GUIDE THE TOKENS, RANK THE MOMENTS: ADAPTIVE ON-POLICY DISTILLATION FOR TEMPORAL VIDEO GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Temporal video grounding requires identifying a queried event and recovering its temporal extent: an answer can contain the right action yet include irrelevant context or omit part of the event. On-policy distillation supplies dense teacher guidance during timestamp generation, but matching token distributions does not explicitly require the student's preferences over complete answers to reflect their localization quality. Our central idea is to use teacher proposals both as answers to imitate and as alternatives whose relative quality can be evaluated through temporal overlap. We propose Geo-OPD, a geometry-aware on-policy distillation framework that combines pointwise imitation of a teacher answer with listwise supervision, aligning the student's relative sequence scores with an annotation-IoU-derived preference distribution over teacher proposals. Adaptive token guidance complements these sequence objectives by balancing soft and sharp views of the same frozen teacher according to temporal discrepancy, while reliability filtering selects teacher supervision. Experiments on Charades-TimeLens, ActivityNet-TimeLens, and QVHighlights-TimeLens show improvements over the reported distillation baselines, with component ablations indicating that the additional benefit of listwise supervision varies across datasets. These results support combining guidance on how to generate temporal answers with explicit supervision on which candidate intervals better localize the queried event.

1 INTRODUCTION

Video understanding is a central problem in multimodal learning, with applications that require locating events within long, untrimmed videos (Ren et al., 2024). Temporal video grounding (TVG) addresses this need by identifying the start and end of an event described by a natural-language query (Gao et al., 2017; Hendricks et al., 2017; Lei et al., 2021). For example, given the query "a person pours water into a glass," the model must locate when the pouring occurs in the video. By connecting language descriptions to specific temporal segments, TVG supports query-based video retrieval and provides temporal context for interpreting individual events (Lei et al., 2021; Krishna et al., 2017). Accurate localization is thus essential for video language models.

Recent studies, including Time-R1 (Wang et al., 2025) and VideoChat-R1.5 (Yan et al., 2025), have advanced temporal grounding through reinforcement post-training and iterative perception, respectively. TimeLens (Zhang et al., 2026) further examines the roles of temporal representations, training strategies, and annotation quality. Alongside these developments, knowledge distillation transfers a teacher's predictive distribution to a student (Hinton et al., 2015). On-policy distillation provides this guidance on trajectories generated by the student. By supervising the prefixes the student actually encounters during generation, it addresses the mismatch between fixed training demonstrations and the student's inference states (Ross et al., 2011; Agarwal et al., 2024). Video-OPD (Li et al., 2026) brings this approach to TVG, using token-level teacher feedback to improve temporal answer generation. These advances make teacher-guided post-training promising for localization.

However, dense guidance during generation does not explicitly align the student's preferences over complete answers with their temporal localization quality. Token-level distillation matches teacher behavior at individual generation steps (Hinton et al., 2015; Agarwal et al., 2024), whereas TVG

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.

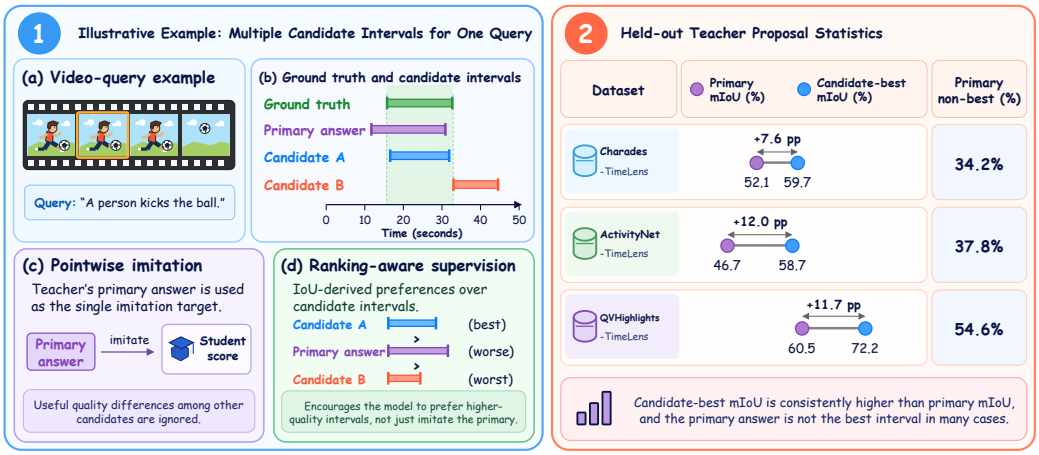


Figure 1: Illustration of quality variation among teacher proposals.

evaluates the interval jointly determined by the generated start and end times (Zhang et al., 2020a; Lei et al., 2021). Its objective does not directly require better-localized answers to receive higher sequence scores. Imitating a complete teacher answer, as in sequence-level distillation (Kim & Rush, 2016), supplies a generation target, but does not explicitly express the relative quality of alternative answers. This leaves useful supervision unexploited: temporal annotations can distinguish which teacher proposals cover the queried event more accurately. Figure 1 illustrates why imitating the primary answer alone does not capture these candidate quality differences. Connecting generation guidance to these quality differences requires supervision over complete candidate answers, with relative preferences represented across the candidate set (Cao et al., 2007).

Motivated by these challenges, we propose Geo-OPD, a geometry-aware on-policy distillation framework that combines learning to generate temporal answers with learning which answers better localize the queried event. Geo-OPD uses teacher proposals in two complementary ways: pointwise imitation reinforces the teacher's primary answer, while listwise supervision aligns the student's relative sequence scores with a quality distribution derived from temporal intersection-over-union (IoU). Together, these objectives connect answer generation to the geometry of complete intervals. They complement token-level guidance on student-generated prefixes, where an offline temporal discrepancy score balances soft and sharp views of the same frozen teacher. Reliability filtering and stratified selection determine which teacher supervision enters training. At inference, the student directly generates the predicted interval.

Extensive experiments on temporal grounding benchmarks demonstrate that Geo-OPD consistently outperforms the compared grounding and post-training methods across the reported evaluation thresholds, including those requiring tighter temporal overlap. Component analyses further examine how answer imitation, relative preference learning, and adaptive token guidance contribute to the overall gains. These results support integrating guidance on how to generate answers with explicit supervision on their localization quality as an effective approach to improving temporal video grounding. Our main contributions are:

- We formulate geometry-aware distillation for TVG using teacher proposals as generation targets and their relative localization quality as explicit supervision.
- We propose Geo-OPD, which integrates pointwise answer imitation and IoU-guided listwise supervision with adaptive token guidance, connecting complete-answer quality to learning on student-generated trajectories.
- Extensive experiments demonstrate consistent improvements over the compared methods across temporal grounding benchmarks, while component analyses characterize the contributions and interactions of the supervision objectives.

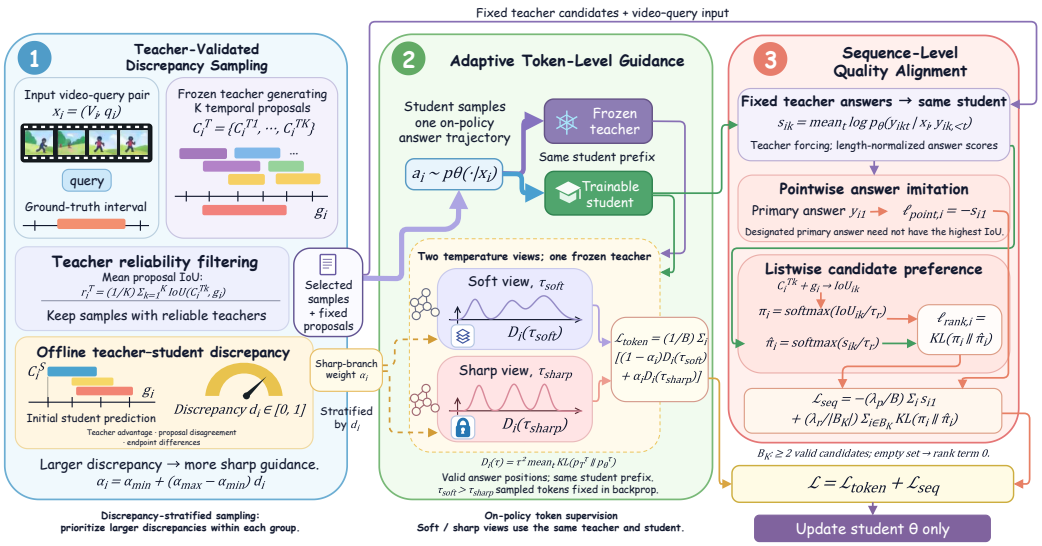


Figure 2: Overview of Geo-OPD.

2 Related Work

Temporal grounding and video language models. Language-based temporal localization maps a query to a video interval (Gao et al., 2017), complementing event-level video descriptions such as dense captioning (Krishna et al., 2017). Task-specific models encode intervals in a 2D start–end map in 2D-TAN (Zhang et al., 2020b) or predict temporal spans in VSLNet (Zhang et al., 2020a). QVHighlights combines query-conditioned moment retrieval and highlight detection, with Moment-DETR predicting temporal coordinates and saliency scores (Lei et al., 2021). Generative approaches express temporal answers through language models. VTimeLLM develops temporal-boundary awareness through staged training (Huang et al., 2024), while VTG-LLM incorporates timestamp knowledge and absolute-time tokens (Guo et al., 2025). Time-R1 studies reinforcement post-training for TVG (Wang et al., 2025), while VideoChat-R1.5 uses iterative perception to refine spatio-temporal reasoning (Yan et al., 2025). TimeLens examines temporal encoding, training recipes, and annotation quality, and provides refined evaluation sets (Zhang et al., 2026). Geo-OPD uses interval quality to guide teacher supervision in this generative interface.

Token-level and sequence-level distillation. Sequence-level knowledge distillation trains students on teacher-generated outputs (Kim & Rush, 2016); sequence distribution matching can also be formulated through generalized divergences (Wen et al., 2023). GKD separates the choice of training trajectories from the divergence used to match teacher and student distributions, including trajectories sampled from the student (Agarwal et al., 2024). MiniLLM minimizes reverse KL for language-model distillation (Gu et al., 2024), and DistiLLM combines skew KL objectives with adaptive off-policy reuse of student-generated outputs (Ko et al., 2024). Video-OPD uses teacher–student log-probability differences as reverse-KL-derived token rewards for policy gradient updates and introduces teacher-validated disagreement focusing (Li et al., 2026). Geo-OPD directly minimizes forward KL between temperature-scaled teacher and student distributions on student-generated prefixes. Its fixed teacher candidates provide an additional off-policy sequence objective.

Learning preferences over complete answers. Listwise learning-to-rank compares distributions over candidate lists, allowing supervision to express relative relevance across multiple items (Cao et al., 2007). Preference optimization for language models also learns from comparisons between complete responses; DPO uses paired preferences and policy likelihood ratios relative to a reference model (Rafailov et al., 2023). BRIO trains summarization models to assign candidate probabilities according to summary quality (Liu et al., 2022). SimPO uses average sequence log probability as a reference-free reward in a pairwise preference objective (Meng et al., 2024). These works link

sequence scores to answer quality across generation tasks. Geo-OPD matches a soft annotation-IoU target to a distribution over the student's mean answer-token log probabilities, combining this temporal-quality supervision with answer imitation and on-policy token guidance.

3 METHOD

In Geo-OPD, the teacher provides generation guidance and the annotations determine temporal quality. Training consists of offline preparation and two training branches. The token branch follows the student's current trajectory; the sequence branch scores a fixed set of teacher answers. Figure 2 summarizes the three components and their joint training objective.

3.1 PROBLEM SETUP

A training example consists of a video–query input $x_i = (V_i, q_i)$ and an annotated interval $g_i = [g_i^s, g_i^e]$. The student p_θ generates an answer $y = (y_1, \ldots, y_T)$ with $p_\theta(y \mid x_i) = \prod_{t=1}^{T} p_\theta(y_t \mid x_i, y_{<t})$. A timestamp parser maps the answer to an interval $c = [c^s, c^e]$. For two valid intervals c and g, temporal quality is

$$I(c, g) = [\min(c^e, g^e) - \max(c^s, g^s)]_+, \qquad u(c, g) = \frac{I(c, g)}{(c^e - c^s) + (g^e - g^s) - I(c, g)}, \tag{1}$$

where $[z]_+ = \max(z, 0)$. This quantity evaluates the entire interval, so moving either endpoint changes the quality of the complete answer.

Let p_T denote the frozen teacher. For each example, its valid proposals are $\{(y_{ik}, c_{ik})\}_{k=1}^{K_i}$, where y_{ik} is the serialized answer and c_{ik} its interval. The index $k = 1$ identifies the designated primary teacher answer, independently of annotation IoU. We write $u_{ik} = u(c_{ik}, g_i)$ for candidate quality. A separate prediction c_i^0 from the initial student supports offline sample selection.

3.2 TEACHER-VALIDATED DISCREPANCY SAMPLING

Reliable teacher proposals are essential because they serve as both imitation targets and candidates for comparison. We measure teacher reliability by average proposal quality,

$$r_i^T = \frac{1}{K_i} \sum_{k=1}^{K_i} u_{ik}, \qquad \mathcal{I}_{\text{rel}} = \{i : r_i^T \geq \rho\}, \tag{2}$$

where ρ is a reliability threshold. This criterion measures reliability over the full proposal set.

Within the reliable pool, an offline score $d_i \in [0, 1]$ summarizes three aspects of temporal discrepancy: the teacher's localization advantage over c_i^0, disagreement between c_i^0 and the teacher proposals, and their endpoint differences. We denote this configurable mapping by

$$d_i = \mathcal{G}_\phi\left(c_i^0, \{c_{ik}\}_{k=1}^{K_i}, g_i; L_i\right), \qquad 0 \leq d_i \leq 1, \tag{3}$$

where L_i is video duration and ϕ specifies the discrepancy configuration. The annotation determines localization advantage; direct teacher–student interval comparisons characterize disagreement. The score is computed offline and held fixed during training.

We divide the reliable pool into discrepancy strata and select a quota from each stratum, prioritizing larger discrepancies within each group. Stratification retains examples across different levels of agreement, whereas within-group prioritization emphasizes examples on which teacher and student predictions differ. Reliability and discrepancy serve different purposes: the first controls the quality of available supervision, and the second allocates a finite training budget within that pool.

3.3 ADAPTIVE TOKEN-LEVEL GUIDANCE

For each selected input, the current student samples one answer $a_i \sim p_\theta(\cdot \mid x_i)$. The teacher and student are evaluated on the same prefixes $a_{i,<t}$. Let $z_{T,it}(v)$ and $z_{\theta,it}(v)$ be their logits for vocabulary item v at such a prefix. With a shared output vocabulary and tokenization, the temperature view

for $m \in \{T, \theta\}$ is

$$p_{m,it}^{(\tau)}(v) = \frac{\exp(z_{m,it}(v)/\tau)}{\sum_{v'} \exp(z_{m,it}(v')/\tau)}. \tag{4}$$

Following temperature-based distillation (Hinton et al., 2015), the two temperatures control the concentration of the same teacher signal. The soft view retains mass over alternative continuations, while the sharp view emphasizes preferred continuations. Their relative weights depend on offline temporal discrepancy:

$$\alpha_i = \alpha_{\min} + (\alpha_{\max} - \alpha_{\min}) d_i, \qquad 0 \le \alpha_{\min} \le \alpha_{\max} \le 1. \tag{5}$$

Larger discrepancies therefore place greater weight on sharp guidance. Both views use the same frozen teacher.

Let the nonempty mask $\mathcal{A}_i$ contain the valid answer positions in the rollout. The forward-KL loss for a temperature view is

$$\mathcal{D}_i(\tau) = \frac{\tau^2}{|\mathcal{A}_i|} \sum_{t \in \mathcal{A}_i} \sum_{v} p_{T,it}^{(\tau)}(v) \log \frac{p_{T,it}^{(\tau)}(v)}{p_{\theta,it}^{(\tau)}(v)}. \tag{6}$$

For a minibatch of B examples, the token objective is

$$\mathcal{L}_{\text{token}} = \frac{1}{B} \sum_{i=1}^{B} \left[(1 - \alpha_i) \mathcal{D}_i(\tau_{\text{soft}}) + \alpha_i \mathcal{D}_i(\tau_{\text{sharp}}) \right], \tag{7}$$

with $\tau_{\text{soft}} > \tau_{\text{sharp}} > 0$. Both branches use the same logits and sampled prefixes. The sampled tokens are fixed inputs during backpropagation, as in on-policy distribution matching (Agarwal et al., 2024); gradients pass through student probabilities.

3.4 Sequence-Level Quality Alignment

The sequence branch trains the same student on fixed teacher proposals. It constrains both the absolute score of the imitation target and the relative scores of alternative answers.

Pointwise answer imitation. For candidate y_{ik}, let $\mathcal{M}_{ik}$ be its nonempty set of scored answer positions. Teacher forcing gives the length-normalized score

$$s_{ik} = \frac{1}{|\mathcal{M}_{ik}|} \sum_{t \in \mathcal{M}_{ik}} \log p_\theta(y_{ik,t} \mid x_i, y_{ik,<t}). \tag{8}$$

The pointwise term minimizes $-s_{i1}$, encouraging generation of the designated teacher answer. Averaging over answer positions controls for differences in token count, as also used in sequence preference scoring (Meng et al., 2024). The score is the mean answer-token log probability.

Listwise candidate preference. To learn quality-based preferences across the candidate set, we turn candidate IoUs into a distribution and normalize student scores over the same set:

$$\pi_{ik} = \frac{\exp(u_{ik}/\tau_r)}{\sum_{j=1}^{K_i} \exp(u_{ij}/\tau_r)}, \qquad \hat{\pi}_{ik} = \frac{\exp(s_{ik}/\tau_r)}{\sum_{j=1}^{K_i} \exp(s_{ij}/\tau_r)}, \tag{9}$$

where $\tau_r > 0$ is the shared ranking temperature. Higher IoU receives greater target mass; equal IoUs receive equal mass. This target preserves quality differences across all proposals. It follows the distributional view of listwise ranking (Cao et al., 2007), with temporal overlap supplying relevance.

The combined sequence objective is

$$\mathcal{L}_{\text{seq}} = -\frac{\lambda_p}{B} \sum_{i=1}^{B} s_{i1} + \frac{\lambda_r}{|\mathcal{B}_K|} \sum_{i \in \mathcal{B}_K} D_{\text{KL}}(\boldsymbol{\pi}_i \,\|\, \hat{\boldsymbol{\pi}}_i), \tag{10}$$

where $\lambda_p, \lambda_r \ge 0$ and $\mathcal{B}_K = \{i : K_i \ge 2\}$ within the minibatch. When this set is empty, the listwise contribution is defined as zero. A single valid candidate still supports pointwise imitation. Candidate intervals, IoUs, and target preferences are fixed with respect to student parameters.

Table 1: Temporal grounding results on three TimeLens-corrected temporal video grounding benchmarks. Gains indicate improvements over Qwen3-VL-8B-Instruct.

Method	Charades			ActivityNet			QVHighlights		
	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
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
Time-R1-7B	57.9	32.0	16.9	44.8	31.0	19.0	65.8	51.5	36.1
MiMo-VL-7B	57.9	42.6	20.5	49.3	38.7	22.4	57.1	42.6	28.4
VideoChat-Flash-7B	60.2	37.9	17.8	35.5	21.8	10.5	45.2	30.6	16.7
Qwen3-VL-8B-Instruct	65.7	43.5	23.8	55.8	43.2	28.4	65.8	55.4	40.3
+GT-SFT	66.5	47.2	26.8	57.2	48.4	29.7	67.6	55.9	41.2
+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	69.8	50.9	28.3	56.4	43.2	28.4	67.0	56.5	41.1
+Geo-OPD (Ours)	**73.0** (↑7.3)	**59.3** (↑15.8)	**33.3** (↑9.5)	**61.5** (↑5.7)	**51.1** (↑7.9)	**34.9** (↑6.5)	**76.3** (↑10.5)	**65.9** (↑10.5)	**50.8** (↑10.5)

The listwise distribution is invariant to a common shift of all candidate scores. It therefore constrains relative preferences among the candidates. Pointwise imitation raises s_{i1} to anchor the primary answer's absolute score. The two objectives interact when the primary candidate is not the best-localized proposal: imitation favors its generation, while listwise supervision reallocates relative scores according to IoU. The weights govern this tradeoff.

3.5 Training and Inference

The final objective is

$$\mathcal{L} = \mathcal{L}_{\text{token}} + \mathcal{L}_{\text{seq}}. \tag{11}$$

At each step, we sample a minibatch from the selected pool and update θ using $\nabla_\theta \mathcal{L}$, with fresh student rollouts and teacher-forced candidate scores. Teacher proposals and offline discrepancy remain fixed. Training annotations determine sample selection and candidate-quality preferences. At inference, p_θ generates an answer, which the timestamp parser converts to an interval.

4 Experiments

4.1 Evaluation Setup

We evaluate on Charades-TimeLens, ActivityNet-TimeLens, and QVHighlights-TimeLens, using the corrected annotations provided by TimeLens (Zhang et al., 2026). Their source benchmarks cover language-based activity localization, event descriptions in untrimmed videos, and query-conditioned moment retrieval, respectively (Gao et al., 2017; Krishna et al., 2017; Lei et al., 2021). All evaluations use the corrected annotations.

We report R@0.3, R@0.5, and R@0.7, the percentage of queries whose first predicted interval reaches the corresponding temporal IoU threshold. Mean IoU (mIoU) averages temporal overlap over evaluation queries, also expressed as a percentage. R@0.7 emphasizes tighter boundary agreement, while mIoU averages overlap across queries. The candidate-ranking analysis measures score ordering over fixed teacher proposals.

Table 1 groups grounding models and post-training approaches around a Qwen3-VL-8B-Instruct student (Bai et al., 2025). The latter include direct ground-truth supervised fine-tuning (GT-SFT), off-policy reverse- and forward-KL distillation (OP-RKD and OP-FKD), GRPO (Shao et al., 2024), and Video-OPD. The off-policy naming follows Video-OPD (Li et al., 2026). GT-SFT supervises the serialized annotation, whereas Geo-OPD uses annotations to evaluate teacher proposals. The

Table 2: Geo-OPD supervision ablation. Adaptive: adaptive token guidance; Pointwise: teacher-answer imitation; Listwise: IoU-guided candidate preference learning.

Supervision Components			Charades				ActivityNet				QVHighlights			
Adapt.	Point.	List.	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
✗	✗	✗	69.8	50.9	28.3	48.3	56.4	43.2	28.4	41.8	67.0	56.5	41.1	52.1
✓	✗	✗	72.4	58.7	32.8	52.0	60.6	49.9	34.0	45.9	73.7	64.0	49.7	58.5
✗	✓	✗	72.8	59.1	33.2	52.3	61.3	50.7	34.4	46.5	73.4	63.9	49.5	58.6
✗	✗	✓	72.4	58.6	31.4	51.9	60.7	49.8	33.7	45.8	73.5	64.0	49.5	58.4
✓	✓	✗	72.6	59.3	33.2	52.3	61.2	50.6	34.8	46.5	73.5	63.8	49.1	58.4
✓	✗	✓	72.2	58.6	32.6	51.8	60.6	49.9	33.9	45.9	74.0	64.2	49.6	58.6
✗	✓	✓	73.0	59.3	33.0	52.4	61.3	51.0	34.0	46.5	73.8	63.7	49.5	58.7
✓	✓	✓	**73.0**	**59.3**	**33.3**	**52.4**	**61.5**	**51.1**	**34.9**	**46.8**	**76.3**	**65.9**	**50.8**	**60.4**

model comparison includes TVG-R1 (Chen et al., 2025), VideoChat-R1.5 (Yan et al., 2025), Time-R1 (Wang et al., 2025), MiMo-VL (Core Team et al., 2025), and VideoChat-Flash (Li et al., 2025).

Implementation and comparison protocol. We initialize the student from Qwen3-VL-8B-Instruct and keep the teacher frozen during distillation. We reproduce Video-OPD (Li et al., 2026) with this student backbone; the Video-OPD results in Table 1 are obtained from our reproduction rather than taken from the original paper. For Geo-OPD, teacher proposals and discrepancy scores are computed offline and held fixed throughout training, while student rollouts are sampled afresh at each update. Training annotations support sample selection and candidate-quality supervision; inference requires only the student and the timestamp parser.

4.2 Overall Grounding Performance

Geo-OPD reaches R@0.5 values of 59.3, 51.1, and 65.9 on Charades-TimeLens, ActivityNet-TimeLens, and QVHighlights-TimeLens, respectively. Relative to the base student in Table 1, these correspond to gains of 15.8, 7.9, and 10.5 percentage points. The improvement extends to R@0.7, where Geo-OPD obtains 33.3, 34.9, and 50.8.

Geo-OPD exceeds the reported GT-SFT results at all three thresholds on each benchmark, with R@0.5 differences of 12.1, 2.7, and 10.0 percentage points. It also exceeds both off-policy distillation variants; its R@0.5 differences from OP-FKD are 12.0, 6.9, and 7.0 points. Relative to the listed Video-OPD row, the corresponding differences are 8.4, 7.9, and 9.4 points. These comparisons establish the ordering of the reported scores; the component comparisons below examine how performance changes across supervision configurations.

4.3 Supervision Components

Table 2 reports all combinations of the three components. Adaptive denotes the discrepancy-conditioned soft/sharp token objective, Pointwise denotes teacher-answer imitation, and Listwise denotes the IoU-derived preference objective.

The Pointwise-only configuration reaches R@0.5 values of 59.1, 50.7, and 63.9, compared with 50.9, 43.2, and 56.5 in the reference row. The Adaptive-only configuration obtains 58.7, 49.9, and 64.0, while the Listwise-only configuration obtains 58.6, 49.8, and 64.0. Each of these configurations exceeds the reference scores. Pointwise-only results are already close to the complete model on Charades-TimeLens and ActivityNet-TimeLens.

Comparing Adaptive + Pointwise with the complete model keeps both components enabled and changes the Listwise flag. R@0.5 changes from 59.3 to 59.3, from 50.6 to 51.1, and from 63.8 to 65.9 across the three datasets; the corresponding mIoU differences are 0.1, 0.3, and 2.0 percentage points. The largest observed additional grounding gain is on QVHighlights-TimeLens, whereas Charades-TimeLens changes little.

4.4 Data Selection and Candidate Diagnostics

Separating reliability from discrepancy. Table 3 compares random sampling, reliability filtering, discrepancy prioritization, and stratification. Random sampling draws from the original eligible pool; reliability-only sampling draws uniformly from the filtered pool. Global prioritization favors the largest discrepancies, while stratified uniform sampling draws within each stratum.

Reliability-only sampling exceeds random sampling on both metrics in all three datasets, but the full selector is not uniformly better than simpler alternatives. On ActivityNet-TimeLens, reliability-only sampling (52.7 R@0.5, 47.1 mIoU) and global prioritization (53.5, 47.2) exceed the full selector (51.1, 46.8). On QVHighlights-TimeLens, it has higher R@0.5 (65.9 vs. 63.4) but lower mIoU (60.4 vs. 63.3) than global prioritization. Stratified uniform sampling leads in Charades and ActivityNet mIoU (53.1 and 48.3).

Table 3: Data-selection ablation on three TimeLens benchmarks. Row 1: random sampling; Rel.: reliability filtering; Strat.: stratification; d_i: discrepancy prioritization. Higher is better.

Data Selection			Charades		ActivityNet		QVHighlights	
Rel.	Strat.	d_i	R@0.5	mIoU	R@0.5	mIoU	R@0.5	mIoU
✗	✗	✗	54.9	50.4	52.3	46.5	65.4	56.3
✓	✗	✗	**59.7**	50.6	52.7	47.1	65.8	58.8
✓	✗	✓	57.1	50.8	**53.5**	47.2	63.4	**63.3**
✓	✓	✗	58.6	**53.1**	49.9	**48.3**	65.8	61.2
✓	✓	✓	59.3	52.4	51.1	46.8	**65.9**	60.4

Candidate score–quality alignment. Table 4 compares the complete model with Adaptive + Pointwise on fixed held-out teacher proposals. Each query has one annotated interval. Both models score identical serialized answers using Equation 8, with the same inputs, answer-token masks, and EOS handling. Valid intervals have finite endpoints satisfying $0 \le c^s < c^e$; queries with invalid timestamps, missing scores, or empty candidate sets are excluded. Duplicate intervals are removed, retaining the primary answer when present and otherwise the first proposal. PairAcc evaluates unordered pairs with IoU differences greater than $\epsilon = 10^{-9}$: score ties within ϵ receive 0.5, and otherwise agreement with IoU ordering receives 1 and reversal receives 0. Scores are averaged within each query and then across queries with at least one eligible pair, and multiplied by 100. The paired difference Δ is full minus without listwise. Its 95% interval uses the 2.5th and 97.5th percentiles of 2,000 paired query-level bootstrap resamples (seed 42), holding the checkpoints fixed.

Table 4: Candidate score–quality alignment on fixed held-out teacher proposals. W/o listwise: Adaptive + Pointwise; full: Geo-OPD. PairAcc is a query-macro percentage; Δ and its confidence interval are in percentage points. N_{rank} counts queries with unequal-IoU candidate pairs.

Dataset	N_{rank}	PairAcc ↑		Δ ↑	95% CI of Δ
		W/o listwise	Full		
Charades	767	61.6	61.3	−0.3	$[-2.0, 1.4]$
ActivityNet	765	56.8	61.2	4.4	$[2.6, 6.2]$
QVHighlights	664	64.1	68.1	4.0	$[2.3, 5.7]$

The reported PairAcc differences are −0.3, 4.4, and 4.0 percentage points on Charades-TimeLens, ActivityNet-TimeLens, and QVHighlights-TimeLens, respectively. The Charades interval $[-2.0, 1.4]$ includes zero, while the ActivityNet and QVHighlights intervals, $[2.6, 6.2]$ and $[2.3, 5.7]$, lie above zero. These intervals support improved ordering on the latter two datasets, with no clear improvement on Charades. This pattern is compatible with Table 2, but PairAcc improves most on ActivityNet and grounding on QVHighlights.

Table 5: Quality differences within held-out teacher proposal sets. N_{valid} includes single-candidate queries and may exceed N_{rank} in Table 4. IoUs are scaled by 100; mean gap is in percentage points; primary non-best is a percentage. Candidate-best: per-set annotation-IoU oracle.

Dataset	N_{valid}	Primary mIoU	Candidate-best mIoU	Mean gap	Primary non-best (%)
Charades	939	52.1	59.7	7.6	34.2
ActivityNet	966	46.7	58.7	12.0	37.8
QVHighlights	914	60.5	72.2	11.7	54.6

Quality variation within proposal sets. Table 5 summarizes proposal quality over all N_{valid} valid queries, including single-candidate queries. Primary and candidate-best mIoU average the IoU of the designated primary answer and the highest-IoU candidate, respectively, multiplied by 100. Mean gap averages their per-query differences in percentage points; primary non-best is the percentage of queries whose best-candidate IoU exceeds primary IoU by more than ϵ.

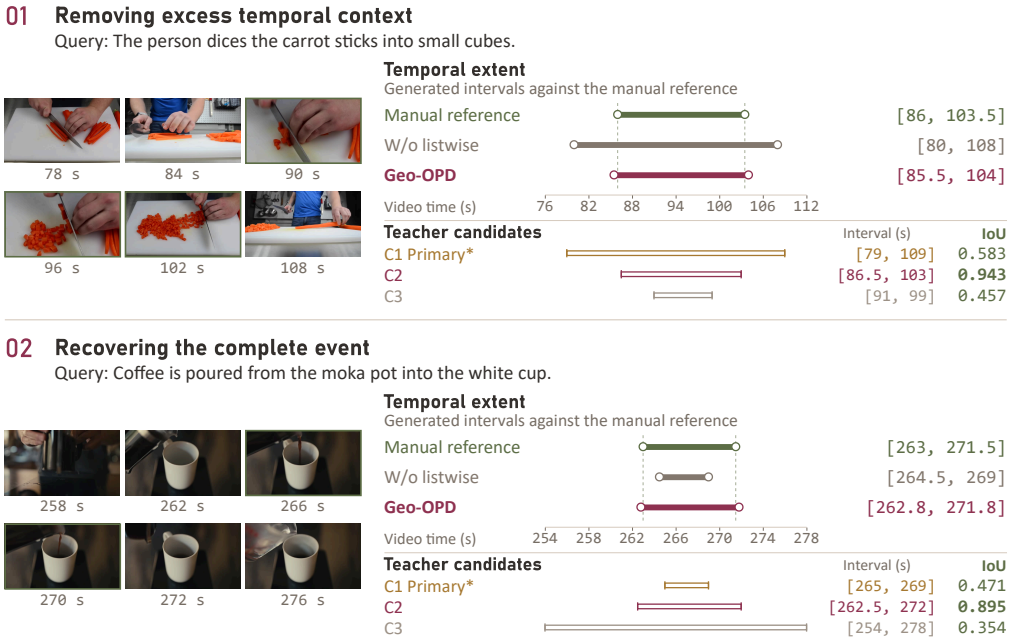


Figure 3: Qualitative comparisons with and without listwise supervision.

Candidate-best mIoU exceeds primary-answer mIoU by 7.6, 12.0, and 11.7 percentage points across the three datasets. The primary answer is non-best for 34.2%, 37.8%, and 54.6% of valid queries. These gaps reveal proposal quality information beyond the primary imitation target.

The proposal statistics and ranking diagnostics address two distinct questions: whether better alternatives exist and whether the student assigns them higher scores. QVHighlights has the highest primary non-best rate (54.6%) and the largest mIoU gain from adding listwise supervision to Adaptive + Pointwise (2.0 points), consistent with the value of comparing alternatives beyond a single imitation target. However, ranking and grounding gains need not be proportional: PairAcc evaluates fixed teacher proposal sets, whereas grounding evaluates student-generated intervals. Geo-OPD learns from these preferences during training but does not rerank teacher proposals at inference.

4.5 Qualitative Analysis

Figure 3 illustrates two temporal grounding cases with and without listwise supervision. In the first case, Geo-OPD reduces excess temporal context; in the second, it recovers more of the event's duration. Both predictions more closely match the manual reference, while the teacher candidate sets contain alternatives with higher IoU than the primary answer.

5 Conclusion

We propose Geo-OPD, a geometry-aware on-policy distillation framework for temporal video grounding. Geo-OPD provides adaptive token-level guidance on student-generated trajectories, using temporal discrepancy to balance soft and sharp distributions from the same frozen teacher. Building on this guidance, we introduce sequence-level quality alignment that combines pointwise imitation of a teacher answer with listwise preference learning over teacher proposals. The listwise objective aligns the student's relative sequence scores with a preference distribution derived from annotation IoU, explicitly favoring better-localized intervals. Together, these mechanisms adapt generation guidance to temporal discrepancy and connect complete-answer preferences to localization quality, addressing the gap between token-level imitation and interval-level evaluation. Experiments on three TimeLens-corrected benchmarks show that Geo-OPD consistently outperforms the compared grounding and post-training methods across the reported temporal IoU thresholds. Geo-OPD offers a promising basis for advancing quality-aware distillation in temporal video grounding.

AI USE STATEMENT

Generative AI assistance was used for manuscript organization, prose revision, mathematical exposition, literature lookup, and LaTeX formatting and verification. It also assisted in drafting experimental protocols and analysis documentation.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5be69a584901a26c521c2b51e40a4c20-Abstract-Conference.html.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL Technical Report, 2025. URL https://arxiv.org/abs/2511.21631. arXiv preprint.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, ICML '07 & ILP '07, pp. 129–136. ACM, June 2007. doi: 10.1145/1273496.1273513. URL http://dx.doi.org/10.1145/1273496.1273513.

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/.

Core Team, Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, Kainan Bao, Hao Tian, Hailin Zhang, Gang Wang, Dawei Zhu, Cici, Chenhong He, Bowen Ye, Bowen Shen, Zihan Zhang, Zihan Jiang, Zhixian Zheng, Zhichao Song, Zhenbo Luo, Yue Yu, Yudong Wang, Yuanyuan Tian, Yu Tu, Yihan Yan, Yi Huang, Xu Wang, Xinzhe Xu, Xingchen Song, Xing Zhang, Xing Yong, Xin Zhang, Xiangwei Deng, Wenyu Yang, Wenhan Ma, Weiwei Lv, Weiji Zhuang, Wei Liu, Sirui Deng, Shuo Liu, Shimao Chen, Shihua Yu, Shaohui Liu, Shande Wang, Rui Ma, Qiantong Wang, Peng Wang, Nuo Chen, Menghang Zhu, Kangyang Zhou, Kang Zhou, Kai Fang, Jun Shi, Jinhao Dong, Jiebao Xiao, Jiaming Xu, Huaqiu Liu, Hongshen Xu, Heng Qu, Haochen Zhao, Hanglong Lv, Guoan Wang, Duo Zhang, Dong Zhang, Di Zhang, Chong Ma, Chang Liu, Can Cai, and Bingquan Xia. MiMo-VL Technical Report, 2025. URL https://arxiv.org/abs/2506.03569. arXiv preprint.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: Temporal Activity Localization via Language Query, 2017. URL https://arxiv.org/abs/1705.02101.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge Distillation of Large Language Models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/8ac015d409635f196f9e3e9dcfb9a94e-Abstract-Conference.html.

Yongxin Guo, Jingyu Liu, Mingda Li, Dingxin Cheng, Xiaoying Tang, Dianbo Sui, Qingbin Liu, Xi Chen, and Kevin Zhao. VTG-LLM: Integrating Timestamp Knowledge into Video LLMs for Enhanced Video Temporal Grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 3302–3310, 2025. doi: 10.1609/aaai.v39i3.32341. URL https://ojs.aaai.org/index.php/AAAI/article/view/32341.

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. URL https://openaccess.thecvf.com/content_iccv_2017/html/Hendricks_Localizing_Moments_in_ICCV_2017_paper.html.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network, 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, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_VTimeLLM_Empower_LLM_to_Grasp_Video_Moments_CVPR_2024_paper.html.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1139. URL https://aclanthology.org/D16-1139/.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards Streamlined Distillation for Large Language Models. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 24872–24895, 2024. URL https://proceedings.mlr.press/v235/ko24c.html.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-Captioning Events in Videos, 2017. URL https://arxiv.org/abs/1705.00754.

Jie Lei, Tamara L. Berg, and Mohit Bansal. QVHighlights: Detecting Moments and Highlights in Videos via Natural Language Queries, 2021. URL https://arxiv.org/abs/2107.09609.

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. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.02994v3.

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, 2025. URL https://arxiv.org/abs/2501.00574v4. arXiv preprint.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. BRIO: Bringing Order to Abstractive Summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2890–2903, 2022. doi: 10.18653/v1/2022.acl-long.207. URL https://aclanthology.org/2022.acl-long.207/.

Yu Meng, Mengzhou Xia, and Danqi Chen. SimPO: Simple Preference Optimization with a Reference-Free Reward. In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e099c1c9699814af0be873a175361713-Abstract-Conference.html.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Advances in Neural Information Processing Systems*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

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. URL https://openaccess.thecvf.com/content/CVPR2024/html/Ren_TimeChat_A_Time-sensitive_Multimodal_Large_Language_Model_for_Long_Video_CVPR_2024_paper.html.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635. PMLR, 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

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, 2024. URL https://arxiv.org/abs/2402.03300. arXiv preprint.

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, 2025. URL https://arxiv.org/abs/2503.13377.

Yuqiao Wen, Zichao Li, Wenyu Du, and Lili Mou. f-divergence minimization for sequence-level knowledge distillation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10817–10834, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.605. URL https://aclanthology.org/2023.acl-long.605/.

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, 2025. URL https://arxiv.org/abs/2509.21100.

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, 2020a. 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 (CVPR)*, pp. 10419–10429, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhang_TimeLens_Rethinking_Video_Temporal_Grounding_with_Multimodal_LLMs_CVPR_2026_paper.html.

Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2D Temporal Adjacent Networks for Moment Localization with Natural Language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 12870–12877, 2020b. doi: 10.1609/aaai.v34i07.6984. URL https://ojs.aaai.org/index.php/AAAI/article/view/6984.