPSD	$\mathbf{55.93 \pm 0.06}$	$\mathbf{54.77 \pm 0.07}$	$\mathbf{66.71 \pm 0.09}$

D.2 Compute-Matched Comparison

We measure wall time from the first rollout to the final update. Offline privileged-video rendering takes 0.6 CPU-hours and is excluded from training GPU-hours. GPU-hours equal wall time multiplied by eight GPUs. The compute-matched Pure GRPO run is trained continuously to approximately the same GPU-hour budget as DTP-OPSD, and its final checkpoint is reported without intermediate selection.

As shown in Table 11, additional GRPO computation improves the one-epoch baseline but does not match DTP-OPSD. Under a comparable GPU-hour budget, compute-matched Pure GRPO reaches 66.09 mIoU, compared with 66.71 for DTP-OPSD. Under this matched budget, longer GRPO training narrows but does not close the observed performance gap.

Table 11: Training cost on eight NVIDIA H20 GPUs and accuracy on QVHighlights-TimeLens. Peak memory is reported per GPU. Values are averaged over the same three runs.

Method	Wall time	GPU-hours	Peak memory	mIoU
Pure GRPO, one epoch	7.3 h	58.4	54.6 GB	65.95
Pure GRPO, compute matched	11.7 h	93.6	54.6 GB	66.09
DTP-OPSD, one epoch	11.7 h	93.6	59.1 GB	**66.71**

E Results Across Model Scales

We evaluate DTP-OPSD on the 2B, 4B, and 8B Qwen3-VL backbones. All scales use the same HI-TVG-300 subset selected by the 8B model, without model-specific filtering or example replacement. The teachability thresholds, JSD coefficient, rescue multiplier, and other method-specific settings are unchanged. Before training, 61.7% and 72.3% of the 8B-selected examples still meet the original teachability rule under the 2B and 4B models, respectively. Table 12 reports the corresponding results.

Table 12: Results across model scales using the same 8B-filtered HI-TVG-300 set. Each fixed SFT initialization is evaluated once, while post-training results are averaged over three runs. Comparisons are made within each scale.

Backbone	Method	Charades mIoU	ActivityNet mIoU	QVHighlights mIoU
Qwen3-VL-2B	SFT initialization	42.18	41.37	54.62
Qwen3-VL-2B	Pure GRPO	43.36	42.11	55.80
Qwen3-VL-2B	DTP-OPSD	**43.91**	**42.68**	**56.51**
Qwen3-VL-4B	SFT initialization	47.02	46.11	56.08
Qwen3-VL-4B	Pure GRPO	48.21	47.03	60.36
Qwen3-VL-4B	DTP-OPSD	**48.79**	**47.66**	**61.14**
Qwen3-VL-8B	SFT initialization	51.85	49.26	57.34
Qwen3-VL-8B	Pure GRPO	55.43	53.62	65.95
Qwen3-VL-8B	DTP-OPSD	**55.93**	**54.77**	**66.71**

DTP-OPSD improves over the corresponding SFT initialization and Pure GRPO across the tested model scales. Because the SFT initializations are trained separately, we focus on within-scale comparisons rather than monotonic trends across backbone sizes.

F Limitations

DTP-OPSD focuses on same-video interference. HI-TVG-300 contains only 300 examples from TimeLens-100K and does not cover every temporal grounding error, video domain, or annotation style. All tested backbones also belong to the Qwen3-VL family. The privileged view requires verified target and distractor intervals, which adds data-preparation cost. Teachability filtering favors examples with a clear privileged-teacher advantage and may underrepresent cases in which both views remain uncertain. Expanding the target cue avoids directly displaying the annotated boundaries but does not remove all temporal information leakage. Teacher rollouts and prefix evaluation increase post-training cost. The method also remains more expensive to train than Pure GRPO. The controller relies on Monte Carlo estimates from eight student and eight teacher completions, so its group-level decisions remain subject to sampling variance. The data ablation changes the training subset while keeping DTP-OPSD fixed, so it measures the contribution of the complete curation pipeline rather than independently isolating every curation stage. The privileged cues, teacher branch, and adaptive controller are used only during training. The final model receives clean videos and adds no parameters or computation at inference.