VIDEO-OPD++: GEOMETRY-CALIBRATED CORRECTIVE ON-POLICY DISTILLATION FOR TEMPORAL VIDEO GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation has recently emerged as an effective post-training paradigm for temporal video grounding, combining student-induced training states with dense teacher feedback. However, each realized Video-OPD update remains centered on the token sampled by the student, without explicitly turning teacher-supported timestamp alternatives that could improve localization into training targets. The interval structure of temporal video grounding makes such alternatives directly evaluable: with the suffix fixed, a timestamp can be replaced and the resulting interval scored by temporal IoU (tIoU). Our analysis shows that, for more than three quarters of student trajectories, the teacher support contains an alternative timestamp action that improves tIoU. Motivated by this observation, we propose **Video-OPD++**, a geometry-calibrated corrective on-policy distillation framework. At each student-induced timestamp state, Video-OPD++ evaluates teacher-supported candidate actions using annotation-derived tIoU utilities under fixed-suffix interventions and reweights the teacher prior to construct a corrective distribution aligned with localization quality. This distribution replaces the original local sampled-action reward at timestamp positions, while a discounted future residual preserves feedback from subsequent token discrepancies. Across three temporal video grounding benchmarks, Video-OPD++ improves average mIoU over Video-OPD by 2.3 points. These results suggest that incorporating temporal video grounding geometry into the distillation target helps the student learn more accurate temporal boundaries.

1 INTRODUCTION

Temporal Video Grounding (TVG) aims to localize the temporal segment in a video that corresponds to a natural-language query (Lin et al., 2023; Wu et al., 2025; Zhang et al., 2026). With the development of multimodal large language models (MLLMs), TVG is increasingly formulated as an autoregressive generation task (Pramanick et al., 2025), where a model predicts the start and end timestamps of the target event in textual form. This formulation provides a unified interface for video understanding and temporal localization, but also makes prediction quality sensitive to the states induced by the model's own previous outputs. Effective post-training must therefore provide informative supervision on the states that the model actually encounters during inference.

Existing post-training paradigms address this requirement only partially. Supervised fine-tuning (Li et al., 2025a) provides explicit token-level targets but optimizes the model on ground-truth prefixes, leading to a mismatch between training states and those induced by the model during inference. Reinforcement-learning approaches such as GRPO (Shao et al., 2024; Wang et al., 2026b; Chen et al., 2025) instead optimize trajectories sampled from the current policy, thereby improving training–inference alignment. However, they typically rely on a sequence-level reward and multiple rollouts per instance, resulting in sparse credit assignment and substantial computational overhead.

Video-OPD (Li et al., 2026) provides an effective compromise between these two paradigms. It samples trajectories from the current student policy and employs a fixed teacher to evaluate every student-generated token under the corresponding student-generated prefix. In this way, Video-OPD preserves the student-induced state distribution while replacing sparse trajectory-level rewards with

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.

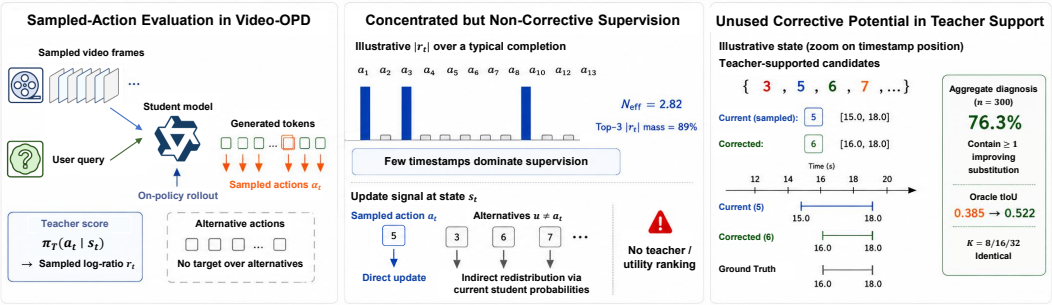


Figure 1: Motivation and mechanism diagnosis of Video-OPD++. Each realized Video-OPD update provides sampled log-ratio feedback only for the realized action, without constructing a target over alternatives. The sampled log-ratio mass is concentrated across completion positions ($N_{\text{eff}} = 2.82$; top-3 mass $= 89\%$), while fixed-suffix analysis shows that teacher support contains an improving substitution in 76.3% of student trajectories, raising oracle tIoU from 0.385 to 0.522. This motivates a task-aware corrective target over teacher-supported timestamp actions.

step-specific teacher feedback. It further requires only one student rollout per training instance, substantially improving the efficiency of post-training for TVG.

Despite these advantages, we identify a limitation in how a sampled Video-OPD update supervises the action space. The log-ratio objective retains teacher guidance in expectation, but at each visited state the realized update assigns a teacher-derived scalar only to the action sampled by the student. The resulting scalar signal can increase or decrease the relative probability of the realized token, but it does not indicate which alternative timestamp token should receive the redistributed probability. Teacher preferences over alternative actions and their effects on the final temporal interval are therefore not explicitly represented as a task-aligned distribution in that update. Video-OPD determines *where* the student should learn—at states induced by its own policy—but does not explicitly determine *what the student should learn instead* when its current temporal action is suboptimal. We refer to this limitation as a *step-wise dense but action-wise pointwise* objective. TVG makes this limitation both consequential and tractable. Its realized OPD signal is concentrated across a few completion positions, while a token substitution can be evaluated deterministically by parsing the modified answer into a temporal interval. On 300 student trajectories sampled under the training decoding policy, a fixed-suffix oracle finds an improving teacher-supported action in 76.3% of cases and raises mean tIoU from 0.385 to 0.522. Useful alternatives therefore already exist near the student trajectory; what is missing is a task-aware target over them.

Motivated by this observation, we propose **Video-OPD++**, a geometry-calibrated corrective on-policy distillation framework for TVG. Video-OPD++ preserves the student-induced state distribution of Video-OPD, while replacing pointwise sampled-action supervision at timestamp positions with an explicit corrective distribution. At each timestamp state, it first extracts a compact set of teacher-supported actions. Each candidate is then substituted into the student completion while the remaining suffix is kept fixed, producing a candidate interval whose temporal utility can be computed directly. Video-OPD++ uses these utilities to reweight the teacher distribution, assigning greater target probability to actions that are both supported by the teacher and more consistent with the annotated temporal interval. Video-OPD++ further separates the responsibilities of local action correction and sequential feedback. At timestamp positions, the geometry-calibrated distribution replaces the original local sampled-action signal and directly specifies the desired action distribution. Discounted feedback from subsequent token discrepancies is retained to preserve dependencies within autoregressive timestamp generation. At non-timestamp positions, the original local Video-OPD objective remains unchanged. This design avoids imposing both pointwise evaluation and distributional correction on the same timestamp decision.

Experiments on Charades-TimeLens, ActivityNet-TimeLens, and QVHighlights-TimeLens (Zhang et al., 2026) show that Video-OPD++ improves mIoU over Video-OPD across all three benchmarks. Video-OPD++ achieves 52.6, 50.6, and 64.0 mIoU, improving the corresponding Video-OPD results by 0.6, 3.3, and 3.0 points. The average mIoU increases from 53.4 to 55.7. On ActivityNet and QVHighlights, the gains hold across all three IoU thresholds, indicating improvements in both coarse temporal coverage and precise localization. Controlled ablations further show that directly matching the teacher support provides only marginal improvements, whereas temporal geometry

calibration substantially increases performance. Removing the discounted future feedback or retaining the original sampled-action signal at timestamp positions yields lower mIoU on all benchmarks.

Our contributions are summarized as follows:

- We identify a sampled-action correction bottleneck in Video-OPD. Although it provides dense feedback across student-generated steps, each realized update remains centered on the sampled action. Fixed-suffix intervention analysis shows that useful teacher-supported alternatives often exist along sampled trajectories.
- We propose Video-OPD++, which constructs a geometry-calibrated corrective distribution over teacher-supported timestamp actions at student-induced states. The method converts pointwise action evaluation into task-aligned corrective supervision.
- We introduce a unified token-partitioned objective in which corrective supervision replaces the original local signal at timestamp positions, discounted future feedback retains information from subsequent token discrepancies, and non-timestamp positions retain vanilla Video-OPD. Video-OPD++ improves average mIoU by 2.3 points across three standard TVG benchmarks.

2 Background and Diagnostic Analysis

2.1 On-Policy Distillation for Temporal Video Grounding

Given a video–query pair (v, q), the rollout policy $\pi_{\theta_{\text{old}}}$ generates a trajectory $\tau = (a_1, \ldots, a_T)$ and visits the student-induced states $s_t = (v, q, a_{<t})$. Video-OPD (Li et al., 2026) uses a fixed teacher π_T to evaluate each sampled action a_t under the same student-generated prefix. Its token-level reward is

$$r_t = \log \pi_T(a_t \mid s_t) - \log \pi_{\theta_{\text{old}}}(a_t \mid s_t). \tag{1}$$

For a fixed state, the expectation of r_t under the student action distribution equals the negative conditional reverse KL. Each realized r_t, however, is a signed sampled log-density ratio attached only to the action observed in the rollout.

Let $\rho_t(\theta) = \pi_\theta(a_t \mid s_t)/\pi_{\theta_{\text{old}}}(a_t \mid s_t)$ denote the importance ratio, m_t the completion mask, and $\text{sg}(\cdot)$ the stop-gradient operator. The core Video-OPD surrogate is

$$\mathcal{L}_{\text{OPD}}(\theta) = -\mathbb{E}_{\tau \sim \pi_{\theta_{\text{old}}}} \left[\sum_{t=1}^{T} m_t \rho_t(\theta)\, \text{sg}(r_t) \right]. \tag{2}$$

This objective preserves the state distribution induced by the student while providing a distinct teacher-derived scalar at every generation step.

2.2 The Sampled-Action Correction Bottleneck

Although Equation 2 provides step-specific feedback, its supervision remains pointwise over the action space. For the per-token term $\ell_t = -m_t \rho_t(\theta)\, \text{sg}(r_t)$, the gradient-descent direction for an arbitrary vocabulary logit is

$$-\frac{\partial \ell_t}{\partial z_\theta(u \mid s_t)} = m_t \rho_t(\theta)\, \text{sg}(r_t) \left[\mathbb{I}(u = a_t) - \pi_\theta(u \mid s_t) \right]. \tag{3}$$

The teacher-dependent scalar distinguishes the sampled action a_t from its aggregate complement in each realized update, but does not rank alternatives within that complement. Their relative updates are inherited from the student's current distribution rather than determined by the teacher's relative preference or temporal grounding utility. At the rollout policy, averaging over $a_t \sim \pi_{\theta_{\text{old}}}(\cdot \mid s_t)$ recovers the local reverse-KL gradient. The correction bottleneck therefore concerns the realized Monte Carlo update, in which the local alternative structure remains implicit rather than represented as an explicit task-aware target. Video-OPD determines *where* the student should learn at states induced by its own policy, without explicitly determining *what it should learn instead* in that update.

2.3 EMPIRICAL DIAGNOSIS

Diagnostic protocol. All analyses in this section are conducted on 300 examples randomly sampled with seed 0 from the same 2.5K TVDF-selected training subset used in all post-training runs. We analyze our reproduced single-round Video-OPD student, initialized from Qwen3-VL-8B-Instruct, together with the frozen GRPO-trained Qwen3-VL-32B teacher. Videos are sampled at 2 FPS. For each example, the student draws one completion from its current policy using the exact decoding configuration adopted during training: temperature 0.7, top-k 20, and top-p 0.8. We fix the random seed for reproducibility and use the same no-thinking training prompt. The teacher does not generate a separate trajectory; instead, it evaluates the exact sampled student completion under the same student-generated prefixes. Student and teacher log-probabilities are aligned over valid completion positions using the same completion mask as in Video-OPD training. Consequently, both the analyzed states and the sampled log-ratio signals match those encountered by the training objective rather than a greedy approximation.

Concentrated supervision. We quantify the effective number of supervised positions in each trajectory using the inverse participation ratio

$$N_{\text{eff}}(\tau) = \frac{\left(\sum_{t:m_t=1} |r_t|\right)^2}{\sum_{t:m_t=1} r_t^2}. \tag{4}$$

We additionally measure the top-three mass independently for each trajectory by sorting all valid completion positions according to $|r_t|$ and computing the fraction of $\sum_t |r_t|$ contributed by the three largest positions. We report the mean of these trajectory-level statistics over the diagnostic set.

As shown in Figure 1, the mean effective token count is 2.82, and the three largest completion-position contributions account for approximately 89% of the absolute sampled log-ratio mass. These statistics show concentration across completion positions, but do not identify token types or establish timestamp dominance. They motivate the separate analysis of timestamp alternatives below.

Corrective potential in teacher support. To determine whether useful alternative actions are available at the visited states, we construct a fixed-suffix token-correction oracle. At each timestamp position, we extract the teacher's top-K candidate actions, replace one token at a time, and keep all remaining student token IDs unchanged. Each modified completion is decoded and parsed into a temporal interval, while candidates that do not produce a valid interval are excluded. The trajectory-level oracle selects the highest tIoU among the original student output and all valid single-token substitutions. Correction coverage is defined as the fraction of trajectories for which this oracle tIoU strictly exceeds that of the original student output.

With $K = 32$, the sampled student trajectories obtain an average tIoU of 0.385. At least one valid teacher-supported substitution improves the original prediction for 76.3% of the trajectories, and selecting the best fixed-suffix substitution increases average oracle tIoU to 0.522, corresponding to an absolute oracle gain of 0.136. Improving substitutions are found at start- and end-timestamp positions for 62.0% and 72.7% of trajectories, respectively. All oracle statistics are identical for $K \in \{8, 16, 32\}$, showing that the useful local action support is already saturated with a small number of high-probability teacher candidates. Additional diagnostic analysis is provided in Appendix B.

These results show that useful teacher-supported local substitutions often exist in the analyzed trajectories. The oracle uses the annotation to select the best valid single-token substitution and thus measures local corrective potential under the fixed-suffix protocol. This observation directly motivates the geometry-calibrated corrective distribution introduced in Section 3.

3 VIDEO-OPD++

Video-OPD++ converts the unused alternatives identified above into an explicit corrective target at student-induced timestamp states. Its training process consists of three successive operations: extracting teacher-supported actions, evaluating them through fixed-suffix temporal interventions, and calibrating the teacher distribution according to their task utility.

Unlike vanilla Video-OPD, Video-OPD++ uses the annotated temporal interval as an action-level utility signal. The annotation is never used as a demonstration prefix: all states remain induced by

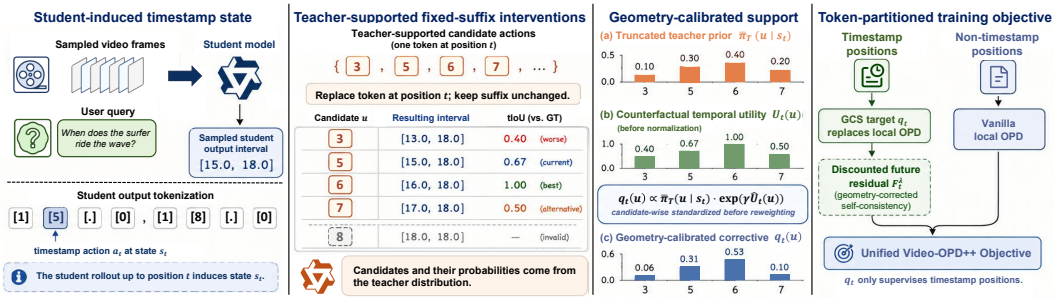


Figure 2: Overview of Video-OPD++. Given a timestamp state induced by the student rollout, we extract teacher-supported candidate actions and evaluate them through fixed-suffix temporal interventions. Their temporal utilities calibrate the truncated teacher prior into a corrective target distribution. At timestamp positions, this distribution replaces the local sampled-action reward while a discounted future residual preserves sequential feedback.

the student trajectory, and the annotation only evaluates local alternatives at those states. Video-OPD++ is therefore an annotation-calibrated form of on-policy *state* distillation rather than teacher forcing on a ground-truth trajectory.

3.1 Teacher-Supported Fixed-Suffix Interventions

Consider a student trajectory $\tau = (a_1, \ldots, a_T)$ with states $s_t = (v, q, a_{<t})$. Let $\mathcal{I}_{\text{ts}}(\tau)$ denote the positions whose decoded spans belong to the generated start or end timestamp. At every $t \in \mathcal{I}_{\text{ts}}(\tau)$, we extract a compact candidate set from the teacher distribution:

$$\mathcal{C}_t = \text{TopK}_K\left(\pi_T(\cdot \mid s_t)\right). \tag{5}$$

The candidate set contains only the teacher's top-K actions. It is not merged with the student top-K set, and the sampled action a_t is not explicitly inserted. Student log-probabilities are subsequently gathered at the same teacher-proposed token indices.

For each candidate $u \in \mathcal{C}_t$, we replace the action at position t while retaining the original student suffix:

$$\tau^{t \leftarrow u} = (a_{<t}, u, a_{>t}), \qquad b_t(u) = \text{Parse}\left(\tau^{t \leftarrow u}\right), \qquad U_t(u) = \text{tIoU}\left(b_t(u), b^\star\right), \tag{6}$$

where $b^\star$ is the annotated temporal interval. We retain the candidates that can be parsed into valid temporal intervals and denote the resulting set by $\mathcal{C}_t^{\text{val}} \subseteq \mathcal{C}_t$.

Equation 6 does not regenerate the actions following position t. It therefore defines an exactly computable local temporal utility under a fixed student suffix, rather than a full counterfactual return under the autoregressive policy. This structure allows Video-OPD++ to evaluate multiple local alternatives without additional candidate rollouts or a learned outcome verifier.

3.2 Geometry-Calibrated Support

The teacher support defines a plausible local action space, but its probabilities do not necessarily reflect the temporal grounding quality of the candidates. We therefore standardize the utilities within each valid support:

$$\widehat{U}_t(u) = \frac{U_t(u) - \mu_t}{\sigma_t + \epsilon}, \qquad u \in \mathcal{C}_t^{\text{val}}, \tag{7}$$

where μ_t and σ_t are respectively the mean and standard deviation of $\{U_t(u)\}_{u \in \mathcal{C}_t^{\text{val}}}$. Thus, $\widehat{U}_t(u)$ represents candidate-relative temporal utility within the local teacher support.

Let $\bar{\pi}_T(\cdot \mid s_t)$ denote the teacher distribution renormalized over $\mathcal{C}_t^{\text{val}}$. We construct the geometry-calibrated corrective target

$$q_t(u) = \frac{\bar{\pi}_T(u \mid s_t) \exp\left(\gamma \widehat{U}_t(u)\right)}{\sum_{w \in \mathcal{C}_t^{\text{val}}} \bar{\pi}_T(w \mid s_t) \exp\left(\gamma \widehat{U}_t(w)\right)}. \tag{8}$$

The teacher prior restricts the target to plausible actions and sets their baseline probability ratios, while temporal utility reweights that prior. When $\gamma = 0$, the target reduces to the truncated teacher distribution. For $\gamma > 0$, actions with greater candidate-relative temporal utility receive a larger multiplicative boost relative to their teacher prior; the final probabilities depend on both terms.

Rather than selecting a single hard target, Equation 8 preserves a soft distribution over multiple plausible alternatives. The student matches this distribution through the Geometry-Calibrated Support loss

$$\mathcal{L}_{\mathrm{GCS}}(\theta) = -\frac{1}{Z_{\mathrm{ts}}}\mathbb{E}_{\tau\sim\pi_{\theta_{\mathrm{old}}}}\left[\sum_{t\in\mathcal{I}_{\mathrm{ts}}(\tau)} m_t \sum_{u\in\mathcal{C}_t^{\mathrm{val}}} q_t(u)\log\pi_\theta(u \mid s_t)\right], \tag{9}$$

where Z_{ts} normalizes over timestamp positions with valid candidate supports. The teacher probabilities and temporal utilities are detached, and gradients propagate only through the student distribution.

Unlike the sampled-action reward, GCS specifies a distributional corrective target over the valid teacher support. Candidate actions assigned greater target probability than student probability are promoted, candidates assigned lower target probability are suppressed, and actions outside the valid teacher support are reduced through full-vocabulary softmax normalization.

KL-regularized interpretation. The target in Equation 8 is the solution to the local optimization problem

$$q_t^\star = \arg\max_{q\in\Delta(\mathcal{C}_t^{\mathrm{val}})} \left\{\gamma\mathbb{E}_{u\sim q}\left[\widehat{U}_t(u)\right] - D_{\mathrm{KL}}\left(q \,\|\, \bar{\pi}_T(\cdot \mid s_t)\right)\right\}. \tag{10}$$

Its closed-form optimizer is precisely Equation 8. GCS can therefore be viewed as teacher-regularized utility reweighting within a local action support. This interpretation does not imply monotonic improvement of the full autoregressive policy, since both the candidate space and the fixed-suffix utility are local. For detailed derivations, please refer to Appendix C.3.

3.3 Token-Partitioned Training Objective

GCS specifies the desired action distribution at the current timestamp state. Autoregressive timestamp generation, however, also contains dependencies on subsequent token decisions. Video-OPD++ therefore retains discounted feedback from later token discrepancies while removing the original local sampled-action reward at timestamp positions.

We define the discounted future residual

$$F_t^\lambda = \sum_{k=t+1}^{T} \lambda^{k-t} r_k, \qquad \lambda \in [0,1]. \tag{11}$$

This residual includes feedback from later digits within the same timestamp as well as from subsequent timestamp decisions; it is therefore a general sequential residual rather than an exclusively start-to-end signal.

Let $h_t = \mathbb{I}[t \in \mathcal{I}_{\mathrm{ts}}(\tau)]$. The final objective is

$$\mathcal{L}_{\text{Video-OPD++}}(\theta) = \mathcal{L}_{\mathrm{GCS}}(\theta) - \mathbb{E}_{\tau\sim\pi_{\theta_{\mathrm{old}}}}\left[\sum_{t=1}^{T} m_t \rho_t(\theta)\,\mathrm{sg}\left((1-h_t)r_t + \eta h_t F_t^\lambda\right)\right], \tag{12}$$

where η controls the contribution of the future residual.

For non-timestamp positions ($h_t = 0$), the objective reduces to the original local Video-OPD surrogate. For timestamp positions ($h_t = 1$), GCS replaces the local sampled-action reward and directly supervises the current action distribution, while F_t^λ retains discounted feedback from subsequent discrepancies. This partition removes only the local sampled-action reward r_t at timestamp positions—replacing it with the distributional GCS target—while retaining the pointwise channel to carry the discounted future residual F_t^λ.

Video-OPD++ reuses the teacher and student logits already computed for on-policy distillation. Candidate evaluation requires only token substitution, temporal-interval parsing, and tIoU computation, adding neither autoregressive candidate rollouts nor inference-time computation. For additional method details, please refer to Appendix C.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Training. We follow the official Video-OPD protocol (Li et al., 2026), using Qwen3-VL-8B-Instruct (Bai et al., 2025a) as the student, a frozen GRPO-trained Qwen3-VL-32B model as the teacher, and the same 2,500 TVDF-selected training examples. We train for one epoch with a learning rate of 1×10^{-6}, a global batch size of 32, and one on-policy student rollout per example. Videos are sampled at 2 FPS. For Video-OPD++, we use a teacher top-K support with $K = 32$, geometry coefficient $\gamma = 4.0$, target temperature $\tau = 1$, future-discount factor $\lambda = 0.25$, and residual weight $\eta = 0.25$. The GCS loss is assigned unit weight. At timestamp positions, GCS replaces the local sampled-action reward while the policy-gradient branch retains only the discounted future residual; non-timestamp positions retain the vanilla Video-OPD objective. Full sequence, rollout, and implementation details are provided in Appendix D.1.

Evaluation. We evaluate on Charades-TimeLens, ActivityNet-TimeLens, and QVHighlights-TimeLens using the corrected temporal annotations from TimeLens (Zhang et al., 2026). Following Video-OPD (Li et al., 2026), we report R@0.3, R@0.5, R@0.7, and mean temporal IoU (mIoU). All variants use the same inference prompt, decoding configuration, temporal-interval parser, and evaluation scripts. Further evaluation details are given in Appendix D.2.

Baselines. We compare against three groups of methods. Proprietary MLLMs include GPT-4o (Hurst et al., 2024), GPT-5 (Singh et al., 2025), and the Gemini-2.0/2.5 series (Comanici et al., 2025). Open-source video MLLMs include VideoChat-Flash-7B (Li et al., 2025b), Qwen2.5-VL-7B (Bai et al., 2025b), VideoChat-R1-7B (Li et al., 2025c), Time-R1-7B (Wang et al., 2026b), TVG-R1-7B (Chen et al., 2025), VideoChat-R1.5-7B (Yan et al., 2025), MiMo-VL-7B (Xiaomi & Team, 2025), and Qwen3-VL-8B-Instruct (Bai et al., 2025a). We further compare with Qwen3-VL-8B post-training frameworks, including OP-RKD, OP-FKD, GRPO (Shao et al., 2024), and Video-OPD (Li et al., 2026). More details are provided in Appendix D.3.

Table 1: Evaluation of Video-OPD++ on three TVG benchmarks. Among all open-source and post-training systems, **bold** values indicate the best performance.

Models for Evaluation	Charades-TimeLens			ActivityNet-TimeLens			QVHighlights-TimeLens		
	R@0.3	R@0.5	R@0.7	R@0.3	R@0.5	R@0.7	R@0.3	R@0.5	R@0.7
Proprietary Models									
GPT-4o	60.6	44.5	23.5	55.2	41.4	25.8	69.0	54.8	38.5
GPT-5	59.3	42.0	22.0	57.4	44.9	30.4	72.4	60.4	46.4
Gemini-2.0-Flash	66.4	53.5	27.1	62.9	54.0	37.7	76.2	66.4	48.3
Gemini-2.5-Flash	68.7	56.1	30.6	66.8	57.5	41.3	78.2	69.4	55.0
Gemini-2.5-Pro	74.1	61.1	34.0	72.3	64.2	47.1	84.1	75.9	61.1
Open-Source Models									
VideoChat-Flash-7B	60.2	37.9	17.8	35.5	21.8	10.5	45.2	30.6	16.7
Qwen2.5-VL-7B	58.1	35.1	18.2	47.2	32.5	20.2	55.0	41.7	29.3
VideoChat-R1-7B	51.9	30.8	11.7	35.0	23.9	11.3	29.3	19.1	9.4
Time-R1-7B	57.9	32.0	16.9	44.8	31.0	19.0	65.8	51.5	36.1
TVG-R1-7B	44.5	23.6	12.4	46.7	31.0	18.6	55.8	41.2	28.0
VideoChat-R1.5-7B	46.4	24.0	10.4	40.6	25.3	16.4	62.2	44.5	28.3
MiMo-VL-7B	57.9	42.6	20.5	49.3	38.7	22.4	57.1	42.6	28.4
Qwen3-VL-8B-Instruct	61.7	41.5	23.1	41.2	30.7	20.0	46.6	38.2	29.5
Post-Training Frameworks									
OP-RKD	67.3	47.0	27.8	59.1	44.9	30.6	70.7	57.4	44.5
OP-FKD	66.9	**47.3**	27.6	58.6	44.2	30.5	71.2	58.9	46.5
GRPO	72.7	44.4	27.6	58.6	42.7	32.1	69.8	53.0	41.5
Video-OPD	73.1	45.8	**32.4**	60.5	45.6	35.8	73.8	60.3	50.4
Video-OPD++	**74.0**	**47.3**	32.2	**64.1**	**49.0**	**38.4**	**77.2**	**64.8**	**53.7**

4.2 MAIN RESULTS

Comparison with existing methods. As shown in Table 1, Video-OPD++ achieves the strongest overall performance among the compared open-source and post-training systems. It obtains the best result on eight of the nine reported recall metrics. The gains are particularly pronounced

on ActivityNet-TimeLens and QVHighlights-TimeLens, where Video-OPD++ consistently outperforms all Qwen3-VL-8B post-training baselines across the three IoU thresholds. It also surpasses GPT-5 on every reported TVG metric, demonstrating that geometry-calibrated on-policy distillation can substantially improve a compact open-source student.

Consistent improvements over Video-OPD. Video-OPD++ improves upon Video-OPD in mIoU across all three benchmarks. As further summarized by the mIoU comparison in Figure 3, the gains are 0.6, 3.3, and 3.0 points on Charades-TimeLens, ActivityNet-TimeLens, and QVHighlights-TimeLens, respectively, raising the average mIoU from 53.4 to 55.7. On ActivityNet and QVHighlights, the improvements hold at every IoU threshold, including the strict R@0.7 criterion, indicating that the method improves both coarse temporal coverage and precise localization. On Charades, Video-OPD++ improves R@0.3, R@0.5, and mIoU, while incurring only a marginal reduction at R@0.7. The overall gains are therefore not attributable to a single dataset or evaluation threshold.

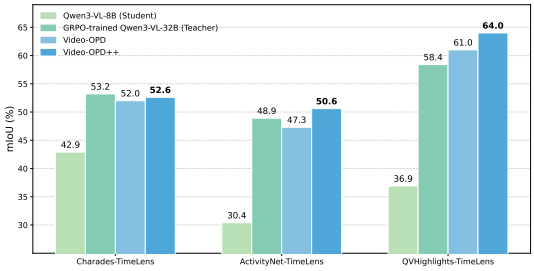


Figure 3: mIoU comparison among the Qwen3-VL-8B student, the GRPO-trained Qwen3-VL-32B teacher, Video-OPD, and Video-OPD++ on three TimeLens benchmarks.

Beyond direct teacher imitation. Figure 3 also compares the 8B student with the GRPO-trained 32B teacher. Video-OPD++ surpasses the teacher in mIoU on ActivityNet and QVHighlights, while remaining close on Charades. This behavior is consistent with the design of Video-OPD++: the teacher defines a plausible local action support, but the student is not trained to reproduce the teacher distribution directly. Instead, annotation-derived temporal utilities reweight the teacher-supported actions according to their grounding quality. The resulting target can therefore preserve the teacher's action prior while correcting temporal preferences that are suboptimal for the downstream metric.

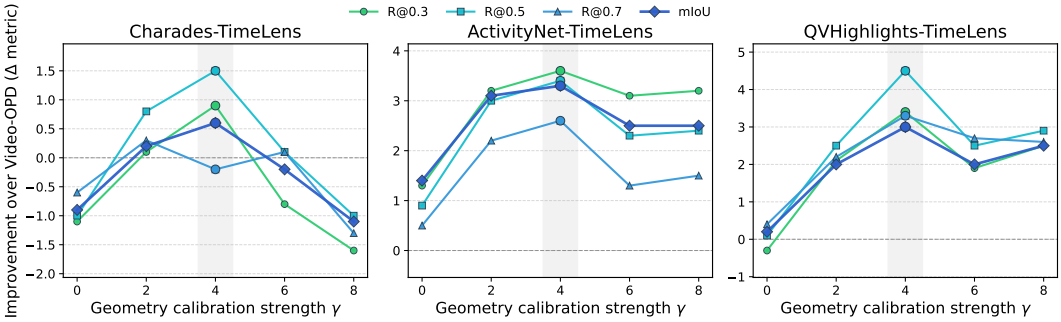


Figure 4: Effect of the geometry-calibration strength γ, measured as absolute gains over Video-OPD. $\gamma = 0$ denotes Support-only, and the shaded region marks our main setting $\gamma = 4$. Geometry calibration consistently improves over Support-only, with $\gamma = 4$ achieving the best overall trade-off; stronger calibration leads to diminishing gains.

4.3 Ablation Studies

Geometry calibration is the primary performance lever. Figure 4 studies the geometry strength γ, where $\gamma = 0$ reduces the target to the truncated teacher distribution and therefore corresponds to Support-only. Support-only yields inconsistent improvements: it degrades performance on Charades and provides only limited gains on QVHighlights, showing that exposing additional teacher-supported actions is insufficient by itself. Introducing temporal-utility calibration substantially improves the results, and $\gamma = 4$ achieves the highest mIoU on all three benchmarks as well as the strongest overall gains across recall thresholds. Increasing γ further consistently weakens performance, suggesting that excessive utility reweighting can overwhelm the teacher prior. These results support the intended balance between teacher plausibility and temporal geometry. Across the three benchmarks, Support-only improves average mIoU by only 0.2 points over Video-OPD, whereas geometry calibration with $\gamma = 4$ yields a further 2.1-point gain. This decomposition shows that the

dominant benefit comes from task-aware temporal reweighting rather than from merely exposing the student to a broader teacher-supported action set.

The future residual preserves sequential feedback. Table 2 evaluates the discounted future residual by setting $\eta = 0$. Removing this term consistently reduces mIoU on all three benchmarks, with the largest degradation occurring on QVHighlights. Nevertheless, GCS without the future residual still outperforms Video-OPD on ActivityNet and QVHighlights, confirming that the geometry-calibrated corrective distribution is itself an effective source of supervision. The additional gains from the full objective show that local action correction and feedback from subsequent timestamp decisions are complementary rather than redundant. Aggregated across the three datasets, the future residual contributes an additional 0.9 mIoU beyond GCS alone, confirming that local action correction and sequential feedback remain complementary.

Table 2: Component ablation of the token-partitioned training objective on three TVG benchmarks. The teacher and base model are included as references. **Bold** values indicate the best performance.

Models for Evaluation	Charades-TimeLens				ActivityNet-TimeLens				QVHighlights-TimeLens			
	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU
Qwen3-VL-32B-GRPO	78.7	49.4	29.4	53.2	65.0	47.5	35.0	48.9	75.6	58.2	44.9	58.4
Qwen3-VL-8B-Instruct	61.7	41.5	23.1	42.9	41.2	30.7	20.0	30.4	46.6	38.2	29.5	36.9
Video-OPD	73.1	45.8	32.4	52.0	60.5	45.6	35.8	47.3	73.8	60.3	50.4	61.0
GCS w/o future ($\eta = 0$)	72.4	46.1	**32.9**	52.0	63.3	48.7	37.6	50.0	75.5	62.6	53.0	62.6
GCS + local sampled reward	73.5	46.2	32.5	52.1	62.3	47.2	37.0	49.1	73.5	60.2	51.1	61.0
Video-OPD++ ($\eta = 0.25$)	**74.0**	**47.3**	32.2	**52.6**	**64.1**	**49.0**	**38.4**	**50.6**	**77.2**	**64.8**	**53.7**	**64.0**

Corrective supervision should replace local sampled-action feedback. Retaining the original sampled-action reward at timestamp positions and treating GCS as an auxiliary loss consistently underperforms the replacement-based Video-OPD++ objective. The degradation is especially substantial on QVHighlights, where augmentation removes the improvement over Video-OPD, whereas the replacement objective yields a gain of more than three mIoU points. This result supports the responsibility decomposition in Section 3.3: GCS should determine the current timestamp action, while the policy-gradient branch should retain only future-dependent feedback. Applying both the pointwise sampled-action signal and the corrective distribution to the same decision introduces competing optimization directions.

Overall contribution of the unified objective. The complete Video-OPD++ objective achieves the best result on 11 of the 12 metrics among Video-OPD and its controlled variants; the only exception is Charades R@0.7, where the variant without the future residual is slightly higher. Taken together, the ablations establish a clear causal chain: teacher support makes plausible alternatives available, geometry calibration assigns them task-aligned preferences, replacement avoids conflict with the original local signal, and the discounted future residual preserves dependencies across autoregressive timestamp decisions. Detailed results of geometry-strength ablation are provided in Appendix E.

5 Conclusion

We identify a sampled-action correction bottleneck in Video-OPD: although it provides dense feedback along student-generated trajectories, each realized update remains centered on the sampled action rather than an explicit task-aware target over plausible alternatives. To address this limitation, we introduce Video-OPD++, which evaluates teacher-supported timestamp actions through fixed-suffix temporal interventions and uses their annotation-derived utilities to construct a geometry-calibrated corrective distribution. A token-partitioned objective replaces the local sampled-action reward with this distribution at timestamp positions, while retaining discounted future feedback and vanilla Video-OPD supervision elsewhere. Across Charades-TimeLens, ActivityNet-TimeLens, and QVHighlights-TimeLens, Video-OPD++ improves average mIoU by 2.3 points over Video-OPD. Controlled ablations further show that geometry calibration is the primary source of improvement, that discounted future feedback remains complementary to local action correction, and that corrective supervision is more effective when it replaces rather than augments the original timestamp reward. These results demonstrate the value of moving from pointwise action evaluation to task-aligned corrective distillation for temporal video grounding, without additional autoregressive candidate rollouts or inference-time computation.

THE USE OF LARGE LANGUAGE MODELS (LLMS)

Large Language Models (LLMs) are employed solely to assist with language refinement and stylistic polishing of the manuscript. They are not involved in research ideation, experimental design and analysis. All conceptual and technical contributions are the sole responsibility of the authors.

ETHICS STATEMENT

Our work aims to improve the post-training of multimodal large language models for temporal video grounding without introducing new data collection, human-subject experiments, or deployment scenarios. It therefore does not introduce task-specific ethical concerns beyond those generally associated with video-language models.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we follow the official Video-OPD training and evaluation protocol and keep the student and teacher models, the 2,500 TVDF-selected training examples, the post-training budget, rollout configuration, and evaluation procedure fixed across controlled comparisons. The main text and appendix specify all model and optimization settings, including the complete Video-OPD++ hyperparameters ($K = 32$, $\gamma = 4.0$, $\tau = 1$, $\lambda = 0.25$, and $\eta = 0.25$), the diagnostic sampling protocol, fixed-suffix candidate construction, candidate-validity rules, timestamp parsing procedure, and complete ablation results. All controlled variants use the same initialization, data order, prompt, and rollout budget. We will release the implementation, configuration files, data indices, and evaluation scripts upon publication.

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*, volume 2024, pp. 21246–21263, 2024.

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.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Peijun Bao, Yong Xia, Wenhan Yang, Boon Poh Ng, Meng Hwa Er, and Alex C Kot. Local-global multi-modal distillation for weakly-supervised temporal video grounding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 738–746, 2024.

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*, pp. 961–970, 2015.

Ruizhe Chen, Tianze Luo, Zhiting Fan, Heqing Zou, Zhaopeng Feng, Guiyang Xie, Hansheng Zhang, Zhuochen Wang, Zuozhu Liu, and Zhang Huaijian. 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.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

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.

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, 2024.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Alex Jinpeng Wang, Linjie Li, Kevin Qinghong Lin, Jianfeng Wang, Kevin Lin, Zhengyuan Yang, Lijuan Wang, and Mike Zheng Shou. Cosmo: Contrastive streamlined multimodal model with interleaved pre-training. *arXiv e-prints*, pp. arXiv–2401, 2024.

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

Hongyu Li, Jinyu Chen, Ziyu Wei, Shaofei Huang, Tianrui Hui, Jialin Gao, Xiaoming Wei, and Si Liu. Llava-st: A multimodal large language model for fine-grained spatial-temporal understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8592–8603. IEEE, 2025a.

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.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, Yu Qiao, Yali Wang, and Limin Wang. Videochat-flash: Hierarchical compression for long-context video modeling, 2025b. URL https://arxiv.org/abs/2501.00574.

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*, 2025c.

Renjie Liang, Yiming Yang, Hui Lu, and Li Li. Efficient temporal sentence grounding in videos with multi-teacher knowledge distillation. *arXiv preprint arXiv:2308.03725*, 2023.

Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. Univtg: Towards unified video-language temporal grounding. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2782–2792. IEEE, 2023.

Daizong Liu, Xiaoye Qu, Yinzhen Wang, Xing Di, Kai Zou, Yu Cheng, Zichuan Xu, and Pan Zhou. Unsupervised temporal video grounding with deep semantic clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 1683–1691, 2022.

Andreea-Maria Oncescu, Joao F Henriques, Yang Liu, Andrew Zisserman, and Samuel Albanie. Queryd: A video dataset with high-quality text and audio narrations. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2265–2269. IEEE, 2021.

Shraman Pramanick, Effrosyni Mavroudi, Yale Song, Rama Chellappa, Lorenzo Torresani, and Triantafyllos Afouras. Enrich and detect: Video temporal grounding with multimodal llms. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24297–24308. IEEE, 2025.

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

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

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Rui Su, Qian Yu, and Dong Xu. Stvgbert: A visual-linguistic transformer based framework for spatio-temporal video grounding. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1513–1522. IEEE, 2021.

Chaolei Tan, Jian-Fang Hu, and Wei-Shi Zheng. Matching and localizing: a simple yet effective framework for human-centric spatio-temporal video grounding. In *CAAI International Conference on Artificial Intelligence*, pp. 305–316. Springer, 2022.

Shihao Wang, Guo Chen, De-an Huang, Zhiqi Li, Minghan Li, Guilin Liu, Jan Kautz, Jose M Alvarez, Lei Zhang, and Zhiding Yu. Videoitg: Multimodal video understanding with instructed temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24640–24650, 2026a.

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, et al. Time-r1: Post-training large vision language model for temporal video grounding. *Advances in Neural Information Processing Systems*, 38:83330–83364, 2026b.

Jianlong Wu, Wei Liu, Ye Liu, Meng Liu, Liqiang Nie, Zhouchen Lin, and Chang Wen Chen. A survey on video temporal grounding with multimodal large language model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

LCT Xiaomi and Core Team. Mimo-vl technical report. *arXiv preprint arXiv:2506.03569*, 1(2):5, 2025.

Ziang Yan, Xinhao Li, Yinan He, Zhengrong Yue, Xiangyu Zeng, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. Videochat-r1. 5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. *arXiv preprint arXiv:2509.21100*, 2025.

Abhay Zala, Jaemin Cho, Satwik Kottur, Xilun Chen, Barlas Oguz, Yashar Mehdad, and Mohit Bansal. Hierarchical video-moment retrieval and step-captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23056–23065, 2023.

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.

APPENDIX

A RELATED WORK

A.1 MULTIMODAL LARGE LANGUAGE MODELS FOR TEMPORAL VIDEO GROUNDING

Temporal Video Grounding (TVG) requires localizing the video segment associated with a natural-language query (Lin et al., 2023; Wu et al., 2025; Zhang et al., 2026). Earlier methods (Liu et al., 2022; Su et al., 2021; Tan et al., 2022) typically treated temporal localization as a specialized discriminative task, whereas recent multimodal large language models formulate it as autoregressive timestamp generation. Existing MLLM-based approaches improve temporal grounding through architectural design, temporal representation learning, task-specific supervised fine-tuning, and post-training (Wang et al., 2026b; Chen et al., 2025; Li et al., 2025c; Yan et al., 2025). In particular, reinforcement-learning methods such as Time-R1 (Wang et al., 2026b) and TVG-R1 (Chen et al., 2025) optimize student-generated trajectories using verifiable temporal-overlap rewards, thereby reducing the mismatch between training and inference states. However, their sequence-level rewards provide limited guidance for individual timestamp actions and generally require multiple rollouts for stable optimization. Video-OPD++ is orthogonal to architectural and data-centric improvements: it keeps the model, training subset, and rollout budget fixed, and instead revisits the action-space supervision provided during post-training.

A.2 KNOWLEDGE DISTILLATION FOR TEMPORAL GROUNDING

Knowledge distillation (Bao et al., 2024; Liang et al., 2023; Ran et al., 2025) transfers the behavior of a stronger teacher to a smaller student through token- or distribution-level objectives. In off-policy distillation (Wang et al., 2026a), the teacher and student distributions are commonly evaluated on fixed demonstration prefixes. Such objectives provide explicit distributional targets, but the optimized states may differ from those encountered by the student during inference. The OP-FKD and OP-RKD variants studied in Video-OPD (Li et al., 2026) instantiate forward- and reverse-KL distillation on off-policy trajectories, respectively. In contrast, Video-OPD evaluates a fixed teacher on prefixes generated by the current student and therefore preserves the student-induced state distribution. Its dense sampled log-ratio feedback also reduces the need for grouped rollouts. Nevertheless, at each visited state, the objective depends only on the action realized by the student and does not construct a target distribution over plausible alternatives.

Video-OPD++ retains the on-policy states and single-rollout budget of Video-OPD, but changes the form of supervision at timestamp positions. Rather than directly imitating the teacher on a ground-truth trajectory, it extracts teacher-supported actions at the current student prefix and evaluates their local temporal utilities. The annotation is therefore used as an action-level utility signal rather than as a demonstration trajectory. This distinction allows the method to preserve training–inference state alignment while providing an explicit corrective target over alternative timestamp actions.

A.3 ON-POLICY DISTILLATION AND TASK-AWARE CORRECTIVE SUPERVISION

On-policy distillation (Agarwal et al., 2024) combines trajectories sampled from the current policy with dense feedback supplied by a fixed or slowly updated teacher. This formulation inherits the state-distribution alignment of on-policy reinforcement learning while avoiding purely sequence-level credit assignment. Video-OPD (Li et al., 2026) first applies this paradigm to TVG by evaluating each sampled student action with a teacher-derived log-ratio reward. Our analysis shows that this supervision is step-wise dense but action-wise pointwise: teacher information over non-sampled alternatives is absent from the optimization target.

Video-OPD++ addresses this limitation by constructing a geometry-calibrated corrective distribution within the teacher's local support. Its fixed-suffix intervention differs from conventional multi-rollout reinforcement learning (Shao et al., 2024): candidate actions are evaluated through deterministic timestamp parsing and exact tIoU rather than additional autoregressive trajectories or a learned reward model. The resulting target can be interpreted as teacher-regularized utility reweighting at student-induced states. It is also complementary to curriculum and data selection strategies such as

TVDF, because it modifies the supervision associated with each visited state rather than the set or ordering of training examples.

B Additional Diagnostic Analysis

This section provides the formal definitions and additional results underlying the empirical diagnosis in Section 2.3. Unless otherwise stated, all statistics are computed over the same 300 trajectories sampled using the training decoding configuration.

B.1 Concentration of the Sampled Log-Ratio Signal

Let

$$\mathcal{V}(\tau) = \{t \mid m_t = 1\} \tag{13}$$

denote the valid completion positions in trajectory τ. We quantify the effective number of supervised positions using the inverse participation ratio

$$N_{\text{eff}}(\tau) = \frac{\left(\sum_{t\in\mathcal{V}(\tau)} |r_t|\right)^2}{\sum_{t\in\mathcal{V}(\tau)} r_t^2}. \tag{14}$$

This quantity is maximized when the absolute supervisory mass is uniformly distributed and decreases when it is concentrated on a small number of positions.

We additionally define the top-three mass as

$$M_3(\tau) = \frac{\sum_{t\in\mathrm{Top3}\left(\{|r_j|\}_{j\in\mathcal{V}(\tau)}\right)} |r_t|}{\sum_{t\in\mathcal{V}(\tau)} |r_t|}. \tag{15}$$

Both N_{eff} and M_3 are first computed independently for each trajectory and then averaged over the diagnostic set.

Under the train-aligned stochastic rollout protocol, the mean effective token count is 2.82, while the three largest positions account for approximately 89% of the total absolute sampled log-ratio mass. These results show that the nominally step-wise supervision of Video-OPD is effectively concentrated on only a few temporal decisions.

B.2 Formal Definition of the Fixed-Suffix Oracle

Let $\mathcal{I}_{\text{ts}}(\tau)$ denote the timestamp positions of a student trajectory τ. At every $t \in \mathcal{I}_{\text{ts}}(\tau)$, the teacher-supported candidate set is

$$\mathcal{C}_t^{(K)} = \mathrm{TopK}_K\left(\pi_T(\cdot \mid s_t)\right). \tag{16}$$

For candidate $u \in \mathcal{C}_t^{(K)}$, we define the fixed-suffix intervention

$$\tau^{t\leftarrow u} = (a_{<t}, u, a_{>t}). \tag{17}$$

Candidates whose modified completions cannot be parsed into valid temporal intervals are excluded. Let $\mathcal{C}_{t,\text{val}}^{(K)}$ denote the remaining candidate set. The corresponding utility is

$$U_t(u) = \mathrm{tIoU}\left(\mathrm{Parse}\left(\tau^{t\leftarrow u}\right), b^\star\right). \tag{18}$$

The trajectory-level token-correction oracle is

$$U_{\text{oracle}}^{(K)}(\tau) = \max\left\{U(\tau), \max_{t\in\mathcal{I}_{\text{ts}}(\tau),\ u\in\mathcal{C}_{t,\text{val}}^{(K)}} U_t(u)\right\}, \tag{19}$$

where

$$U(\tau) = \mathrm{tIoU}\left(\mathrm{Parse}(\tau), b^\star\right) \tag{20}$$

is the utility of the original student output.

Correction coverage is defined as

$$\mathrm{Cov}^{(K)} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left[U^{(K)}_{\mathrm{oracle}}(\tau_i) > U(\tau_i)\right]. \tag{21}$$

The strict inequality excludes substitutions that merely tie with the original prediction.

Using $K = 32$, the sampled student trajectories achieve an average tIoU of 0.385. The oracle finds a strictly improving teacher-supported substitution for 76.3% of trajectories and raises average tIoU to 0.522, corresponding to an absolute gain of 0.136. Improving substitutions occur at start- and end-timestamp positions for 62.0% and 72.7% of trajectories, respectively.

The oracle statistics are identical for $K \in \{8, 16, 32\}$, indicating that useful local alternatives are already covered by the highest-probability portion of the teacher distribution.

B.3 Robustness to the Rollout Decoding Strategy

The main diagnosis uses stochastic trajectories sampled with the exact training decoding configuration. For completeness, we also repeat the analysis using greedy student trajectories. Table A1 shows that the quantitative values vary slightly, but all qualitative conclusions remain unchanged: the supervisory signal is concentrated, improving teacher-supported substitutions are common, and the oracle is already saturated at a small K.

Table A1: Robustness of the mechanism diagnosis to the student rollout decoding strategy. The sampled configuration matches the rollout distribution used during training.

Rollout	Base tIoU	N_{eff}	Top-3 Mass	Coverage	Oracle tIoU	ΔtIoU	Start Cov.	End Cov.
Greedy	0.404	2.52	91.0%	74.7%	0.545	0.141	55.0%	67.0%
Train-aligned sampling	0.385	2.82	89.0%	76.3%	0.522	0.136	62.0%	72.7%

C Additional Method Details and Derivations

C.1 Candidate-Relative Temporal Utility

For a candidate u at position t, let

$$U_t(u) = \mathrm{tIoU}\left(b_t(u), b^\star\right) \tag{22}$$

denote its fixed-suffix utility, and let

$$U_{\mathrm{cur}} = \mathrm{tIoU}\left(b(\tau), b^\star\right) \tag{23}$$

denote the utility of the original student completion. An equivalent implementation may first define

$$\Delta U_t(u) = U_t(u) - U_{\mathrm{cur}}. \tag{24}$$

Because U_{cur} is constant across candidates at the same position,

$$\Delta U_t(u) - \frac{1}{|\mathcal{C}_{t,\mathrm{val}}|}\sum_{w}\Delta U_t(w) = U_t(u) - \frac{1}{|\mathcal{C}_{t,\mathrm{val}}|}\sum_{w}U_t(w). \tag{25}$$

The candidate-wise standardization of ΔU_t is therefore identical to that of U_t:

$$\widehat{U}_t(u) = \frac{U_t(u) - \mu_t}{\sigma_t + \epsilon}. \tag{26}$$

Consequently, the geometry signal represents the relative ordering of candidates within the teacher support rather than an absolute improvement guarantee over the current action.

C.2 Full-Vocabulary Gradient of GCS

For a valid timestamp position, the per-position GCS loss is

$$\ell_{\mathrm{GCS},t} = -m_t \sum_{u\in\mathcal{C}_{t,\mathrm{val}}} q_t(u)\log\pi_\theta(u \mid s_t). \tag{27}$$

Extend q_t to the complete vocabulary $\mathcal{V}$ by

$$\widetilde{q}_t(v) = \begin{cases} q_t(v), & v \in \mathcal{C}_{t,\mathrm{val}}, \\ 0, & v \notin \mathcal{C}_{t,\mathrm{val}}. \end{cases} \tag{28}$$

Since $\pi_\theta(\cdot \mid s_t)$ is the full-vocabulary softmax, differentiating with respect to student logit $z_\theta(v \mid s_t)$ gives

$$\frac{\partial \ell_{\mathrm{GCS},t}}{\partial z_\theta(v \mid s_t)} = m_t \left[\pi_\theta(v \mid s_t) - \widetilde{q}_t(v)\right]. \tag{29}$$

After applying the reduction used by $\mathcal{L}_{\mathrm{GCS}}$, this expression is scaled by the corresponding timestamp normalization factor.

Equation 29 makes the distribution-matching update explicit. Supported actions with $q_t(v) > \pi_\theta(v \mid s_t)$ are promoted, supported actions with smaller target probability are suppressed, and all actions outside the support receive suppressive gradients through full-vocabulary normalization. The gradient therefore moves the student distribution toward the support-relative target defined jointly by the teacher prior and temporal utility.

C.3 Derivation of the KL-Regularized Target

At a timestamp state s_t, consider the optimization problem

$$q_t^\star = \arg\max_{q \in \Delta(\mathcal{C}_{t,\mathrm{val}})} \left\{ \gamma \sum_u q(u) \widehat{U}_t(u) - \sum_u q(u) \log \frac{q(u)}{\bar{\pi}_T(u \mid s_t)} \right\}. \tag{30}$$

The first term favors candidates with greater temporal utility, whereas the second regularizes the target toward the truncated teacher prior.

Introducing a Lagrange multiplier ξ for $\sum_u q(u) = 1$ gives

$$\mathcal{J}(q, \xi) = \gamma \sum_u q(u) \widehat{U}_t(u) - \sum_u q(u) \log \frac{q(u)}{\bar{\pi}_T(u \mid s_t)} + \xi \left(\sum_u q(u) - 1 \right). \tag{31}$$

The stationary condition is

$$\frac{\partial \mathcal{J}}{\partial q(u)} = \gamma \widehat{U}_t(u) - \log \frac{q(u)}{\bar{\pi}_T(u \mid s_t)} - 1 + \xi = 0. \tag{32}$$

Therefore,

$$q(u) = \bar{\pi}_T(u \mid s_t) \exp\left(\gamma \widehat{U}_t(u) + \xi - 1 \right). \tag{33}$$

Normalizing over the valid support yields

$$q_t^\star(u) = \frac{\bar{\pi}_T(u \mid s_t) \exp\left(\gamma \widehat{U}_t(u) \right)}{\sum_w \bar{\pi}_T(w \mid s_t) \exp\left(\gamma \widehat{U}_t(w) \right)}. \tag{34}$$

The objective in Equation 30 is strictly concave in q on the simplex, so Equation 34 is its unique optimizer.

This result establishes GCS as teacher-regularized utility reweighting within the local candidate support. It does not imply monotonic improvement of the complete autoregressive policy, because both the candidate domain and the utility are defined locally under a fixed student suffix.

C.4 Piecewise Form of the Token-Partitioned Objective

The compact objective in the main text can equivalently be written according to token type. Let

$$h_t = \mathbb{I}\left[t \in \mathcal{I}_{\mathrm{ts}}(\tau)\right]. \tag{35}$$

At non-timestamp positions, Video-OPD++ retains

$$\ell_t^{\text{non-ts}} = -m_t \rho_t(\theta)\, \mathrm{sg}(r_t). \tag{36}$$

At timestamp positions, the local sampled-action term is removed and replaced by

$$\ell_t^{\mathrm{ts}} = -\eta m_t \rho_t(\theta)\,\mathrm{sg}\left(F_t^{\lambda}\right) + \ell_{\mathrm{GCS},t}. \tag{37}$$

Thus,

$$\mathcal{L}_{\text{Video-OPD++}} = \mathbb{E}_{\tau}\left[\sum_{t:h_t=0} \ell_t^{\text{non-ts}} + \sum_{t:h_t=1} \ell_t^{\mathrm{ts}}\right]. \tag{38}$$

The two component ablations correspond to:

- **GCS w/o future:** set $\eta = 0$, leaving GCS as the only timestamp-position objective.
- **GCS + local sampled reward:** retain the local term $-m_t\rho_t(\theta)\,\mathrm{sg}(r_t)$ at timestamp positions in addition to GCS and the future residual.

D Detailed Experimental Setup

D.1 Training Details

Models and training data. We initialize the student from Qwen3-VL-8B-Instruct and use the GRPO-trained Qwen3-VL-32B model released with Video-OPD as the fixed teacher. The teacher remains frozen throughout post-training.

We use exactly the same 2,500 TVDF-selected examples as the official Video-OPD run. The original Video-OPD training pool combines temporal grounding data from HiREST (Zala et al., 2023), QuerYD (Oncescu et al., 2021), HowTo-Interlink7M (Jinpeng Wang et al., 2024), VTimeLLM (Huang et al., 2024), DiDeMo (Anne Hendricks et al., 2017), and TimeLens-100K (Zhang et al., 2026), resulting in 96,586 candidate training instances. TVDF is used to select the final 2,500-example subset. We introduce no additional training examples, data-selection rounds, or iterative post-training stages.

Visual and sequence configuration. Videos are sampled at 2 FPS. The number of sampled frames is capped at 768, and the per-video visual-token budget is set to 768. Each frame is encoded into at most 768 visual tokens, with a lower bound of 4 frames per video and 4 tokens per frame. The maximum generation length is $8{,}192$, and the maximum total sequence length, including the multimodal prompt and completion, is $16{,}000$. All methods use the same no-thinking prompt and temporal-answer format.

On-policy rollout generation. Each example produces one completion sampled from the current student policy. We use temperature 0.7, top-k 20, and top-p 0.8, matching the rollout configuration used during training. The sampled completion and its token IDs are shared by all supervision branches.

The teacher does not generate a separate trajectory. Instead, it performs teacher forcing over the exact student completion and evaluates every completion token under the corresponding student-generated prefix. Student and teacher log-probabilities are aligned over the same valid completion positions and use the same completion mask as the Video-OPD objective.

Optimization. We train for one epoch with a learning rate of 1×10^{-6} and a global batch size of 32. Training is conducted on eight H20 GPUs with a per-device batch size of 1 and four gradient-accumulation steps. Only the student language model is updated; the vision encoder and the modality aligner are kept frozen. Both the student and the teacher are executed with DeepSpeed ZeRO-3.

All controlled variants share the same student initialization, teacher checkpoint, 2,500-example training subset, data order, prompt, rollout configuration, and optimization schedule. Thus, the compared variants differ only in the objective component under study.

Teacher-supported candidate construction. For every detected timestamp position, we extract the teacher's top-K actions over the full vocabulary, with $K = 32$. The candidate set contains only teacher-proposed actions: it is not merged with the student top-K set, and the sampled action is not explicitly inserted.

For each candidate, we replace the token at the current timestamp position while retaining all remaining student token IDs. The resulting fixed-suffix completion is decoded and parsed into a temporal interval. Candidates that do not produce a valid interval are excluded. Candidate evaluation therefore requires no additional autoregressive rollout.

Geometry calibration. For each valid candidate u, we compute its temporal utility as the tIoU between the fixed-suffix candidate interval and the annotated interval. Utilities are standardized independently within each valid candidate support. We then construct

$$q_t(u) \propto \bar{\pi}_T(u \mid s_t) \exp\Big(\gamma \widehat{U}_t(u)\Big), \tag{39}$$

where $\bar{\pi}_T$ is the teacher distribution renormalized over the valid support.

We set $\gamma = 4.0$ and use no additional target-temperature scaling, corresponding to $\tau = 1$. Both the teacher probabilities and candidate utilities are detached. The GCS cross-entropy loss is assigned coefficient 1.0, and gradients propagate only through the student probabilities.

Sequential residual and token partitioning. We use $\lambda = 0.25$ to construct the discounted future residual and set its coefficient to $\eta = 0.25$. At timestamp positions, the local sampled-action reward is removed and replaced by the GCS distributional target. The policy-gradient branch at these positions contains only discounted feedback from subsequent token discrepancies. At non-timestamp positions, we retain the original local Video-OPD surrogate.

The complete objective is

$$\mathcal{L}_{\text{Video-OPD++}} = \mathcal{L}_{\text{GCS}} - \mathbb{E}_{\tau \sim \pi_{\theta_{\text{old}}}} \left[\sum_{t=1}^{T} m_t \rho_t(\theta) \, \text{sg}\big((1 - h_t) r_t + \eta h_t F_t^{\lambda}\big) \right], \tag{40}$$

where h_t indicates whether position t belongs to a timestamp. Ground-truth intervals are used only to compute candidate utilities and never serve as demonstration prefixes.

Inference cost. Video-OPD++ modifies only the post-training objective. Teacher-supported candidate construction, fixed-suffix interventions, and temporal-utility computation are not used at inference time. The final student therefore follows the same autoregressive inference procedure as Video-OPD and incurs no additional inference-time computation.

D.2 Evaluation Details

Benchmarks. We evaluate Temporal Video Grounding on three benchmarks: Charades-TimeLens, ActivityNet-TimeLens, and QVHighlights-TimeLens. These benchmarks use the corrected temporal annotations provided by TimeLens (Zhang et al., 2026), which revise annotation errors in the original Charades-STA (Gao et al., 2017), ActivityNet (Caba Heilbron et al., 2015), and QVHighlights (Lei et al., 2021) evaluation sets.

Metrics. Given a predicted interval $\hat{b}$ and a ground-truth interval $b^{\star}$, temporal IoU is defined as the duration of their intersection divided by the duration of their union. We report R@0.3, R@0.5, and R@0.7, corresponding to the proportion of examples whose tIoU is no smaller than 0.3, 0.5, and 0.7, respectively. We additionally report mIoU, the mean tIoU over the complete evaluation set.

Inference and parsing. All methods and ablation variants use the same evaluation prompt, video preprocessing, decoding configuration, timestamp-output format, interval parser, and invalid-output handling. The generated answer is parsed into a start and end timestamp before computing tIoU. These settings are held fixed across all controlled comparisons, ensuring that differences are attributable to post-training rather than inference-time changes.

D.3 Baseline Details

Proprietary MLLMs. We compare against GPT-4o (Hurst et al., 2024), GPT-5 (Singh et al., 2025), Gemini-2.0-Flash, Gemini-2.5-Flash, and Gemini-2.5-Pro (Comanici et al., 2025). These models provide a reference for the performance of contemporary proprietary multimodal systems on the TimeLens-corrected TVG benchmarks.

Open-source video MLLMs. The open-source comparison includes VideoChat-Flash-7B (Li et al., 2025b), Qwen2.5-VL-7B (Bai et al., 2025b), VideoChat-R1-7B (Li et al., 2025c), Time-R1-7B (Wang et al., 2026b), TVG-R1-7B (Chen et al., 2025), VideoChat-R1.5-7B (Yan et al., 2025), MiMo-VL-7B (Xiaomi & Team, 2025), and Qwen3-VL-8B-Instruct (Bai et al., 2025a). These models cover supervised, reinforcement-trained, and general-purpose video MLLMs with comparable parameter scales.

Post-training frameworks. The most direct comparisons are post-training methods based on Qwen3-VL-8B. OP-RKD performs off-policy reverse-KL distillation, while OP-FKD performs off-policy forward-KL distillation; both follow the implementations introduced in Video-OPD (Li et al., 2026). We also compare against GRPO (Shao et al., 2024) and the official Video-OPD model (Li et al., 2026).

E Complete Geometry-Strength Ablation

Figure 4 presents the relative trends as γ varies. Table A2 reports the complete numerical results. The setting $\gamma = 0$ reduces GCS to matching the truncated teacher support without temporal-utility calibration.

Table A2: Complete ablation on the geometry-calibration strength γ. $\gamma = 0$ denotes Support-only. **Bold** values indicate the best result in each column.

	Charades-TimeLens				ActivityNet-TimeLens				QVHighlights-TimeLens			
Setting	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU
Video-OPD	73.1	45.8	32.4	52.0	60.5	45.6	35.8	47.3	73.8	60.3	50.4	61.0
$\gamma = 0$ (Support-only)	72.0	44.8	31.8	51.1	61.8	46.5	36.3	48.7	73.5	60.4	50.8	61.2
$\gamma = 2$	73.2	46.6	**32.7**	52.2	63.7	48.6	38.0	50.4	75.9	62.8	52.6	63.0
$\gamma = 4$	**74.0**	**47.3**	32.2	**52.6**	**64.1**	**49.0**	**38.4**	**50.6**	**77.2**	**64.8**	**53.7**	**64.0**
$\gamma = 6$	72.3	45.9	32.5	51.8	63.6	47.9	37.1	49.8	75.7	62.8	53.1	63.0
$\gamma = 8$	71.5	44.8	31.1	50.9	63.7	48.0	37.3	49.8	76.3	63.2	53.0	63.5

Support-only achieves an average mIoU of 53.7, only 0.2 points above Video-OPD. Introducing geometry calibration with $\gamma = 4$ raises average mIoU to 55.7, an improvement of 2.1 points over Support-only. Larger values reduce performance, suggesting that excessive utility-based reweighting can overwhelm the teacher prior.

F Prompt and Timestamp Processing

F.1 Prompt Template

We use the same no-thinking prompt format for all training and evaluation variants.

System instruction:

```
You are a video analysis expert.
```

User instruction:

```
To accurately pinpoint the event ``{query}'' in the video,
determine the precise time period of the event.  Provide the
start and end times (in seconds, precise to two decimal
places) in the format ``start time to end time''.  For
example:  ``12.54 to 17.83''.
Video duration:  {duration} seconds.
```

No explicit chain-of-thought or auxiliary reasoning instruction is included.

F.2 Timestamp-Token Identification

For every sampled completion, timestamp positions are identified as follows:

1. We retain the exact completion token IDs generated by the student.
2. We incrementally decode the completion to associate every token with its character span.

3. We parse the first valid start–end timestamp pair using the same parser as evaluation.
4. Character spans belonging to the two timestamps are mapped back to completion-token positions, forming $\mathcal{I}_{\mathrm{ts}}(\tau)$.
5. For candidate evaluation, one timestamp token is replaced while all remaining completion token IDs are kept unchanged.
6. The modified sequence is decoded and parsed again. Unparsable, reversed, zero-length, or out-of-range intervals are treated as invalid candidates.

This procedure ensures that GCS acts only on tokens that determine the predicted temporal interval. Formatting and other non-timestamp tokens continue to use the vanilla Video-OPD objective.