GLoR: Global-Local Ranking for Video Temporal Grounding

Anonymous authors
Paper under double-blind review

Abstract

Video temporal grounding (VTG) requires a model to return the temporal segment in an untrimmed video that corresponds to a natural-language query. In VTG, especially in long-video scenarios, candidate segments with lower relevance may receive higher confidence scores than target segments. Consequently, the model may fail not because the correct moment is absent, but because an inferior candidate is promoted to Top-1, turning an otherwise successful localization into an incorrect prediction. To address this issue, we propose a method named the Global-Local Ranking Grounder (GLoR), which decouples global candidate ranking from local boundary localization. GLoR adopts a global branch for candidate scoring and ranking and a local branch for boundary prediction. The global branch uses relevance evidence to determine the degree of match between a candidate segment and the query, while the local branch preserves high-frequency temporal variations and obtains more precise temporal boundaries by predicting probability distributions over start and end boundaries. In addition, IoU-based ranking and candidate-quality calibration further improve the accuracy of candidate scores. We evaluate GLoR on the TaCoS, Ego4D-NLQ, and Soccer-GMR datasets. Experimental results show that GLoR achieves state-of-the-art (SOTA) performance on TaCoS, Ego4D-NLQ, and Soccer-GMR, fully demonstrating its outstanding capability in candidate ranking and precise temporal localization.

1 Introduction

Video temporal grounding (VTG) requires a model to retrieve the temporal segment in an untrimmed video that matches a natural-language query (Hendricks et al., 2017; Gao et al., 2017; Chen et al., 2018). Beyond serving as a standalone moment-retrieval task, accurate temporal grounding provides fine-grained temporal evidence for long-form video understanding and multimodal large language model reasoning. UniTime employs temporal grounding as a preliminary retriever for long-video question answering (Li et al., 2025), while VTimeCoT uses relevant-moment retrieval to support multimodal video reasoning (Zhang et al., 2025). A reliable VTG model must determine which event satisfies the query and when that event starts and ends. However, many VTG methods first decode multiple candidate intervals and then rank them according to their confidence scores to produce the final prediction. SOONet filters long-video candidates through pre-ranking and re-ranking (Pan et al., 2023), BAM-DETR uses localization quality to adjust candidate priority (Lee & Byun, 2024), and Sim-DETR reveals competition among queries associated with related target moments (Tang et al., 2025). These studies indicate that even when a suitable target segment has already been generated, a less relevant candidate may receive a higher confidence score and be ranked above the target, resulting in localization failure. This problem becomes more pronounced in long videos, where the target event often occupies only a small fraction of the video and extended duration together with temporal redundancy introduces more competing candidates. Soccer videos provide a representative example: many intervals contain players, the ball, and recurring camera views, whereas only a small subset truly satisfies the query.

Existing VTG methods have substantially improved candidate generation and cross-modal interaction, yet their confidence supervision does not fully resolve candidate misranking. For example, 2D-TAN learns scores over structured temporal proposals (Zhang et al., 2020b), Moment-DETR combines set matching with saliency supervision (Lei et al., 2021), and FlashVTG refines dense point confidence using neighboring and cross-scale context (Cao et al., 2025). These designs im-

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.

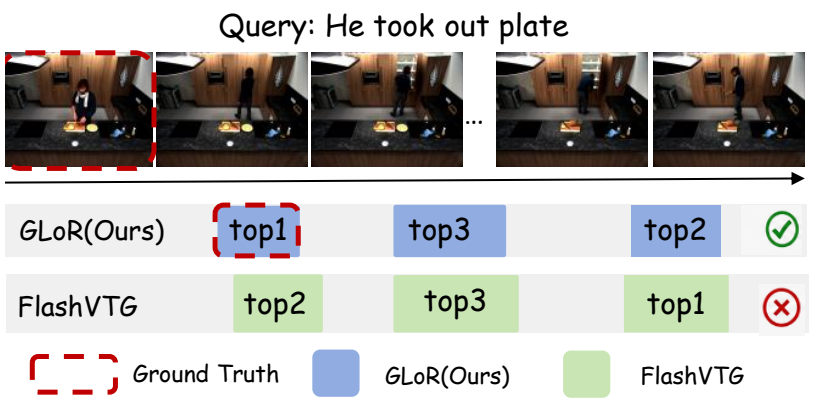

Figure 1: Candidate misranking on TaCoS. The dense baseline ranks a plausible but incorrect interval first, whereas GLoR promotes the ground-truth interval.

prove temporal coverage and relevance estimation, but do not directly require a decoded interval with higher localization quality to receive a higher score. Moreover, scoring and boundary regression commonly operate on shared temporal representations, although ranking benefits from query-conditioned global context while boundary estimation depends on local transitions. This mismatch leaves candidate ordering as an important source of localization error.

To address this problem, we propose the *Global-Local Ranking Grounder* (GLoR), a ranking-oriented VTG model that decouples global candidate ranking from local boundary estimation over a shared cross-modal temporal pyramid. The global branch aggregates long-range and cross-scale context to assess whether a candidate is supported by the queried event, whereas the local branch preserves temporal variations and predicts probability distributions over start and end boundaries. IoU-aware ranking and candidate-quality calibration further improve confidence reliability.

GLoR introduces *Global Relevance Fields* (GRF) to estimate whether each global position lies inside and near the center of the queried event, providing structured support for candidate ranking (Tian et al., 2019). In parallel, *Local Boundary Distributions* (LBD) represent uncertain start and end locations as categorical distributions and preserve boundary-sensitive temporal variations without top-down global aggregation (Li et al., 2020; 2021). This separation allows the global branch to compare candidates using long-range context while the local branch retains precise boundary cues. IoU-aware ranking and quality calibration then align candidate confidence with decoded interval quality and suppress high-scoring hard negatives.

We evaluate GLoR on TaCoS (Regneri et al., 2013), Ego4D-NLQ (Grauman et al., 2022), and Soccer-GMR (Ding et al., 2026). For Soccer-GMR, we extend videos to five minutes, preserving the annotations while increasing same-domain candidate density. GLoR achieves state-of-the-art performance on TaCoS and Ego4D-NLQ and outperforms all compared methods under this controlled protocol; ranking diagnostics attribute the gains mainly to more reliable Top-1 selection. Our contributions are threefold: (1) a ranking-oriented global-local decomposition for candidate ordering and temporal localization; (2) GRF and LBD for event support and boundary uncertainty, coupled with decoded-candidate ranking and calibration objectives; and (3) extensive evaluation across the three benchmarks, including the controlled five-minute Soccer-GMR protocol.

2 Related Work

Video Temporal Grounding. Early VTG methods generate temporal proposals and learn their compatibility with a query. MCN ranks sliding windows (Hendricks et al., 2017), whereas TALL jointly learns cross-modal alignment and boundary regression (Gao et al., 2017). Subsequent methods reduce proposal dependence and improve temporal structure. 2D-TAN propagates context over a two-dimensional start–end map (Zhang et al., 2020b), while VSLNet directly predicts a query-conditioned temporal span (Zhang et al., 2020a). Transformer-based methods further expand cross-modal context: Moment-DETR formulates grounding as set prediction (Lei et al., 2021), QD-DETR

strengthens query dependence (Moon et al., 2023), and EaTR introduces event-aware decoding (Jang et al., 2023). FlashVTG uses multi-scale temporal feature layering to improve coverage across moment durations (Cao et al., 2025). Unlike these end-to-end grounders, GLoR separates candidate ranking from boundary estimation and directly supervises the quality of decoded candidates.

Long-Video Temporal Grounding. Long-video grounding must preserve global context while controlling a large temporal search space. Retrieval-first approaches reduce the range processed by the localizer. CONE selects candidate windows before fine-grained alignment (Hou et al., 2023), RGNet jointly learns clip retrieval and frame-level grounding (Hannan et al., 2024), and DeCafNet delegates full-video scanning to a lightweight encoder (Lu et al., 2025). ReVisionLLM recursively narrows the temporal scope of hour-long videos (Hannan et al., 2025). State-space models provide another means of efficiently aggregating long-range context. VideoMamba applies selective state-space scanning to video sequences (Li et al., 2024), while HieraMamba constructs hierarchical anchor representations for temporal grounding (An & Grauman, 2026). Additional comparison systems include CG-DETR, UniVTG, SnAG, and HiTeA (Moon et al., 2026; Lin et al., 2023; Mu et al., 2024; Xu et al., 2026). These approaches primarily improve search efficiency or long-range representation. In contrast, GLoR retains dense candidates to avoid irreversible early retrieval errors and learns their final ordering.

Candidate Ranking and Localization Quality. The final VTG prediction is selected according to candidate confidence. Localization can therefore fail even when the candidate set contains a suitable interval. Query-dependent features improve relevance estimation (Moon et al., 2023), while hard-negative learning makes plausible mismatches more informative during training (Wang et al., 2022). Query relevance alone, however, does not guarantee that a high score corresponds to an accurately localized interval. Dense detection faces an analogous mismatch between classification confidence and localization quality. FCOS uses centerness to suppress unreliable locations (Tian et al., 2019); GFL represents localization through distributed targets (Li et al., 2020); and GFLv2 estimates localization quality from predicted distributions (Li et al., 2021). ActionFormer estimates deterministic boundary distances from multi-scale temporal points (Zhang et al., 2022), TriDet models relative boundary distributions around neighboring instants (Shi et al., 2023), and BAM-DETR refines sparse anchors and boundaries through parallel decoder pathways with localization-quality ranking (Lee & Byun, 2024). GLoR instead combines dense query-conditioned reference points, inside/center event support, directional boundary distributions, and candidate-quality supervision.

3 METHOD

3.1 PROBLEM FORMULATION

In video temporal grounding, given a video V, a query sentence Q, and video duration D, the model predicts a temporal segment $m = (s, e)$, where $0 \leq s < e \leq D$. We use pre-extracted video features $X \in \mathbb{R}^{T \times d_v}$ and query-token features $Q \in \mathbb{R}^{L \times d_q}$. The feature pyramid produces L_p levels $\{F^l\}_{l=0}^{L_p-1}$. At temporal location t of level l, the model predicts a confidence score $s_{l,t}$ and two temporal distances $d^s_{l,t}$ and $d^e_{l,t}$. PointGenerator expresses the reference point p_i and pyramid stride u_i in units of the base feature grid, and Δ is the duration represented by one base-grid step; hence the reference timestamp used by GRF is $t_i = \Delta p_i$. The decoded candidate in seconds is

$$m_i = \Delta \left(p_i - u_i d^s_i,\ p_i + u_i d^e_i \right). \tag{1}$$

At inference, both endpoints are clipped to $[0, D]$, matching the implementation and ensuring a valid temporal interval.

3.2 MODEL OVERVIEW

As shown in Figure 2, frozen feature encoders extract video and query features offline. The model projects these pre-extracted features into a shared embedding space. A cross-attention layer then aligns the two modalities, and the interacted video features are hierarchically transformed to construct a multi-scale feature pyramid. TFL converts the temporally fine-grained representation into L_p shared pyramid levels $\mathcal{P} = \{F^l\}_{l=0}^{L_p-1}$, where $F^l \in \mathbb{R}^{T_l \times d}$. PointGenerator assigns a temporal reference point p_i and stride u_i to every temporal location at each pyramid level.

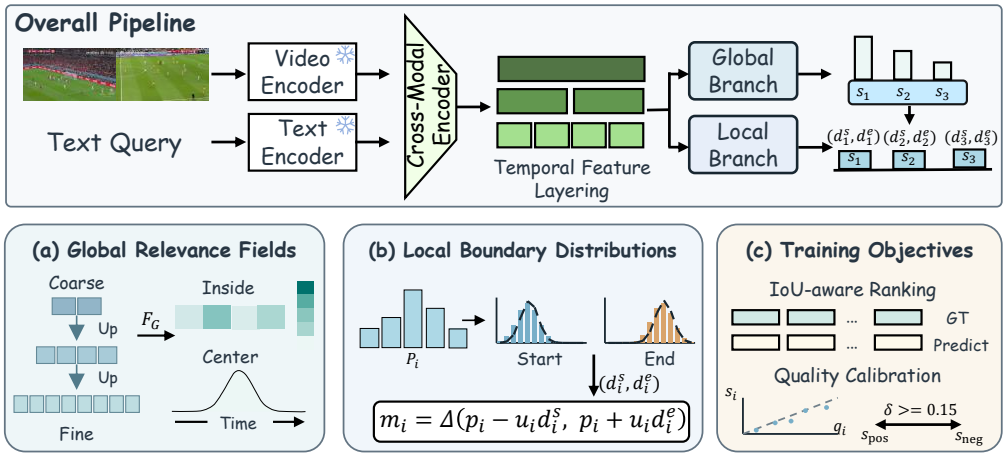

Figure 2: Overview of GLoR. A shared cross-modal temporal pyramid is divided into a global ranking stream and a local boundary stream. GRF produces inside and center relevance fields for ranking. LBD predicts independent start and end distance distributions for localization. Candidate-level training aligns the final ranking score with decoded interval quality.

GLoR splits each shared pyramid feature into a global ranking stream and a local boundary stream. The global stream aggregates dilated and coarse-to-fine context and uses GRF to produce ranking evidence. The local stream preserves temporal variations and uses LBD to predict start and end distance distributions; it receives only a detached relevance prior rather than global contextual features. Expected distances decode each candidate, whose score is supervised by candidate-level objectives. All reported GLoR models disable the inherited semantic-evidence injection, so both streams begin from the same cross-modal pyramid without a second token–frame reweighting stage.

3.3 Global Relevance Fields

Candidate ranking requires context beyond a local point classifier. A point can be visually active but outside the queried event, or it can overlap an event boundary without providing reliable evidence for the full interval. We first enhance each pyramid level with a dilated depthwise temporal operator D_l:

$$F^l_{\text{ctx}} = F^l + \alpha D_l(F^l), \tag{2}$$

where α is a shared learnable residual scale; its parameterization is detailed in Appendix A. Global context is propagated from coarse to fine levels. For $l < L_p - 1$, the fusion gate is

$$g_l = \sigma\left(G_l\left([F^l_{\text{ctx}}; \text{Up}(F^{l+1}_{\text{G}})]\right)\right), \tag{3}$$

and the global feature is

$$F^l_{\text{G}} = F^l_{\text{ctx}} + \alpha g_l \odot \text{Up}(F^{l+1}_{\text{G}}). \tag{4}$$

At the coarsest level, $F^{L_p-1}_{\text{G}} = F^{L_p-1}_{\text{ctx}}$. A level-wise class head and a confidence head over the concatenated global features $F^{\text{cat}}_{\text{G}} = [F^0_{\text{G}}; \ldots; F^{L_p-1}_{\text{G}}]$ yield the base ranking logit:

$$z_{\text{base},i} = \eta H_{\text{cls}}(F^l_{\text{G},i}) + (1-\eta) H_{\text{conf}}(F^{\text{cat}}_{\text{G}})_i, \tag{5}$$

where η is a learned unconstrained affine mixing scalar rather than a convex-mixture coefficient.

We introduce *Global Relevance Fields* (GRF) to make the score stream aware of event support. The inside field estimates whether a point lies within a queried event, while the center field estimates whether it lies near the event's temporally reliable core:

$$a^l_{\text{in},i} = H^l_{\text{in}}(F^l_{\text{G},i}), \tag{6}$$

$$a^l_{\text{ctr},i} = H^l_{\text{ctr}}(F^l_{\text{G},i}). \tag{7}$$

Their average forms an event-support residual:

$$a_{\text{evt},i} = \tfrac{1}{2}\left(a_{\text{in},i} + a_{\text{ctr},i}\right). \tag{8}$$

This residual refines confidence without hard-pruning candidates.

Let $\mathcal{G}$ denote the ground-truth intervals. The inside target is binary:

$$y_{\mathrm{in},i} = \max_{[s_g,e_g]\in\mathcal{G}} \mathbf{1}[s_g \le t_i \le e_g]. \tag{9}$$

The center target is a Gaussian soft label:

$$y_{\mathrm{ctr},i} = \max_{[s_g,e_g]\in\mathcal{G}} \exp\left(-\frac{(t_i - c_g)^2}{2\sigma_g^2}\right), \tag{10}$$

where $c_g = (s_g + e_g)/2$ and $\sigma_g = \max(0.2(e_g - s_g), \Delta)$. For $r \in \{\mathrm{in}, \mathrm{ctr}\}$, the relevance-field loss is

$$\mathcal{L}_r = \frac{\sum_i M_i(1 + 2y_{r,i})\,\mathrm{BCEWithLogits}(a_{r,i}, y_{r,i})}{\sum_i M_i}, \tag{11}$$

where M_i masks invalid pyramid positions. The global stream thereby models both event membership and center support for ranking.

3.4 Local Boundary Distributions

Boundary estimation has a different objective from global ranking. It should preserve local feature transitions, because the exact start or end can lie between two sampled temporal positions. We first apply a level-specific depthwise separable temporal operator L_l to the shared pyramid feature:

$$F^l_{\mathrm{B},0} = F^l + \alpha L_l(F^l). \tag{12}$$

The local branch excludes dilated and top-down features, using only a detached inside prior:

$$F^l_{\mathrm{B}} = F^l_{\mathrm{B},0} \odot \left[1 + \lambda_g\,\mathrm{sg}\left(\sigma(a^l_{\mathrm{in}})\right)\right], \tag{13}$$

where sg denotes stop-gradient. The prior amplifies globally supported locations without allowing boundary supervision to alter the global ranking representation.

Instead of direct offset regression, LBD predicts independent start and end distributions over $B = 32$ relative-distance bins with centers $\{r_b\}_{b=0}^{B-1}$ uniformly spaced from 0.25 to 31:

$$P^s_i = \mathrm{softmax}(H_s(F^l_{\mathrm{B},i})), \tag{14}$$

$$P^e_i = \mathrm{softmax}(H_e(F^l_{\mathrm{B},i})). \tag{15}$$

The offsets are decoded by the distribution expectations:

$$d^s_i = \sum_{b=0}^{B-1} r_b P^s_i(b), \tag{16}$$

$$d^e_i = \sum_{b=0}^{B-1} r_b P^e_i(b). \tag{17}$$

For a point assigned to ground-truth interval $[s_g, e_g]$, timestamps are converted to the base feature grid and the start/end distances are normalized by u_i and clipped to $[0.25, 31]$. A Gaussian label then distributes mass across neighboring bins:

$$y(b) = \frac{\exp(-(r_b - d^\star)^2/(2\sigma_b^2))}{\sum_{k=0}^{B-1}\exp(-(r_k - d^\star)^2/(2\sigma_b^2))}. \tag{18}$$

where σ_b is the fixed bandwidth. Only valid points inside a ground-truth interval are supervised; overlapping intervals use the nearest center. Out-of-support targets saturate at the endpoint bins without level-specific range filtering (Appendix A). Let $\mathcal{V}$ be this valid set. The distribution loss is

$$\mathcal{L}_{\mathrm{dist}} = -\frac{1}{2|\mathcal{V}|}\sum_{i\in\mathcal{V}}\sum_{b=0}^{B-1}\left[y^s_i(b)\log P^s_i(b) + y^e_i(b)\log P^e_i(b)\right]. \tag{19}$$

Rather than force a one-hot target on an unresolved temporal grid, we use the normalized distribution entropy as bounded boundary certainty:

$$q^{\mathrm{B}}_i = 1 - \frac{\mathcal{H}(P^s_i) + \mathcal{H}(P^e_i)}{2\log B}. \tag{20}$$

3.5 TRAINING OBJECTIVES

Point-level supervision does not require decoded-interval scores to follow temporal quality. The original dense classification, coordinate-regression, and saliency terms form $\mathcal{L}_{\text{base}}$: coordinate regression supervises intervals decoded from expected distances, while $\mathcal{L}_{\text{dist}}$ supervises the boundary distributions. GLoR further applies the two relevance-field losses and aligns confidence with the quality order of *decoded* candidates.

Candidate score construction. The global stream contributes semantic relevance through $z_{\text{base},i}$ and $a_{\text{evt},i}$. The local stream contributes only the bounded boundary-certainty scalar from Eq. 20:

$$z_i = z_{\text{base},i} + \lambda_e a_{\text{evt},i} + \lambda_b \operatorname{logit}(\operatorname{clip}(q_i^{\text{B}}, 0.05, 0.95)). \tag{21}$$

$$s_i = \sigma(z_i). \tag{22}$$

Boundary certainty provides only a constrained adjustment to semantic ranking.

IoU-aware ranking. The target quality of candidate i is its maximum ground-truth overlap:

$$q_i = \max_{g \in \mathcal{G}} \text{IoU}(m_i, g). \tag{23}$$

Decoded candidates are sorted by their current scores before NMS. The first $K_r = 64$ candidates form $\mathcal{I}_r$, and their detached target qualities define a ranking distribution:

$$p_i = \frac{\exp(q_i/\tau_r)}{\sum_{j \in \mathcal{I}_r} \exp(q_j/\tau_r)}. \tag{24}$$

The predicted score distribution is

$$h_i = \frac{\exp(s_i/\tau_r)}{\sum_{j \in \mathcal{I}_r} \exp(s_j/\tau_r)}, \tag{25}$$

where $\tau_r = 0.15$. The ranking loss is

$$\mathcal{L}_{\text{rank}} = -\sum_{i \in \mathcal{I}_r} p_i \log h_i. \tag{26}$$

Unlike binary point classification, this objective explicitly requires a higher-IoU decoded interval to receive a higher relative score.

Candidate-quality calibration. The first $K_q = 128$ score-sorted pre-NMS candidates form $\mathcal{I}_q$, for which score-IoU alignment is enforced by

$$\mathcal{L}_{\text{cal}} = \frac{1}{|\mathcal{I}_q|} \sum_{i \in \mathcal{I}_q} \text{SmoothL1}(s_i, q_i). \tag{27}$$

Candidates with $q_i \geq 0.5$ are positives and candidates with $q_i < 0.1$ are hard negatives. Let s_{pos} be the highest positive score and let $\mathcal{N}$ contain at most eight highest-scoring hard negatives. We use

$$\mathcal{L}_{\text{hard}} = \frac{1}{|\mathcal{N}|} \sum_{j \in \mathcal{N}} \text{softplus}(\delta + s_j - s_{\text{pos}}), \tag{28}$$

where $\delta = 0.15$. If a sample has no candidate with $q_i \geq 0.5$, the highest-IoU candidate is used as the positive fallback. If $\mathcal{N}$ is empty, we define $\mathcal{L}_{\text{hard}} = 0$ for that sample.

The complete objective is

$$\mathcal{L} = \mathcal{L}_{\text{base}} + \lambda_{\text{in}}\mathcal{L}_{\text{in}} + \lambda_{\text{ctr}}\mathcal{L}_{\text{ctr}} + \lambda_{\text{dist}}\mathcal{L}_{\text{dist}} + 0.1\mathcal{L}_{\text{rank}} + 0.1(\mathcal{L}_{\text{cal}} + \mathcal{L}_{\text{hard}}). \tag{29}$$

We set $\lambda_{\text{in}} = \lambda_{\text{ctr}} = \lambda_{\text{dist}} = 0.1$. Candidate IoUs are detached, and discrete pre-NMS Top-K membership receives no gradient; the candidate objectives therefore update selected scores without changing their IoU targets through boundary regression.

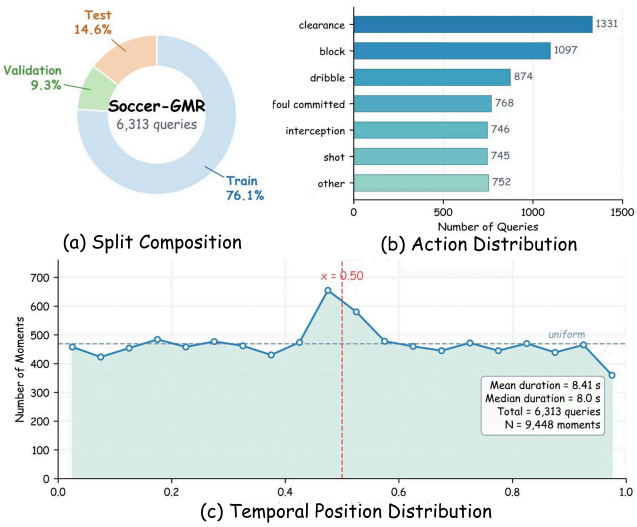

Figure 3: Statistics of controlled five-minute Soccer-GMR. Exact temporal offsets preserve source annotations while increasing in-domain candidate density.

4 Experiments

4.1 Datasets

We evaluate three complementary benchmarks. **TaCoS** contains 127 untrimmed cooking videos with temporally aligned descriptions and uses the standard splits (Regneri et al., 2013). **Ego4D-NLQ** comprises minute-long egocentric videos with daily-activity queries, viewpoint changes, and extended irrelevant operations (Grauman et al., 2022). **Soccer-GMR** provides densely annotated soccer scenes with recurring events (Ding et al., 2026). For the controlled five-minute protocol, a deterministic source-match hash assigns all clips from one match to one split. We then pair adjacent clips within each split, remove the second clip's overlapping prefix, and shift its annotations by the retained duration. Figure 3 summarizes these protocol statistics.

4.2 Evaluation Metrics

We report Recall@1 at IoU thresholds 0.3, 0.5, and 0.7, mean Average Precision (mAP), and mean Intersection over Union (mIoU). R1@0.3 measures Top-1 event retrieval, R1@0.5 standard localization, and R1@0.7 strict boundary accuracy. For multi-moment queries, R1 and mIoU use the best-matching ground truth, whereas mAP evaluates all relevant windows.

Similar-Distractor Error Rate (SDER). We report SDER on Soccer-GMR as a ranking diagnostic. Let $\hat{m}_q$ denote the Top-1 prediction for query q, $\mathcal{G}_q$ its ground-truth windows, and $\bar{v}(\cdot)$ the mean-pooled CLIP feature within a temporal window. With $\theta = 0.3$ and $\gamma = 0.85$, we define

$$\mathrm{SDER}(\theta, \gamma) = \frac{100}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathbf{1}\left[\max_{g \in \mathcal{G}_q} \mathrm{IoU}(\hat{m}_q, g) < \theta\right] \mathbf{1}\left[\max_{g \in \mathcal{G}_q} \cos(\bar{v}(\hat{m}_q), \bar{v}(g)) \geq \gamma\right]. \tag{30}$$

SDER measures the percentage of original queries whose Top-1 prediction fails to cover any relevant temporal window but remains visually similar to at least one relevant window; lower is better. It is used only as a diagnostic metric for identifying similar distractors and applies to both single-moment and multi-moment queries. SDER does not replace the standard VTG metrics.

4.3 Implementation Details

We follow dataset-specific feature protocols. Soccer-GMR and TaCoS use pre-extracted SlowFast video features and the dense baseline's text protocol (Feichtenhofer et al., 2019; Radford et al., 2021); Ego4D-NLQ uses public EgoVLP video features and 512-dimensional CLIP text features

Table 1: Validation and test results on controlled five-minute Soccer-GMR.

	Validation					Test				
Method	R1@.3	R1@.5	R1@.7	mAP	mIoU	R1@.3	R1@.5	R1@.7	mAP	mIoU
Moment-DETR 2021	9.86	6.46	3.06	3.61	6.86	7.06	4.89	2.39	3.18	4.80
CG-DETR 2026	29.25	22.96	13.10	14.04	19.99	23.89	17.48	9.55	11.99	16.46
UniVTG 2023	27.38	22.11	10.88	13.92	18.47	30.84	24.32	12.60	14.68	20.37
FlashVTG 2025	25.17	20.41	12.41	12.94	18.06	27.36	21.06	12.27	13.56	19.10
SnAG 2024	29.25	24.97	15.16	16.98	21.56	29.57	23.71	14.70	15.84	21.07
DeCafNet 2025	**30.27**	25.04	15.71	17.67	21.94	30.67	**24.45**	14.98	16.07	21.11
GLoR	30.10	**25.51**	**17.86**	**18.30**	**22.52**	**32.79**	24.21	**15.74**	**18.27**	**22.40**

Table 2: TaCoS test results.

Method	R1@.3	R1@.5	R1@.7	mIoU
2D-TAN 2020b	40.01	27.99	12.92	27.22
VSLNet 2020a	35.54	23.54	13.15	24.99
Moment-DETR 2021	37.97	24.67	11.97	25.49
CG-DETR 2026	52.23	39.61	22.23	36.48
UniVTG 2023	51.44	34.97	17.35	33.60
FlashVTG 2025	53.71	41.76	24.74	37.61
GLoR	**54.86**	**43.36**	**28.74**	**39.39**

Table 3: Ego4D-NLQ validation results.

Method	R1@.3	R1@.5	mIoU
2D-TAN 2020b	4.33	1.83	3.39
Moment-DETR 2021	4.34	1.81	3.53
UniVTG 2023	7.28	3.95	4.91
FlashVTG 2025	13.43	8.86	9.50
HiTeA 2026	10.39	6.04	8.12
RGNet 2024	14.21	9.19	9.73
DeCafNet 2025	15.04	10.42	10.61
GLoR	**15.57**	**10.70**	**11.13**

(Lin et al., 2022; Radford et al., 2021). TaCoS and Ego4D-NLQ follow their official evaluation protocols. Soccer-GMR uses original-query evaluation: each sample includes all relevant windows and one ranked list. All methods share videos, shifted annotations, features, splits, and post-processing.

4.4 Comparison Results

Table 1 reports the Soccer-GMR results. On validation, GLoR leads R1@0.5, R1@0.7, mAP, and mIoU by 0.47, 2.15, 0.63, and 0.58 points over the respective runners-up, while trailing DeCafNet by 0.17 points at R1@0.3. On test, it leads R1@0.3, R1@0.7, mAP, and mIoU by 1.95, 0.76, 2.20, and 1.29 points, respectively, and trails the best R1@0.5 by 0.24 points. These gains support improved candidate ordering and accurate boundaries. Figure 4 provides qualitative examples.

Table 2 shows that GLoR leads all four TaCoS metrics, improving over FlashVTG by 1.15, 1.60, 4.00, and 1.78 points on R1@0.3, R1@0.5, R1@0.7, and mIoU, respectively. The largest gain at R1@0.7 supports distributional boundary prediction under strict overlap.

On Ego4D-NLQ, Table 3 shows that GLoR exceeds DeCafNet by 0.53, 0.28, and 0.52 points on R1@0.3, R1@0.5, and mIoU, respectively, demonstrating transfer to long egocentric videos.

Overall, the consistent gains across all three protocols support the generality of combining global candidate ranking with local boundary estimation.

4.5 Candidate Ranking Error Analysis

Table 4 compares H@1@0.3 and SDER on the Soccer-GMR validation and test splits. GLoR obtains the lowest SDER, at 68.54% and 65.91%, reducing the error rate relative to DeCafNet by 0.34 and 1.60 points, respectively. Together with the Top-1 results, this indicates more reliable selection and fewer visually similar incorrect predictions. Appendix C reports additional ranking diagnostics.

4.6 Ablation Study

We progressively add global-local decoupling (GLD), global relevance fields (GRF), local boundary distributions (LBD), and candidate-level training objectives (CTO) to the same dense baseline. Ta-

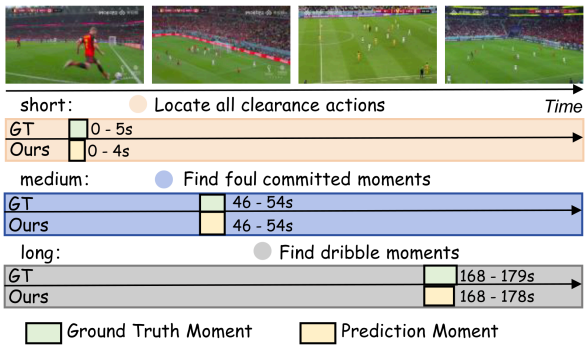

Figure 4: Qualitative results under the controlled five-minute Soccer-GMR protocol.

Table 4: Similar-distractor ranking on the Soccer-GMR validation and test splits.

	Validation		Test	
Method	H@1@.3↑	SDER↓	H@1@.3↑	SDER↓
Moment-DETR 2021	9.86	81.12	7.06	78.28
CG-DETR 2026	29.25	69.56	23.89	73.94
UniVTG 2023	27.38	70.75	<u>30.84</u>	67.64
FlashVTG 2025	25.17	73.47	27.36	71.01
SnAG 2024	29.25	69.73	29.57	67.63
DeCafNet 2025	**30.27**	<u>68.88</u>	30.67	<u>67.51</u>
GLoR	<u>30.10</u>	**68.54**	**32.79**	**65.91**

ble 5 reports the Soccer-GMR validation results. Starting from GLD, GRF improves R1@0.7, mAP, and mIoU by 1.87, 0.10, and 0.51 points, supporting explicit interior and center evidence. LBD then improves R1@0.3, R1@0.5, mAP, and mIoU, showing the benefit of distributional boundaries. Finally, CTO produces the best R1@0.3, R1@0.5, R1@0.7, and mIoU. The intermediate trade-offs indicate that GRF and LBD capture complementary aspects of candidate quality, while candidate-level supervision converts them into more reliable Top-1 predictions.

Table 5: Ablation on the Soccer-GMR validation split.

GLD	GRF	LBD	CTO	R1@.3	R1@.5	R1@.7	mAP	mIoU
✓				26.53	22.96	14.12	17.45	19.24
✓	✓			26.53	22.96	15.99	17.55	19.75
✓	✓	✓		28.91	23.30	14.46	**18.53**	20.98
✓	✓	✓	✓	**30.10**	**25.51**	**17.86**	18.30	**22.52**

5 Conclusion

In this paper, we presented GLoR, a ranking-oriented VTG model that decouples global event scoring from local distributional boundary prediction. The global branch uses relevance evidence to estimate the match between each candidate and the query, while the local branch predicts start and end boundary distributions to obtain precise temporal boundaries. Together with IoU-aware ranking and candidate-quality calibration, this design improves Top-1 selection over the matched dense baseline. GLoR obtains the strongest results among the listed methods on TaCoS and Ego4D-NLQ and leads four of the five reported metrics on both splits of the controlled five-minute Soccer-GMR protocol. The complete ranking analysis further shows that candidate coverage and list-wide ordering remain open directions. Future work will explore diverse candidate generation, compositional query understanding, and efficient long-range reasoning.

REPRODUCIBILITY STATEMENT

The main paper specifies the feature protocols, model architecture, evaluation metrics, and controlled Soccer-GMR construction. Appendix A provides the complete optimization, training-schedule, checkpoint-selection, and inference settings. Appendix C reports seed-wise results and statistical analyses. The supplementary material provides anonymized reference code for the core GLoR components, including hierarchical global context aggregation, GRF, LBD, score fusion, and candidate-level training objectives.

AI USE STATEMENT

AI-assisted tools were used for retrieval and discovery of related work, text refinement, and code assistance. Details of these uses are described in the paper. The authors verified the results and take responsibility for the final content.

REFERENCES

Joungbin An and Kristen Grauman. Hieramamba: Video temporal grounding via hierarchical anchor-mamba pooling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16954–16965, 2026.

Zhuo Cao, Bingqing Zhang, Heming Du, Xin Yu, Xue Li, and Sen Wang. Flashvtg: Feature layering and adaptive score handling network for video temporal grounding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 9208–9218, 2025.

Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. Temporally grounding natural sentence in video. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 162–171. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1015.

Yiming Ding, Siyu Cao, Luyuan Jiao, Yixuan Li, Zitong Wang, Zhiyong Liu, and Lu Zhang. Retrieving any relevant moments: Benchmark and models for generalized moment retrieval. *arXiv preprint arXiv:2605.02623*, 2026. doi: 10.48550/arXiv.2605.02623.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6202–6211, 2019.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5267–5275, 2017. doi: 10.1109/ICCV.2017.563.

Kristen Grauman et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18995–19012, 2022.

Tanveer Hannan, Md Mohaiminul Islam, Thomas Seidl, and Gedas Bertasius. RGNet: A unified clip retrieval and grounding network for long videos. In *Computer Vision – ECCV 2024*, pp. 352–369. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72664-4_20.

Tanveer Hannan, Md Mohaiminul Islam, Jindong Gu, Thomas Seidl, and Gedas Bertasius. ReVisionLLM: Recursive vision-language model for temporal grounding in hour-long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19012–19022, 2025. doi: 10.1109/CVPR52734.2025.01771.

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.

Zhijian Hou, Wanjun Zhong, Lei Ji, Difei Gao, Kun Yan, W. K. Chan, Chong-Wah Ngo, Mike Zheng Shou, and Nan Duan. Cone: An efficient coarse-to-fine alignment framework for long video temporal grounding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 8013–8028. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.445.

Jinhyun Jang, Jungin Park, Jin Kim, Hyeongjun Kwon, and Kwanghoon Sohn. Knowing where to focus: Event-aware transformer for video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13846–13856, 2023.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002. doi: 10.1145/582415.582418.

Pilhyeon Lee and Hyeran Byun. BAM-DETR: Boundary-aligned moment detection transformer for temporal sentence grounding in videos. In *Computer Vision – ECCV 2024*, pp. 220–238. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72627-9_13.

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

Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. VideoMamba: State space model for efficient video understanding. In *Computer Vision – ECCV 2024*, pp. 237–255. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-73347-5_14.

Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21002–21012, 2020.

Xiang Li, Wenhai Wang, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss v2: Learning reliable localization quality estimation for dense object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11632–11641, 2021.

Zeqian Li, Shangzhe Di, Zhonghua Zhai, Weilin Huang, Yanfeng Wang, and Weidi Xie. Universal video temporal grounding with generative multi-modal large language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2159.

Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, Chengfei Cai, Hongfa Wang, Dima Damen, Bernard Ghanem, Wei Liu, and Mike Zheng Shou. Egocentric video-language pretraining. In *Advances in Neural Information Processing Systems*, volume 35, pp. 7575–7586, 2022.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2794–2804, 2023.

Zijia Lu, A. S. M. Iftekhar, Gaurav Mittal, Tianjian Meng, Xiawei Wang, Cheng Zhao, Rohith Kukkala, Ehsan Elhamifar, and Mei Chen. Decafnet: Delegate and conquer for efficient temporal grounding in long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24066–24076, 2025.

WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23023–23033, 2023. doi: 10.1109/CVPR52729.2023.02205.

WonJun Moon, Sangeek Hyun, Subeen Lee, and Jae-Pil Heo. Correlation-guided calibration of query dependency for video temporal grounding. *Pattern Recognition*, 174:112984, 2026. doi: 10.1016/j.patcog.2025.112984.

Fangzhou Mu, Sicheng Mo, and Yin Li. Snag: Scalable and accurate video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18930–18940, 2024.

Yulin Pan, Xiangteng He, Biao Gong, Yiliang Lv, Yujun Shen, Yuxin Peng, and Deli Zhao. Scanning only once: An end-to-end framework for fast temporal grounding in long videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13767–13777, 2023. doi: 10.1109/ICCV51070.2023.01266.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013. doi: 10.1162/tacl_a_00207.

Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, and Dacheng Tao. TriDet: Temporal action detection with relative boundary modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18857–18866, 2023.

Jiajin Tang, Zhengxuan Wei, Yuchen Zhu, Cheng Shi, Guanbin Li, Liang Lin, and Sibei Yang. Sim-DETR: Unlock detr for temporal sentence grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22760–22771, 2025. doi: 10.1109/ICCV51701.2025.02113.

Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9627–9636, 2019. doi: 10.1109/ICCV.2019.00972.

Ellen M. Voorhees. The TREC-8 question answering track report. In *Proceedings of the Eighth Text REtrieval Conference*, volume 500-246 of *NIST Special Publication*, pp. 77–82, 1999.

Zhenzhi Wang, Limin Wang, Tao Wu, Tianhao Li, and Gangshan Wu. Negative sample matters: A renaissance of metric learning for temporal grounding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3):2613–2623, 2022. doi: 10.1609/aaai.v36i3.20163.

Xinyi Xu, Hongsong Wang, Guo-Sen Xie, Caifeng Shan, and Fang Zhao. HiTeA: Hierarchical temporal alignment for training-free long-video temporal grounding. In *International Conference on Learning Representations*, 2026.

Chen-Lin Zhang, Jianxin Wu, and Yin Li. ActionFormer: Localizing moments of actions with transformers. In *Computer Vision – ECCV 2022*, pp. 492–510. Springer Nature Switzerland, 2022. doi: 10.1007/978-3-031-19772-7_29.

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. Association for Computational Linguistics, 2020a. doi: 10.18653/v1/2020.acl-main.585.

Jinglei Zhang, Yuanfan Guo, Rolandos Alexandros Potamias, Jiankang Deng, Hang Xu, and Chao Ma. VTimeCoT: Thinking by drawing for video temporal grounding and reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24203–24213, 2025. doi: 10.1109/ICCV51701.2025.02243.

Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(7):12870–12877, 2020b. doi: 10.1609/aaai.v34i07.6984.

A Detailed Experimental Settings

Common optimization and inference settings. We train GLoR with AdamW using an initial learning rate of 2×10^{-4}, weight decay of 10^{-4}, and gradient-norm clipping at 0.1. A StepLR scheduler multiplies the learning rate by 0.5 after epoch 70. Validation is performed every three epochs with an evaluation batch size of one. The hidden width is 256, the input dropout is 0.5, and the temporal heads use the dataset-specific dropout shown in Table 6. At inference, we retain the 50 highest-scoring candidates before NMS, apply temporal NMS at IoU 0.7, and retain at most 10 predictions per query. Checkpoints on TaCoS and Soccer-GMR are selected by validation R1@0.3. The Ego4D-NLQ checkpoint is selected by the mean of validation R1@0.5 and R1@0.7.

Table 6: Dataset-specific training settings. Δt is the temporal interval represented by one input video token. Feature encoders are frozen.

Dataset	Video feature	Text feature	d_v/d_q	Δt (s)
TaCoS	SlowFast+CLIP	CLIP	2816/512	2.000
Soccer-GMR	SlowFast+CLIP	CLIP	2816/512	2.000
Ego4D-NLQ	EgoVLP	CLIP	256/512	0.534759

Dataset	Batch	Epochs	Dropout	Selection
TaCoS	32	150	0.1	Val R1@.3
Soccer-GMR	32	130	0.1	Val R1@.3
Ego4D-NLQ	4	150	0.0	Mean Val R1@.5/.7

GLoR-specific settings. The local boundary head uses 32 bins uniformly spaced over $[0.25, 31]$ and Gaussian targets with standard deviation 1.0. Ground-truth timestamps are divided by the feature interval Δ, and each start/end distance is further divided by the point stride u_i before being clipped to the bin support. The LBD loss supervises valid points inside an assigned interval and does not apply PointGenerator's level-specific regression-range filter; out-of-support targets therefore saturate at the nearest endpoint bin. The IoU-aware ranking objective uses the highest-scoring pre-NMS $K_r = 64$ candidates and temperature 0.15, while candidate-quality calibration uses $K_q = 128$ and a hard-negative margin of 0.15. Candidate IoUs are detached, and Top-K membership is a discrete selection recomputed in each forward pass.

The residual scale α is shared across levels, initialized to 0.01, and clipped to $[0, 0.1]$. The affine class/confidence mixing scalar η is initialized to 0.5 and is not constrained to $[0, 1]$. The inside-prior weight is initialized to 0.05 and clipped to $[0, 0.5]$; the GRF and boundary-certainty score weights are initialized to 0.05 and 0.02 and clipped to $[0, 0.25]$. The global-inside, global-center, boundary-distribution, IoU-ranking, and candidate-quality losses each have weight 0.1; the remaining base losses follow the dense baseline. These settings are shared across the three datasets.

B Controlled Five-Minute Soccer-GMR Protocol

The controlled protocol uses positive-query clips from the official Soccer-GMR annotations. Each source clip is approximately 150 seconds long. A deterministic hash of the source identifier first assigns every clip from one match to a single train, validation, or test split. Within each split, we pair temporally adjacent clips from the same match and half when their overlap is between 8 and 12 seconds. Removing the second clip's overlapping prefix produces a 290-second video. This source-level assignment prevents cross-split match leakage, while same-match pairing preserves continuity and increases the density of competing in-domain events.

For a pair whose first clip begins at a and second clip begins at b, windows from the first clip retain their timestamps, while windows from the second clip are shifted by $b - a$ (normally 140 seconds). Query strings are normalized and matched exactly. If the same query occurs in both clips, all corresponding windows are retained as one multi-moment query; duplicate or overlapping windows within 0.25 seconds are merged. Thus, one evaluation sample contains one original query and all of its relevant temporal windows, and a method returns one ranked candidate list for that

query. The protocol contains 1,078 videos, 6,313 queries, 9,448 annotated windows, and 16 action classes. Table 7 gives the split statistics.

Table 7: Statistics of the controlled five-minute Soccer-GMR protocol.

Split	Videos	Queries	Windows
Train	830	4,804	7,213
Validation	104	588	866
Test	144	921	1,369
Total	1,078	6,313	9,448

All compared methods use the same paired videos, split assignments, shifted annotations, offline features, original-query evaluation unit, Top-K evaluation cutoffs, and NMS protocol. Each method retains its native candidate-generation mechanism, and we quantify the resulting candidate coverage with Oracle Recall. R1 and mIoU compare the Top-1 prediction with the best-matching relevant window, while mAP evaluates the ranked list against all relevant windows.

For Soccer-GMR, mAP follows the official positive-query localization protocol. We retain at most ten score-ranked predictions per query after temporal NMS and compute average precision independently at temporal-IoU thresholds $\{0.50, 0.55, \ldots, 0.95\}$. Predictions are matched one-to-one to relevant windows at each threshold, and mAP is the mean over queries and thresholds. R1 and mIoU use the pre-NMS Top-1 prediction, matching the values reported in the main tables.

We additionally audit the construction at the source-video, paired-video, and source-query levels. Every pairwise split comparison has zero shared identifiers for all three units, which rules out cross-split leakage introduced by pairing. All resulting videos are 290 seconds long. The fractions of annotated windows whose normalized centers fall within $[0.475, 0.525]$ are 8.98%, 8.31%, and 5.04% for train, validation, and test, respectively. The concentration around the join is therefore most visible in training. These windows retain source annotations from the temporally overlapping adjacent clips; the construction does not synthesize a boundary-centered label.

C Robustness and Statistical Analysis

Random-seed robustness. We train three Soccer-GMR models with seeds 2024, 2025, and 3408. Every run uses 130 epochs and the same validation R1@0.3 checkpoint rule. Table 8 reports the exact selected-checkpoint results and the sample mean and standard deviation. All three checkpoints were selected without access to test labels.

Table 8: Three-seed robustness on the controlled Soccer-GMR protocol. mAP is measured after NMS; the other metrics use the pre-NMS ranked predictions, matching the main tables.

Seed	R1@.3	R1@.5	R1@.7	mAP	mIoU
			Validation		
2024	30.10	25.51	17.86	18.30	22.52
2025	29.93	25.51	16.16	18.10	21.86
3408	29.59	25.00	17.86	18.49	21.93
Mean $\pm$ std	29.87$\pm$0.26	25.34$\pm$0.29	17.29$\pm$0.98	18.30$\pm$0.20	22.10$\pm$0.36
			Test		
2024	32.79	24.21	15.74	18.27	22.40
2025	32.25	25.73	15.09	18.47	22.19
3408	32.18	24.02	14.79	18.27	21.93
Mean $\pm$ std	32.41$\pm$0.33	24.65$\pm$0.94	15.21$\pm$0.49	18.34$\pm$0.12	22.17$\pm$0.24

Paired bootstrap significance. We compare GLoR with FlashVTG using 10,000 paired bootstrap resamples over original queries (bootstrap seed 20260915). We use percentile 95% confidence inter-

vals and two-sided paired bootstrap p-values with plus-one correction. Holm correction is applied across the five metrics within each split. Table 9 shows that every confidence interval excludes zero and every corrected p-value is below 0.05.

Table 9: Paired bootstrap differences between GLoR and FlashVTG on Soccer-GMR. Positive values favor GLoR.

Split	Metric	Δ (95% CI)	Holm p
Val	R1@.3	4.93 [1.36, 8.50]	.0074
Val	R1@.5	5.10 [1.70, 8.50]	.0072
Val	R1@.7	5.44 [2.38, 8.50]	.0010
Val	mAP	3.00 [1.44, 4.59]	.0010
Val	mIoU	4.45 [2.03, 6.94]	.0018
Test	R1@.3	5.43 [2.50, 8.36]	.0020
Test	R1@.5	3.15 [0.21, 6.08]	.0424
Test	R1@.7	3.47 [0.87, 6.08]	.0156
Test	mAP	1.85 [0.53, 3.19]	.0156
Test	mIoU	3.30 [1.15, 5.52]	.0128

SDER threshold sensitivity. Table 10 varies the CLIP-similarity threshold while keeping the IoU error threshold at 0.3. GLoR yields lower SDER than FlashVTG at every tested threshold on both splits, so the ranking conclusion is not tied to the main threshold of 0.85.

Table 10: SDER (%, lower is better) under different CLIP-similarity thresholds.

	CLIP similarity threshold					
Model/split	.70	.75	.80	.85	.90	.95
FlashVTG/Val	74.83	74.83	74.32	73.47	69.73	56.80
GLoR/Val	**69.90**	**69.73**	**69.39**	**68.54**	**64.46**	**52.21**
FlashVTG/Test	72.64	72.53	71.88	71.01	67.75	55.48
GLoR/Test	**67.21**	**67.10**	**66.56**	**65.91**	**61.89**	**50.27**

Candidate availability and whole-list ranking. SDER diagnoses a particular Top-1 failure mode but does not characterize the complete candidate list. We therefore evaluate every method at the same cutoff of five predictions using Oracle R@5@0.3, thresholded MRR@5, and IoU-graded NDCG@5. MRR and NDCG are standard measures of ranked retrieval quality (Voorhees, 1999; Järvelin & Kekäläinen, 2002).

For query q, let $v_{q,r} = \max_{g \in \mathcal{G}_q} \mathrm{IoU}(m_{q,r}, g)$ be the quality of the candidate at rank r, let $\theta = 0.3$, and let $r_q^\star = \min\{r \le K : v_{q,r} \ge \theta\}$ when such a candidate exists. We define

$$\mathrm{OracleR@K@}\theta = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathbf{1}\left[\max_{r \le K} v_{q,r} \ge \theta\right], \tag{31}$$

$$\mathrm{MRR@K} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathbf{1}[r_q^\star \text{ exists}] \frac{1}{r_q^\star}. \tag{32}$$

For graded ranking quality, $\mathrm{DCG@K}_q = \sum_{r=1}^{K} (2^{v_{q,r}} - 1)/\log_2(r+1)$ and $\mathrm{NDCG@K}_q$ divides this value by the DCG of the ideal IoU ordering; NDCG@K is averaged over queries. We additionally report selection efficiency, $\mathrm{R1@0.3}/\mathrm{OracleR@5@0.3}$, as an explanatory ratio rather than a standard benchmark metric. It measures how often an available valid candidate is promoted to rank one.

Table 11 shows that GLoR has the highest selection efficiency on both splits. Retrieval-based methods obtain higher Oracle recall and, consequently, higher MRR or NDCG. The evidence therefore supports a specific claim: GLoR improves Top-1 selection from its generated candidates, rather than universally dominating candidate generation or every whole-list ranking measure.

Table 11: Candidate availability and Top-5 ordering on the controlled Soccer-GMR protocol. Efficiency is Top-1 recall divided by Oracle Top-5 recall.

Split	Method	R1@.3	Oracle R@5@.3	Efficiency	MRR@5	NDCG@5
Val	Moment-DETR	9.86	22.96	0.430	0.145	0.229
	CG-DETR	29.25	61.05	0.479	0.409	0.507
	UniVTG	27.38	47.79	0.573	0.347	0.231
	FlashVTG	25.17	46.94	0.536	0.322	0.198
	SnAG	29.25	67.52	0.433	**0.439**	**0.559**
	DeCafNet	**30.27**	64.12	0.472	0.429	0.557
	GLoR	<u>30.10</u>	50.17	**0.600**	0.369	0.235
Test	Moment-DETR	7.06	20.96	0.337	0.121	0.194
	CG-DETR	23.89	62.54	0.382	0.374	0.481
	UniVTG	<u>30.84</u>	57.44	0.541	0.406	0.252
	FlashVTG	27.36	49.08	0.558	0.345	0.215
	SnAG	29.57	75.57	0.431	0.487	0.576
	DeCafNet	30.67	75.79	0.446	**0.494**	**0.579**
	GLoR	**32.79**	54.29	**0.604**	0.399	0.249

D Ablation and Hyperparameter Settings

Validation ablation. Table 12 reports the validation counterpart of the sequential ablation in the main paper. GLD establishes separate global ranking and local boundary streams. GRF adds explicit inside and center support, LBD represents boundary uncertainty, and CTO aligns confidence with decoded-candidate quality. Their complementary effects remain consistent on the validation split.

Table 12: Validation ablation on the controlled five-minute Soccer-GMR protocol.

GLD	GRF	LBD	CTO	R1@.3	R1@.5	R1@.7	mAP	mIoU
✓				26.53	22.96	14.12	17.45	19.24
✓	✓			26.53	22.96	15.99	17.55	19.75
✓	✓	✓		28.91	23.30	14.46	**18.53**	20.98
✓	✓	✓	✓	**30.10**	**25.51**	**17.86**	18.30	**22.52**

Candidate-budget sensitivity. We vary the ranking and quality-calibration budgets on the Soccer-GMR validation split while keeping all other settings fixed. Table 13 shows that GLoR remains stable across the tested budgets. The smaller budget favors R1@0.3, R1@0.5, and mIoU, the default budget gives the best R1@0.7, and the larger budget gives the best mAP. Thus, the gains do not depend on a single candidate cutoff.

Table 13: Candidate-budget sensitivity on the Soccer-GMR validation split.

K_r	K_q	R1@.3	R1@.5	R1@.7	mAP	mIoU
32	64	**31.97**	**26.19**	17.35	18.55	**22.97**
64	128	30.10	25.51	**17.86**	18.30	22.52
128	256	30.27	25.00	15.48	**18.59**	21.47

Loss-weight sensitivity. We jointly vary λ_{rank} and λ_{qual} with $K_r = 64$ and $K_q = 128$. The zero-weight setting removes CTO while retaining GLD, GRF, and LBD. As shown in Table 14, the default weight of 0.1 achieves the strongest result on four of the five metrics, supporting the selected balance between candidate-level supervision and the base objectives.

Table 14: Candidate-level loss-weight sensitivity on the Soccer-GMR validation split.

$\lambda_{\text{rank}} = \lambda_{\text{qual}}$	R1@.3	R1@.5	R1@.7	mAP	mIoU
0	28.91	23.30	14.46	**18.53**	20.98
0.05	28.57	23.98	15.65	17.61	20.66
0.10	**30.10**	**25.51**	**17.86**	18.30	**22.52**

Boundary-certainty score ablation. The boundary-distribution entropy measures prediction sharpness rather than correctness: a confident but misplaced distribution can still have low entropy. We therefore remove only the entropy-derived score term at inference while retaining the trained LBD distributions, expected-distance decoding, and all candidate boundaries. Table 15 shows that removing this term leaves or slightly improves every Top-1 metric, except for a negligible 0.02-point test mAP reduction. We consequently treat boundary entropy as an auxiliary scoring diagnostic rather than evidence that boundary regression itself is improved.

Table 15: Inference-only ablation of the entropy-derived boundary-certainty score on controlled five-minute Soccer-GMR. Boundary predictions and model weights are unchanged.

Split	Score	R1@.3	R1@.5	R1@.7	mAP	mIoU
Val	Full score	30.10	25.51	17.86	18.30	22.52
Val	Without entropy	30.10	25.51	**18.03**	**18.32**	**22.60**
Test	Full score	32.79	24.21	15.74	**18.27**	22.40
Test	Without entropy	**32.90**	**24.43**	15.74	18.25	**22.47**

E Relation to Prior Dual-Path Designs

Table 16 separates the proposed design from representative classification–regression and dual-path localization systems. Branch separation is not, by itself, our novelty claim. GLoR combines dense multi-scale candidates with event-support fields, directional boundary distributions, and losses defined on the quality order of decoded intervals. This combination differs from the deterministic point regression of ActionFormer (Zhang et al., 2022) and the sparse anchor-query refinement of BAM-DETR (Lee & Byun, 2024).

Table 16: Architectural distinction from representative localization designs.

Method	Candidate representation	Ranking evidence
ActionFormer	Dense pyramid points	Point classification
BAM-DETR	Sparse learned anchor queries	Global-anchor quality
GLoR	Dense multiscale points	Context + inside/center GRF

Method	Boundary representation	Candidate-level ordering
ActionFormer	Deterministic distances	No
BAM-DETR	Boundary-focused offsets	Quality-aware proposal ranking
GLoR	Start/end categorical distributions	IoU ordering + calibration

F Computational Cost

We profile one model forward pass with batch size one, 145 video tokens, and 32 text tokens. The video and text input dimensions are fixed to 2816 and 512, respectively. Offline feature extraction, training losses, decoding, and NMS are excluded. We use `fvcore.nn.FlopCountAnalysis`; one multiply–accumulate is counted once as a GMAC, and we report the conventional two-FLOP-per-MAC conversion as GFLOPs. Panel (a) of Table 17 is the strictly matched comparison used in

our analysis. All methods in this panel receive the same input tensor sizes and are measured by the same profiler. Panel (b) reports additional reference measurements at the native execution lengths required by the released implementations. These native-resolution values are not used for direct efficiency claims against GLoR.

Table 17: Model-forward complexity. Panel (a) uses a unified batch size and temporal input length. Panel (b) is included only to document the native execution cost of implementations that were not profiled at the matched length. GFLOPs counts each multiply–add as two floating-point operations.

Method	Params (M)	GMACs	GFLOPs
(a) Matched input: $B = 1$, $T_v = 145$, $T_q = 32$			
Moment-DETR(Lei et al., 2021)	4.823	0.487	0.974
CG-DETR(Moon et al., 2026)	12.618	0.947	1.894
FlashVTG(Cao et al., 2025)	13.112	1.864	3.728
UniVTG(Lin et al., 2023)	43.361	7.039	14.078
GLoR	15.970	2.000	4.010
(b) Native-resolution references (not directly comparable)			
SnAG(Mu et al., 2024) ($T_v = 256$)	10.902	1.958	3.916
DeCafNet(Lu et al., 2025) ($T_v = 288$)	9.462	2.589	5.178
2D-TAN(Zhang et al., 2020b) ($T_v = 256$)	60.705	1066.255	2132.510

Under the strictly matched protocol, the current GLoR implementation adds 2.858M parameters and 0.136 GMACs (0.282 GFLOPs) over FlashVTG. The additional computation comes from the global relevance fields, local boundary-distribution heads, and candidate-quality computations.

We also measure wall-clock model-forward cost on one NVIDIA A100-SXM4-40GB in FP32. Every run uses batch size one, 100 warm-up iterations, and 1,000 timed iterations. Table 18 follows the same matched/native separation as the operation-count table. GLoR requires 18.17 ms per query and 75.3 MiB peak allocated memory at the matched length. It is slower than FlashVTG because the relevance and distribution heads execute at every dense reference point, but still processes 54.93 queries per second without offline feature extraction.

Table 18: Model-forward runtime on one A100 GPU in FP32. Feature extraction, training losses, decoding, and NMS are excluded.

Method	Mean (ms)	P95 (ms)	Query/s	Peak MiB
(a) Matched input: $B = 1$, $T_v = 145$, $T_q = 32$				
Moment-DETR	3.59	3.63	276.13	32.2
CG-DETR	9.29	9.60	107.37	62.7
FlashVTG	10.35	10.75	96.37	63.9
UniVTG	3.12	3.14	317.80	195.2
GLoR	18.17	24.29	54.93	75.3
(b) Native-resolution references (not directly comparable)				
SnAG ($T_v = 256$)	36.72	40.15	27.21	183.8
DeCafNet ($T_v = 288$)	33.25	34.43	30.04	179.2