EVENTFIELDNET: STRUCTURED EVENT SCORING FOR VIDEO MOMENT RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Highly overlapping candidates pose a ranking challenge in video moment retrieval: truncated, overextended, and well-aligned intervals can share query-relevant content yet differ in event coverage and excess context. We propose **EventFieldNet**, which makes semantic and temporal ranking criteria explicit through three complementary fields that augment learned base span scores. **Evidence** aggregates learned clip-level relevance across the entire candidate interval; **Support** assesses event persistence within the interval relative to local context derived from the model's own predictions; and **Transition** separately evaluates directional entry and exit at the two boundaries. The fields receive role-specific supervision; in particular, Evidence can be constrained by **reciprocal counterfactual video–query pairing** built from moment annotations instead of by clip-level saliency GT, with the same retrieval accuracy. With the same InternVideo2 features as the strongest baseline and no additional pretraining, EventFieldNet improves average mAP on QVHighlights by **1.7 points**, with larger gains at stricter IoU thresholds.

1 INTRODUCTION

Video moment retrieval localizes the intervals of an untrimmed video that match a natural-language query (Gao et al., 2017; Hendricks et al., 2017; Lei et al., 2021). Current detectors generate many candidate intervals and return them ranked by a predicted score (Lei et al., 2021; Moon et al., 2023; 2026b; Gordeev et al., 2026). Retrieval accuracy therefore depends not only on recognizing the queried event, but also on ranking the candidate that covers its full temporal extent above its overlapping neighbors.

Ranking is particularly challenging among highly overlapping candidates around the same event. Consider the query "a person puts a book on the table" in Figure 1. A truncated interval may keep the most recognizable part of the action but miss its onset or completion, whereas an overextended interval contains the event together with surrounding activity. Both can share much of their visual content with a well-aligned interval, making precise temporal extent difficult to judge from content similarity alone. Boundary errors can also be one-sided: a candidate may end at the right time but start too late, or start at the right time but end too early. For SG-DETR (Gordeev et al., 2026), the strongest baseline by average mAP in our InternVideo2 comparison without additional task pretraining, test mAP falls from 73.2 at IoU 0.5 to 55.8 at IoU 0.75: many ranked predictions find the event but not its extent. This gap between coarse retrieval and precise temporal localization motivates examining how event coverage and boundary alignment enter candidate ranking.

Prior work already separates semantic matching from localization quality. DRN combines a matching score with predicted IoU (Zeng et al., 2020), BAM-DETR ranks proposals with an IoU-supervised quality score (Lee & Byun, 2024), and Sim-DETR aligns moment queries with frames inside their target intervals (Tang et al., 2025). Closest to our work, DE^{2}TR separates semantic and boundary evidence inside its moment queries, using the former for confidence and the latter for boundary refinement (Zhang et al., 2026). An aggregate localization-quality target supervises overall overlap, while the contributions of event coverage and directional boundary alignment remain implicit in the ranking function. We make these criteria explicit through separately supervised scoring terms, alongside interval-wide semantic evidence. This formulation introduces a structured inductive bias for ranking highly overlapping candidates. Completeness assessment (Zhao et al., 2017)

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: Schematic illustration of ranking ambiguity among overlapping candidates. Truncated and overextended candidates share query-relevant content with the reference interval but differ in coverage and surrounding context; single-boundary shifts are one-sided errors. The three criteria motivate our fields: (a) Evidence, query–event agreement; (b) Support, event coverage in temporal context; (c) Transition, alignment with event entry and exit.

and contextual boundary prediction (Wang et al., 2020) have established precedents; our focus is on their integration as supervised components of candidate ranking.

We propose **EventFieldNet**, which adds three complementary scoring fields to a learned base span score. Each field is a query-conditioned score over candidate intervals, computed from the pretrained video and text features by its own interaction and readout (Transition also reads boundary cues from the input adapter). *Evidence* averages clip-level query relevance over the whole candidate, so every clip, not only the most salient one, contributes to its semantic assessment. *Support* measures whether the candidate sustains the event and compares its content with a local context region selected from the model's own provisional predictions rather than from annotations. *Transition* scores the start and the end separately from directional feature changes, so an early entry and a late exit are judged independently. The field scores are added to the base span score, and the sum ranks all candidates on a common temporal grid.

We construct counterfactual contrasts to supervise the intended role of each field. For Evidence, the counterfactual swaps the query or the video: over the annotated moments, the matched pair must score higher than both swapped pairs. For Support, it removes part of the event: the best-matching candidate must score higher than its truncations, by a margin that grows with the lost coverage, and about the same as outward expansions that lose none of it. For Transition, we contrast positive endpoints with nearby shifts and encourage the correct-versus-opposite endpoint score gap to be larger under the matched query than under an alternative query. All three contrasts are built from moment annotations alone; unlike direct saliency-GT supervision, they also apply to datasets that annotate only moments, such as Charades-STA (Gao et al., 2017) and TACoS (Regneri et al., 2013).

With the same InternVideo2 features as SG-DETR and no additional pretraining, EventFieldNet raises QVHighlights test average mAP from 54.1 to 55.84 (mean of five seeds), with a larger gain at mAP@0.75 (+1.71) than at mAP@0.5 (+0.85). With InternVideo2 features, removing any one field from a trained model lowers validation mAP in every seed, and the fields act in different regimes: Evidence and Support matter most for short moments, while the effect of Transition grows from 0.25 at IoU 0.5 to 4.49 at IoU 0.95. Constraining Evidence with the counterfactual loss instead of a saliency-GT loss gives the same validation mAP (57.46 vs. 57.45).

Our contributions are as follows.

- We formulate candidate ranking with separately supervised terms for interval-wide semantic relevance, prediction-referenced support, and directional boundary alignment, combined with a learned span score.
- We supervise each field with counterfactual contrasts tailored to its intended role—video–query pairing for Evidence, event coverage for Support, and query-conditioned boundary alignment for Transition—using moment annotations alone. For Evidence, this constraint works as well as one built from saliency GT.
- With identical features, EventFieldNet improves QVHighlights average mAP by 1.74 points over the strongest baseline without additional pretraining, and our analysis shows that the three fields contribute in complementary regimes of moment length and IoU strictness.

2 Related Work

Moment retrieval and candidate ranking. Moment-DETR formulates moment retrieval as set prediction and jointly predicts clip-level saliency for highlight detection (Lei et al., 2021). Later detectors strengthen query dependence (Moon et al., 2023; 2026b), model feedback between the two tasks (Yang et al., 2024), initialize queries from video-specific events (Jang et al., 2023), separate foreground from background (Jung et al., 2025), and exploit saliency or phrase-level structure (Gordeev et al., 2026; Kang et al., 2025). Other methods score discretized start–end maps (Zhang et al., 2020a), while multimodal language models generate timestamps directly (Huang et al., 2024a; Ren et al., 2024). Candidate scoring has also received explicit attention: FlashVTG refines confidence through intra-scale and inter-scale score fusion (Cao et al., 2025), and RGTR incorporates an IoU-aware scoring head (Sun et al., 2025a). EventFieldNet focuses on the criteria used to evaluate candidates, augmenting a learned span score with separately supervised semantic and temporal terms.

Semantic matching and localization quality. In temporal action localization, SSN models action classification and completeness separately (Zhao et al., 2017), BSN and BMN combine boundary probabilities with proposal confidence (Lin et al., 2018; 2019), and ActionFormer pairs per-moment classification with boundary regression (Zhang et al., 2022). In object detection, IoU-Net predicts localization confidence (Jiang et al., 2018), whereas GFL and VarifocalNet fold localization quality into the classification score (Li et al., 2020; Zhang et al., 2021). For language grounding, DRN combines matching scores with predicted IoU (Zeng et al., 2020), and BAM-DETR ranks proposals with an IoU-supervised quality score (Lee & Byun, 2024). Sim-DETR uses query-to-frame alignment and IoU-aware query ranking within its decoder (Tang et al., 2025), while DE^{2}TR separates semantic and boundary evidence within moment queries (Zhang et al., 2026). EventFieldNet distinguishes coverage relative to temporal context (Support) from directional entry/exit alignment (Transition), while retaining interval-wide semantic relevance (Evidence), each as an explicitly supervised ranking term.

Temporal context and boundary modeling. CBP aggregates contextual information to predict semantic boundaries and uses boundary scores to modulate anchor confidence at inference (Wang et al., 2020). TD-DETR varies background context to reduce spurious correlations (Zhou et al., 2025), while CVA combines query-aware context diversification, multi-scale context modeling, and boundary-focused contrastive learning (Moon et al., 2026a). Support instead compares each candidate with a reference region selected from the model's provisional predictions and with its surrounding context, as an explicit term of the ranking score.

Negative pairs and counterfactual supervision. MMN learns bidirectional moment–sentence matching with cross-video negatives and combines its matching score with an IoU score for prediction (Wang et al., 2022). QD-DETR uses mismatched video–query pairs to encourage query-dependent saliency (Moon et al., 2023), while counterfactual contrastive learning constructs positive and negative transformations for weakly supervised grounding (Zhang et al., 2020b). Related causal and debiasing approaches address temporal-location biases and spurious cross-modal associations (Yang et al., 2021; Zhou et al., 2025). EventFieldNet instead assigns role-specific contrasts,

built from moment annotations, to its scoring fields: query and video swaps for Evidence, coverage perturbations for Support, and boundary-position, entry/exit-role, and query perturbations for Transition.

3 METHOD

3.1 OVERVIEW

Given pretrained clip features $V \in \mathbb{R}^{N\times d_v}$ and query-token features $Q \in \mathbb{R}^{L\times d_q}$, EventFieldNet scores every candidate interval $I = (i, j)$, covering clips i through j, on a grid $\Omega \subseteq \{(i, j) : 0 \le i \le j < N\}$. Its ranking score is

$$F(i,j) = C(i,j) + E(i,j) + S(i,j) + \tfrac{1}{2}\big[T_s(i) + T_e(j)\big]. \tag{1}$$

The base span score C is produced by a conventional path (Figure 2, top): a query-guided input adapter and a multimodal temporal encoder yield contextual features H, and a span head combines endpoint logits, interval-pooled features, the query, and the interval length. The raw score is centered and scaled over the candidates of each example and bounded by $\tanh$, giving it the same range $[-1, 1]$ as each field contribution. This provides a common amplitude bound; the terms can differ in their variation across candidates and influence on the final ranking.

The three *fields* E, S, and T make the criteria of Figure 1 explicit: whether the candidate expresses the queried event, whether it covers the event relative to its temporal context, and whether its boundaries coincide with entry into and exit from the event. Each field has its own video–text interaction over the pretrained inputs rather than over H, producing embeddings $Z^E, Z^S, Z^T \in \mathbb{R}^{N\times 64}$ (Appendix A.1), and a readout whose temporal scope matches its criterion: clips inside the candidate for Evidence, relations within and around the candidate for Support, and changes at its endpoints for Transition. The fields are trained with the counterfactual objectives of Section 3.3 in addition to the ranking loss.

3.2 SCORING FIELDS

Evidence: semantic relevance over the whole interval. Evidence converts each clip embedding into a bounded relevance score and averages it over the candidate:

$$e_t = \tanh\big(w_E^\top \widetilde{Z}_t^E\big), \qquad E(i,j) = \frac{1}{j-i+1}\sum_{t=i}^{j} e_t, \tag{2}$$

where $\widetilde{Z}_t^E = \gamma_E Z_t^E$ with a fixed gain γ_E. Because the nonlinearity precedes averaging, every clip in the candidate contributes, and a candidate cannot score highly on a few salient clips while the rest is unrelated to the query.

Support: persistence inside the candidate and contrast with its context. Support has two parts. The first measures whether the event persists across the candidate through relations between adjacent clips. With RMS-normalized embeddings h_t^S and a small network g that reads $[a; b; b-a; a\odot b]$,

$$s_t = \tanh\big[g(h_t^S, h_{t+1}^S) - g(h_t^S, 0) - g(0, h_{t+1}^S) + g(0,0)\big], \qquad S_0(i,j) = \frac{1}{j-i}\sum_{t=i}^{j-1} s_t, \tag{3}$$

with $S_0(i, i) = 0$. The four-term difference cancels additive single-clip terms in g, giving s_t an inductive bias for modeling interactions between adjacent clips.

The second part explicitly compares the candidate with nearby content, complementing its internal relations. We define this temporal context from the model's own predictions. We form the provisional score $F_0 = C + E + S_0 + \frac{1}{2}(T_s + T_e)$, select up to eight reference intervals by temporal NMS on the stop-gradient of F_0, and assign each candidate I the reference with the largest IoU. Around this reference we define a local region R_I and an outer ring O_I. A learned readout of Z^S assigns each clip a positive mass m_t; with $M(X) = \sum_{t\in X} m_t$ and $D(X) = M(X)/|X|$,

$$c_I = \frac{M(I\cap R_I)}{M(R_I)}, \qquad p_I = \frac{D(I)}{D(I)+D(O_I)}, \qquad S(I) = S_0(I) + \big(1 - |S_0(I)|\big)\tanh\big(w_r^\top x_I\big), \tag{4}$$

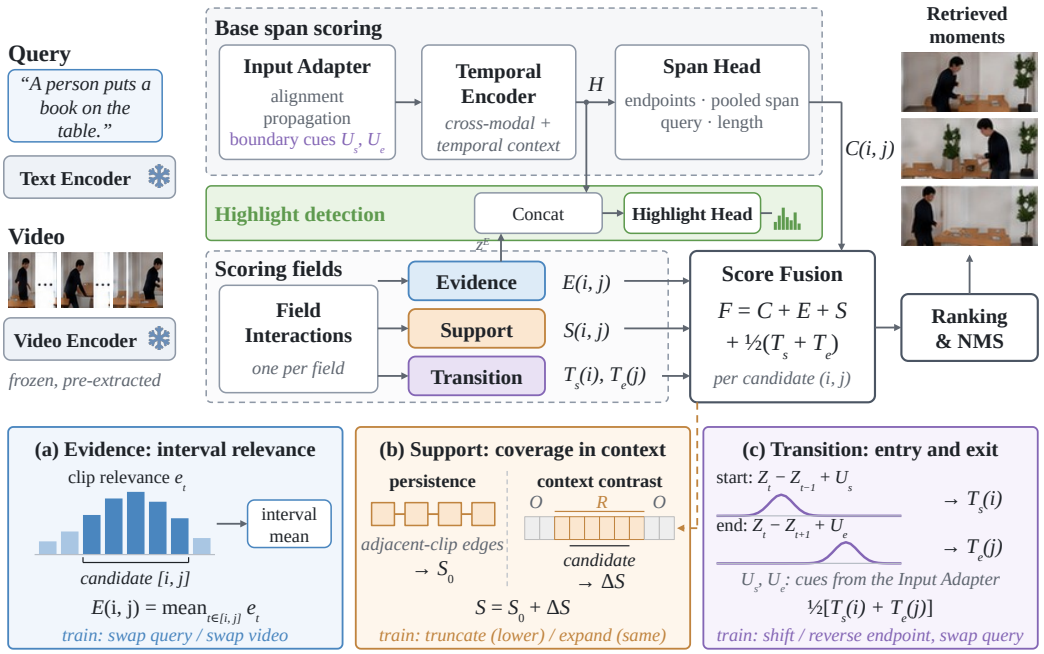

Figure 2: **EventFieldNet.** Top: the Input Adapter, Temporal Encoder, and Span Head give the base span score C; separate Field Interactions over the frozen video and text features give the Evidence, Support, and Transition fields, which Score Fusion adds to C. The Highlight Head reads H and the Evidence embedding; its output does not enter F. Bottom: field readouts. (a) Evidence averages clip relevance over the candidate. (b) Support combines adjacent-clip persistence S_0 with a context contrast ΔS computed in a region R and outer ring O around references selected from the provisional score F_0 (dashed arrow). (c) Transition scores start and end separately from directional differences and adapter cues. "train" lines list each field's counterfactual supervision, used only during training. Frames and curves are illustrative.

where the second term of $S(I)$ is the context correction $\Delta S(I)$, c_I is the share of the region's event mass that the candidate covers, p_I contrasts the candidate's mass density with that of the outer ring, and x_I collects c_I, p_I, and $c_I p_I$, each centered at its value under uniform mass so that geometric overlap alone gives no correction. The factor $1 - |S_0|$ keeps S in $[-1, 1]$. References are selected once, without annotations and without recursion; region sizes and eligibility rules are given in Appendix A.4.

Transition: directional entry and exit. An event boundary is a change in a particular direction: entering the event at the start and leaving it at the end. Transition therefore scores the two endpoints with separate directional readouts,

$$T_s(t) = \tanh\big[b_s(U_s(t)) + w_s^\top (Z_t^T - Z_{t-1}^T)\big], \qquad T_e(t) = \tanh\big[b_e(U_e(t)) + w_e^\top (Z_t^T - Z_{t+1}^T)\big], \quad (5)$$

where U_s and U_e are start and end cues from the input adapter, built from signed left and right differences of video–query agreement (Appendix A.2). A candidate (i, j) receives $\frac{1}{2}[T_s(i) + T_e(j)]$, so an early start and a late end are penalized independently. These scores are separate from the endpoint logits of the base span head.

3.3 Counterfactual Field Supervision

The ranking loss rewards the total score F and does not specify what each field should measure. We supervise each field with counterfactual contrasts designed for its ranking criterion. Interval perturbations can change content, duration, and boundary alignment together; the supervision specifies the response encouraged for each field. For a field score f, a factual input x, and a counterfactual $\tilde{x}$, a *sensitivity* constraint $[\, m - (f(x) - f(\tilde{x}))\,]_+$ asks the field to prefer the factual input by margin

m, and an *invariance* constraint asks it to score x and $\tilde{x}$ alike. All counterfactuals are built from moment annotations and are used only during training.

Evidence: query and video swaps. Evidence contrasts matched and swapped video–query pairs. We pair each example a with an example b from a different video and with a sufficiently different query, and evaluate the same Evidence function on the four combinations. Let r_{ab} be the clip scores e_t of (V_a, Q_b) pooled over the annotated moments of V_a with overlap weights, each moment weighted equally. The loss

$$\ell_E^{(a)} = \tfrac{1}{2}\big([m_E - r_{aa} + r_{ab}]_+ + [m_E - r_{aa} + r_{ba}]_+\big), \qquad m_E = 0.2, \tag{6}$$

applies a query swap (same video) and a video swap (same query); averaging over both members of each pair gives $\mathcal{L}_E$. Swapped pairs are weak negatives, since a random query may still partly describe another video. Pairing and pooling details are in Appendix A.1.1.

Support: coverage removal. For Support, we perturb intervals to remove or preserve annotated event coverage. For each annotated event G with best-matching candidate I^+, a counterfactual I^- moves one boundary of I^+. With the lost coverage $\delta c = (|I^+ \cap G| - |I^- \cap G|)/|G|$ and $\delta s = S(I^+) - S(I^-)$, truncations that remove event content receive the sensitivity loss $[0.2\,\delta c - \delta s]_+$, whose margin grows with the removed coverage, and outward expansions that keep all of G receive the invariance loss $\mathrm{SmoothL1}_{0.05}(\delta s, 0)$. Support thus learns to detect missing content without being asked to penalize every added context clip; excess context is judged by the full score F. Averaging over perturbations, events, and queries gives $\mathcal{L}_S$.

Transition: boundary shifts, reversals, and query swaps. For Transition, we construct contrasts from endpoint shifts, opposite endpoints, and alternative queries. For endpoint type $\sigma \in \{s, e\}$, let b^+ be the positive grid position supplied by the highest-IoU candidate for an annotated event. A *position* counterfactual shifts this endpoint by two clips to b^- and receives $[0.2 - T_\sigma(b^+) + T_\sigma(b^-)]_+$. A *direction* counterfactual uses the opposite endpoint $\bar{b}$ of the same event as a negative for endpoint type σ. When an eligible alternative query Q^- is available, it additionally swaps the query:

$$D_\sigma(Q) = \tfrac{1}{2}\big[T_\sigma(b^+ \mid Q) - T_\sigma(\bar{b} \mid Q)\big], \qquad \ell_{T,\mathrm{dir}} = \big[0.2 - D_\sigma(Q) + D_\sigma(Q^-)\big]_+, \tag{7}$$

which encourages a larger correct-versus-opposite endpoint score gap under the matched query than under the alternative query. Without an eligible Q^-, the direction loss uses the same-query endpoint margin $[0.2 - 2D_\sigma(Q)]_+$. Negatives near another annotated boundary of the same type are excluded; averaging over endpoint types, events, and queries gives $\mathcal{L}_T$ (Appendix A.1.3).

3.4 Training and Inference

The ranking loss combines IoU-target distribution matching, pairwise comparison between the best candidate and its contained or containing alternatives, and quality calibration, $\mathcal{L}_{\mathrm{rank}} = \mathcal{L}_{\mathrm{KL}} + \frac{1}{2}\mathcal{L}_{\mathrm{pair}} + \frac{1}{4}\mathcal{L}_{\mathrm{quality}}$ (Appendix A.3). The endpoint logits of the span head receive $\mathcal{L}_{\mathrm{endpoint}}$. For joint highlight detection, a separate head maps $[H_t; \widetilde{Z}_t^E]$ to a clip-level logit trained on the saliency GT with $\mathcal{L}_{\mathrm{HD}}$, which also updates the shared and Evidence representations; the highlight output does not enter F, and the branch is removed in retrieval-only training. The total objective is

$$\mathcal{L} = 0.5\,\mathcal{L}_{\mathrm{rank}} + 0.1\big(\mathcal{L}_E + \mathcal{L}_S + \mathcal{L}_T + \mathcal{L}_{\mathrm{endpoint}} + \mathcal{L}_{\mathrm{HD}}\big). \tag{8}$$

Support receives two training signals: its counterfactual coverage objective and the ranking loss, which improves the total score without requiring Support to keep its coverage role. To let ranking adjust Support while reducing the chance that it overrides $\mathcal{L}_S$, we attenuate only the ranking gradient on the Support parameters (its interaction projection, edge scorer, and local context readout): for these parameters θ_S, the update uses $0.1\,\lambda_{\mathrm{rank}}\nabla_{\theta_S}\mathcal{L}_{\mathrm{rank}} + \lambda_S\nabla_{\theta_S}\mathcal{L}_S$ plus the gradients of the remaining losses, with $\lambda_{\mathrm{rank}} = 0.5$ and $\lambda_S = 0.1$ from Equation 8. This is neither a smaller Support learning rate nor a change to the forward score in Equation 1; the factor is fixed at 0.1 in all experiments. At inference, we compute F_0, select references, apply the Support correction to obtain F, and run temporal NMS at IoU 0.5 over all grid candidates, returning up to 30 intervals on the input clip grid. No annotations or counterfactual inputs are used at inference.

4 EXPERIMENTS

We evaluate EventFieldNet on QVHighlights and address three questions: how it compares with prior methods that use the same features (Section 4.2), where its gains come from and how much each field contributes (Section 4.3), and whether a counterfactual constraint on Evidence can replace a saliency-GT constraint (Section 4.4).

4.1 SETUP

Benchmark and metrics. QVHighlights (Lei et al., 2021) pairs 10,310 queries with 150-second videos and annotates both moments and clip-level saliency scores. We train on the 7,218 training queries, select checkpoints on the 1,550 validation queries, and report the 1,542 test queries with the official evaluation. Moment retrieval (MR) uses Recall@1 (R1) at IoU 0.5 and 0.7, mAP at IoU 0.5 and 0.75, and mAP averaged over IoU $0.5{:}0.05{:}0.95$ (Avg). R1 scores only the top-ranked prediction, whereas mAP scores the ranked list against every annotated moment. Highlight detection (HD) uses mAP and HIT@1 at the "Very Good" level.

Features. Our main setting uses the InternVideo2-1B video and text features (Wang et al., 2024) released by SG-DETR (Gordeev et al., 2026) on a 2-second clip grid, so the comparison with SG-DETR has identical inputs. We also report the widely used CLIP+SlowFast features (Radford et al., 2021; Feichtenhofer et al., 2019; Lei et al., 2021); Appendix B.3 describes how they enter the model. No task-level pretraining is used.

Implementation. The temporal encoder has width 384 and each field has width 64. We train with AdamW (learning rate 10^{-4}, weight decay 0.05, batch size 64) for 24 epochs, with 3 warm-up epochs, learning-rate decay by 0.3 at epochs 10 and 20, and gradient clipping at 1.0. For every run, the checkpoint with the highest validation average mAP produces both the MR and HD predictions. We train every model, including the retrained ablations, with five random seeds (2041–2045) and report the mean; per-seed results are in Appendix B.6.

4.2 COMPARISON WITH THE STATE OF THE ART

InternVideo2. Table 1 compares methods within each feature setting. With the same InternVideo2 features and no pretraining, EventFieldNet improves over SG-DETR by 1.74 average mAP (54.1 to 55.84) and by 1.71 mAP@0.75, and it leads this group in average mAP, both mAP thresholds, R1@0.5, and HIT@1. The lowest of the five seeds reaches 55.05. R1@0.7 is on par with SG-DETR (56.57 vs. 56.6), and HD mAP is 0.4 below it. SG-DETR with InterVid-MR task pretraining remains higher (58.8); we do not use this pretraining.

CLIP+SlowFast. Among methods trained only on QVHighlights, EventFieldNet has the best average mAP (49.38) and mAP@0.75 (50.56), while DualGround has higher R1, mAP@0.5 (by 0.05), and HD mAP. Methods with target-data augmentation reach up to 49.87. Methods with other visual inputs are listed in Appendix Table 9, including A^2FA+F^2SGD (Tan et al., 2025) (63.4 with CLIP features), which predicts boundaries as points in a 2D frame-similarity space rather than ranking candidates.

In both settings the clearest gains are in mAP at strict IoU, while R1 stays within 1.3 points of the strongest baselines; mAP rewards every correctly ranked interval in the list, whereas R1 checks only the first.

4.3 ANALYSIS AND ABLATIONS

Gains grow with overlap strictness. Figure 3(a) shows the per-seed test difference to SG-DETR. Every seed improves mAP@0.5, mAP@0.75, and average mAP, and the gain at mAP@0.75 (+1.71) is twice that at mAP@0.5 (+0.85); from the ten-threshold gain of 1.74, the nine thresholds from 0.55 to 0.95 gain 1.84 on average. The improvement is thus concentrated at strict overlap, where the correct interval must be separated from truncated or overextended neighbors.

Table 1: Comparison on the QVHighlights test split. Rows are grouped by input features. **Bold**: best within a group among methods trained only on QVHighlights. Gray: additional training data (target-data augmentation or task pretraining), shown for reference. Baseline rows are published values. –: not reported. *Intermediate InternVideo2 layers with side tuning.

	MR R1		MR mAP			HD (≥ Very Good)	
Method	@0.5	@0.7	@0.5	@0.75	Avg	mAP	HIT@1
CLIP + SlowFast features							
QD-DETR (2023)	62.40	44.98	62.52	39.88	39.86	38.94	62.40
CG-DETR (2026b)	65.43	48.38	64.51	42.77	42.86	40.33	66.21
BAM-DETR (2024)	62.71	48.64	64.57	46.33	45.36	–	–
FlashVTG (2025)	66.08	50.00	67.99	48.70	47.59	–	–
Sim-DETR (2025)	67.64	50.91	67.81	47.59	46.93	–	–
DualGround (2025)	**68.20**	**51.72**	**69.23**	47.71	49.02	**41.15**	66.30
DE^{2}TR (2026)	67.24	51.39	67.92	49.83	48.74	41.08	**66.34**
EventFieldNet (ours)†	67.74	50.45	69.18	**50.56**	**49.38**	40.20	65.89
AMR (2025)	68.22	51.88	68.53	49.82	48.43	–	–
CVA (2026a)	70.05	55.32	69.49	48.45	47.49	44.43	66.01
DE^{2}TR + BAS (2026)	67.98	52.46	68.36	51.24	49.87	40.51	66.75
InternVideo2 features (identical release for SG-DETR and ours)							
FlashVTG (2025)	70.69	53.96	72.33	53.85	52.00	–	–
DualGround (2025)	71.87	**56.94**	72.41	54.38	52.73	**44.02**	70.80
SDST* (2025)	70.82	56.23	71.31	54.99	53.31	43.40	69.13
SG-DETR (2026)	72.2	56.6	73.2	55.8	54.1	43.8	69.1
EventFieldNet (ours)†	**72.68**	56.57	**74.05**	**57.51**	**55.84**	43.41	**71.04**
SG-DETR + PT (2026)	74.2	60.4	76.2	60.8	58.8	44.7	71.0

†Mean over five runs with random seeds 2041–2045; per-seed results and standard deviations are in Table 7.

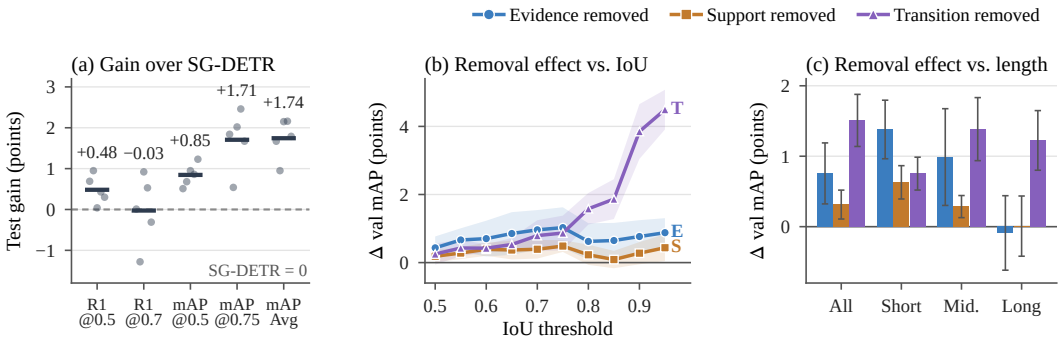

Figure 3: Where the gains come from. (a) Test difference to SG-DETR with identical InternVideo2 features; dots are seeds and horizontal lines are means. (b, c) Drop in validation mAP when one field is removed from each trained InternVideo2 model, (b) at each IoU threshold and (c) for average mAP overall and by moment length (short <10 s, middle 10–30 s, long >30 s). Bands and error bars: SD over five seeds. CLIP+SlowFast results and exact values are in Appendix B.4.

Retrained ablations. Table 2 retrains the model with components removed. Without the three fields, the base path alone reaches 52.13 ± 2.30 validation average mAP, 5.33 points below the full model and with a much larger spread across seeds. Keeping the fields but removing their counterfactual losses also costs accuracy: 1.25 points without $\mathcal{L}_E$, 2.04 without $\mathcal{L}_S$, 1.73 without $\mathcal{L}_T$, and 4.33 without all three, which leaves the model one point above the base path. Most of the gain from the fields therefore depends on their supervision rather than on the added parameters alone. Removing $\mathcal{L}_E$ also lowers HD mAP by 3.06, since the highlight head reads the Evidence embedding.

Table 2: Retrained ablations on QVHighlights validation (InternVideo2; mean ± SD over five seeds). L1–L4 remove the named field losses; L5 replaces the counterfactual $\mathcal{L}_E$ with a saliency-GT loss (Section 4.4).

		Moment retrieval		Highlight detection	
Setting		Avg mAP	R1@0.7	mAP	HIT@1
Architecture (retrained)					
EventFieldNet (full)		57.46 ±0.20	60.37 ±0.40	43.84 ±0.10	73.30 ±0.52
Base only (no fields)		52.13 ±2.30	53.54 ±3.00	42.13 ±1.30	70.07 ±2.10
Field supervision (all three fields kept, retrained)					
L1	w/o $\mathcal{L}_E$	56.21 ±0.34	59.06 ±0.57	40.78 ±1.25	67.16 ±1.63
L2	w/o $\mathcal{L}_S$	55.42 ±0.71	58.18 ±0.89	42.57 ±0.46	69.52 ±0.94
L3	w/o $\mathcal{L}_T$	55.73 ±0.39	58.51 ±0.61	42.19 ±0.31	71.58 ±0.67
L4	w/o $\mathcal{L}_E, \mathcal{L}_S, \mathcal{L}_T$	53.13 ±1.08	55.96 ±1.27	37.24 ±0.74	66.41 ±1.18
L5	$\mathcal{L}_E$ on saliency GT	57.45 ±0.21	60.70 ±0.24	43.97 ±0.14	73.70 ±0.48

Reliance on each field. To measure how much a trained model relies on each field, we subtract that field from the final score and keep every other tensor fixed, including the Support references. With InternVideo2 features, removing Transition, Evidence, or Support lowers validation average mAP by 1.51, 0.76, and 0.31, and each removal lowers it in all five seeds. Figure 3(b) splits this effect by IoU threshold, and the three fields behave differently. The Transition effect grows with the required overlap: removing it lowers mAP by 0.25 at IoU 0.5 but by 4.49 at IoU 0.95, which is 15% of the full model's mAP at that threshold. Evidence contributes at every threshold (0.43 to 1.03), and Support is largest at IoU 0.75 (0.49). By moment length (Figure 3(c)), Evidence and Support matter most for short moments (1.38 and 0.63), where a one-clip error changes IoU the most, whereas Transition matters most for middle and long moments (1.38 and 1.22). With CLIP+SlowFast features, the Transition effect rises in the same way (0.41 at IoU 0.5 to 1.57 at IoU 0.95), while Evidence helps only below IoU 0.8 (Appendix B.4). These deletions measure reliance; the retrained ablations above measure what each component adds.

4.4 Counterfactual versus Saliency Supervision of Evidence

QVHighlights annotates a saliency score for each clip inside the annotated moments, a direct constraint on Evidence: a clip with a higher saliency score should receive a higher score e_t. Charades-STA and TACoS annotate only moments, so we test whether the counterfactual loss (Equation 6) can replace a saliency-GT loss that applies the same hinge, with margin 0.2 and the same loss weight, to pairs of clips with different saliency scores. Only $\mathcal{L}_E$ changes; the architecture, seeds, checkpoint selection, and all other losses are shared (Appendix B.1). The two losses give the same validation average mAP (57.46 vs. 57.45, row L5 of Table 2) and test average mAP (55.84 vs. 55.64, Table 3), whereas removing $\mathcal{L}_E$ costs 1.25 validation points (L1). The counterfactual loss can therefore replace the saliency target of $\mathcal{L}_E$. Both models keep $\mathcal{L}_{\text{HD}}$, which also updates the Evidence embedding, so the comparison isolates the target of $\mathcal{L}_E$; on Charades-STA and TACoS, training has no $\mathcal{L}_{\text{HD}}$ and the counterfactual loss is the only auxiliary constraint on Evidence (Appendix B.5).

5 Conclusion

We presented EventFieldNet, which makes the criteria for ranking moment candidates explicit: Evidence scores interval-wide relevance, Support scores coverage in temporal context, and Transition scores directional entry and exit, each added to a learned span score and trained with its own counterfactual contrast from moment annotations. Retraining without these contrasts removes most of the gain, and for Evidence the moment-based contrast matches a saliency-GT target. With the same InternVideo2 features as SG-DETR, EventFieldNet improves QVHighlights test average mAP by 1.74 points, with the largest gains at strict IoU. More broadly, supervising each scoring term with a contrast aimed at one failure mode is a simple inductive bias for other temporal localization models.
Limitations. Predictions lie on a fixed clip grid without boundary regression, which limits short moments, and Support relies on predicted references.

AI USE STATEMENT

Generative AI tools were used to assist with writing and language polishing, including improving clarity and readability; to retrieve and discover relevant literature; and to draft parts of the manuscript. They were also used to assist with writing portions of the code, organizing the experimental code, and producing scripts for tables and figures, and to provide feedback on the experimental design, analyses, and drafts, including suggestions for ablations. The authors verified all reported results and take full responsibility for the scientific claims, experimental results, references, code, and final content of this paper.

ETHICS STATEMENT

This work studies video moment retrieval using the existing QVHighlights, Charades-STA, and TACoS benchmarks. Video datasets can contain identifiable individuals and reflect biases in the activities, language, and recording conditions they represent. Benchmark accuracy does not establish fairness or suitability across populations and applications. Video retrieval can also facilitate searching sensitive content; its use should respect privacy, consent, and the access and usage conditions of the underlying data. Our experiments evaluate retrieval performance and do not assess deployment in privacy-sensitive settings.

REPRODUCIBILITY STATEMENT

Section 3 describes the scoring fields, training objectives, and inference procedure; Appendix A provides their detailed construction and Appendix A.4 specifies the reference geometry and the exact ranking-loss weights. Section 4.1 reports the QVHighlights data splits, metrics, features, and training hyperparameters. Appendices B.1 and B.3 document checkpoint selection, the choice of the Evidence objective, and feature processing, while Appendix B.6 reports the individual seeds and their results. The field-removal protocol and the settings for the other benchmarks are described in Appendices B.4 and B.5, respectively. We will release the code, configuration files, and trained checkpoints upon publication.

REFERENCES

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 Winter Conference on Applications of Computer Vision (WACV)*, pp. 9208–9218, February 2025.

Xin Dong, Wenjia Geng, Wenfeng Deng, and Yansong Tang. CoSTL: Comprehensive spatial-temporal representation learning for moment retrieval and highlight detection. *arXiv preprint arXiv:2606.01149*, 2026.

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.

Aleksandr Gordeev, Vladimir Dokholyan, Irina Tolstykh, and Maksim Kuprashevich. Saliency-guided DETR for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 907–916, 2026.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5803–5812, 2017.

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*, 2024a.

Cheng Huang, Yi-Lun Wu, Hong-Han Shuai, and Ching-Chun Huang. Semantic fusion augmentation and semantic boundary detection: A novel approach to multi-target video moment retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6783–6792, January 2024b.

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.

Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. In *European Conference on Computer Vision*, pp. 816–832, 2018.

Minjoon Jung, Youwon Jang, Seongho Choi, Joochan Kim, Jin-Hwa Kim, and Byoung-Tak Zhang. Background-aware moment detection for video moment retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 8575–8585, 2025.

Minseok Kang, Minhyeok Lee, Minjung Kim, Donghyeong Kim, and Sangyoun Lee. Empower words: DualGround for structured phrase and sentence-level temporal grounding. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 92472–92499. Curran Associates, Inc., 2025.

Pilhyeon Lee and Hyeran Byun. BAM-DETR: Boundary-aligned moment detection transformer for temporal sentence grounding in videos. In *European Conference on Computer Vision*, pp. 220–238, 2024.

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, 2021.

Pandeng Li, Chen-Wei Xie, Hongtao Xie, Liming Zhao, Lei Zhang, Yun Zheng, Deli Zhao, and Yongdong Zhang. MomentDiff: Generative video moment retrieval from random to real. In *Advances in Neural Information Processing Systems*, volume 36, pp. 65948–65966, 2023.

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.

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*, 2023.

Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. BSN: Boundary sensitive network for temporal action proposal generation. In *European Conference on Computer Vision*, 2018.

Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. BMN: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

Ye Liu, Siyuan Li, Yang Wu, Chang-Wen Chen, Ying Shan, and Xiaohu Qie. UMT: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3042–3051, 2022.

Ye Liu, Jixuan He, Wanhua Li, Junsik Kim, Donglai Wei, Hanspeter Pfister, and Chang Wen Chen. R^2-tuning: Efficient image-to-video transfer learning for video temporal grounding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 421–438, 2024a.

Zhihang Liu, Jun Li, Hongtao Xie, Pandeng Li, Jiannan Ge, Sun-Ao Liu, and Guoqing Jin. Towards balanced alignment: Modal-enhanced semantic modeling for video moment retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 3855–3863, 2024b.

Hongxu Ma, Guanshuo Wang, Fufu Yu, Qiong Jia, and Shouhong Ding. MS-DETR: Towards effective video moment retrieval and highlight detection by joint motion-semantic learning. In *Proceedings of the 33rd ACM International Conference on Multimedia*. Association for Computing Machinery, 2025.

Sungho Moon, Seunghun Lee, Jiwan Seo, and Sunghoon Im. CVA: Context-aware video-text alignment for video temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2026a.

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.

WonJun Moon, Sangeek Hyun, SuBeen Lee, and Jae-Pil Heo. Correlation-guided calibration of query dependency for video temporal grounding. *Pattern Recognition*, 174:112984, 2026b.

Seojeong Park, Jiho Choi, Kyungjune Baek, and Hyunjung Shim. MomentMix augmentation with length-aware DETR for temporally robust moment retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1127–1136, March 2026.

David Pujol-Perich, Sergio Escalera, and Albert Clapés. Sparse-dense side-tuner for efficient video temporal grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21515–21524, October 2025.

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 International Conference on Machine Learning*, 2021.

Ran Ran, Jiwei Wei, Xiangyi Cai, Xiang Guan, Jie Zou, Yang Yang, and Heng Tao Shen. CDTR: Semantic alignment for video moment retrieval using concept decomposition transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6684–6692, 2025a.

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

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.

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*, 2024.

Hao Sun, Mingyao Zhou, Wenjing Chen, and Wei Xie. TR-DETR: Task-reciprocal transformer for joint moment retrieval and highlight detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4998–5007, 2024.

Xiaolong Sun, Liushuai Shi, Le Wang, Sanping Zhou, Kun Xia, Yabing Wang, and Gang Hua. Diversifying query: Region-guided transformer for temporal sentence grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7131–7139, 2025a.

Xiaolong Sun, Le Wang, Sanping Zhou, Liushuai Shi, Kun Xia, Mengnan Liu, Yabing Wang, and Gang Hua. Moment quantization for video temporal grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20137–20146, 2025b.

Jiawei Tan, Hongxing Wang, Junwu Weng, Jiaxin Li, Zhilong Ou, and Kang Dang. Anchor-aware similarity cohesion in target frames enables predicting temporal moment boundaries in 2D. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24180–24189, June 2025.

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.

Jingwen Wang, Lin Ma, and Wenhao Jiang. Temporally grounding language queries in videos by contextual boundary-aware prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. InternVideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, 2024.

Zhenzhi Wang, Limin Wang, Tao Wu, Tianhao Li, and Gangshan Wu. Negative sample matters: A renaissance of metric learning for temporal grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

Zhengxuan Wei, Jiajin Tang, and Sibei Yang. Augmenting moment retrieval: Zero-dependency two-stage learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3401–3412, 2025.

Yicheng Xiao, Zhuoyan Luo, Yong Liu, Yue Ma, Hengwei Bian, Yatai Ji, Yujiu Yang, and Xiu Li. Bridging the gap: A unified video comprehension framework for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Jin Yang, Ping Wei, Huan Li, and Ziyang Ren. Task-driven exploration: Decoupling and inter-task feedback for joint moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18308–18318, 2024.

Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. Deconfounded video moment retrieval with causal intervention. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021.

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

Chen-Lin Zhang, Jianxin Wu, and Yin Li. ActionFormer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, pp. 492–510, 2022.

Haoyang Zhang, Ying Wang, Feras Dayoub, and Niko Sünderhauf. VarifocalNet: An IoU-aware dense object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8514–8523, 2021.

Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2D temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 12870–12877, 2020a.

Yifan Zhang, Chengxu Liu, Yujie Dun, and Xueming Qian. DE^{2}TR: Dual evidence detection transformer for video temporal grounding. In *European Conference on Computer Vision*, pp. 463–482, 2026.

Zhu Zhang, Zhou Zhao, Zhijie Lin, Jieming Zhu, and Xiuqiang He. Counterfactual contrastive learning for weakly-supervised vision-language grounding. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 18123–18134. Curran Associates, Inc., 2020b.

Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

Xinyang Zhou, Fanyue Wei, Lixin Duan, Angela Yao, and Wen Li. The devil is in the spurious correlations: Boosting moment retrieval with dynamic learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20981–20990, October 2025.

A Additional Method Details

This appendix gives the complete field parameterizations, supervision, and network realization omitted from the concise main method.

A.1 Field Construction and Supervision

The three fields organize query-conditioned information at different temporal scopes: clips within an interval, relations within and around it, and changes at its endpoints. We describe each field together with the supervision applied to its scores. Ground-truth intervals are used only to construct training objectives, whereas candidate scoring and local reference selection depend on model inputs and predictions.

Field-specific interactions. Let $\widehat{V}$ and $\widehat{Q}$ denote row-wise normalized video and query features. For each field $k \in \{E, S, T\}$, separate projections produce video queries $X^k = \widehat{V}W_v^k$, text keys $K^k = \widehat{Q}W_k^k$, and text values $Y^k = \widehat{Q}W_u^k$. Their interaction is

$$A^k = \mathrm{softmax}_{\mathrm{text}}\left(\frac{X^k(K^k)^\top}{\sqrt{d}}\right), \qquad Z^k = \mathrm{GELU}(X^k) \odot (A^k Y^k), \tag{9}$$

where $\odot$ denotes element-wise multiplication and $d = 64$. Text padding is excluded from attention and video padding from subsequent readouts. Equation 9 shows the deterministic computation; the ordinary Evidence and Transition branches additionally apply dropout during training. The Support projections are independently parameterized. These interactions consume the pretrained input features rather than the temporal encoder output H.

A.1.1 Candidate-wide Semantic Evidence

As illustrated in Figure 2(a), Evidence evaluates the content of the entire candidate rather than selecting only its most discriminative clips. A learned linear readout converts the scaled evidence embedding $\widetilde{Z}_t^E = \gamma_E Z_t^E$ into a bounded clip score:

$$e_t = \tanh\left(w_E^\top \widetilde{Z}_t^E\right), \qquad E(i,j) = \frac{1}{j-i+1}\sum_{t=i}^{j} e_t. \tag{10}$$

Here γ_E is a fixed input gain, not a candidate-dependent attention weight. The nonlinearity precedes interval averaging. All clips in a candidate therefore contribute to its semantic assessment, although a high average alone does not establish complete event coverage.

Reciprocal counterfactual supervision. To learn query-dependent evidence, we pair examples with different video identities and sufficiently different queries. For a reciprocal pair (a, b), the same evidence function evaluates (V_a, Q_a), (V_a, Q_b), (V_b, Q_a), and (V_b, Q_b). These are training-time evaluations with shared parameters, not four inference networks.

Let $\mathcal{G}_a$ be the annotated event intervals for example a, and $J_{a,t}$ the annotation-time interval associated with clip t. For $G \in \mathcal{G}_a$, define overlap weights

$$\omega_{a,G,t} = \frac{|J_{a,t} \cap G|}{\sum_u |J_{a,u} \cap G|}, \qquad r_{ab} = \frac{1}{|\mathcal{G}_a|}\sum_{G\in\mathcal{G}_a}\sum_t \omega_{a,G,t}\, e_t(V_a, Q_b). \tag{11}$$

Only annotation-covered, non-padding clips enter the weights; each annotated interval contributes equally. The pooling weights depend on temporal overlap, not on the predicted evidence strength or saliency scores.

For each accepted example a with reciprocal partner b, we use

$$\ell_E^{(a)} = \frac{1}{2}\left([m_E - r_{aa} + r_{ab}]_+ + [m_E - r_{aa} + r_{ba}]_+\right), \qquad m_E = 0.2, \tag{12}$$

where $[x]_+ = \max(x, 0)$. The first term contrasts queries for the same video, and the second contrasts videos for the same query. Averaging over both members of accepted pairs gives $\mathcal{L}_E$. Dropout is disabled in these auxiliary Evidence evaluations and kept in the ordinary forward pass. The pairing produces weak negatives and does not guarantee semantic incompatibility.

A.1.2 CONTEXTUAL TEMPORAL SUPPORT

As illustrated in Figure 2(b), Support combines compatibility between neighboring clips with a candidate-dependent comparison against local temporal context. We first compute an internal support score S_0, then add a bounded contextual correction.

Internal temporal relations. Let h_t^S be the token-wise RMS-normalized version of Z_t^S. A shared network g reads the pair descriptor $[a; b; b-a; a \odot b]$. For adjacent clips, we subtract single-token reference responses before bounding the joint score:

$$s_t = \tanh\big[g(h_t^S, h_{t+1}^S) - g(h_t^S, 0) - g(0, h_{t+1}^S) + g(0,0)\big]. \tag{13}$$

Any additive component of the form $u(a) + v(b) + c$ cancels in this four-term difference relative to the zero-input references. The interval readout averages its internal edges:

$$S_0(i,j) = \begin{cases} \dfrac{1}{j-i}\displaystyle\sum_{t=i}^{j-1} s_t, & j > i, \\ 0, & j = i. \end{cases} \tag{14}$$

A single-clip candidate has no internal edge and receives a neutral base support value. Support averages adjacent-pair responses, while Evidence averages clip-level scores. This difference in readout structure does not imply that their learned semantic information is disjoint.

Prediction-derived contextual support. To complement internal relations, we explicitly compare the candidate with nearby content using prediction-derived reference regions. We form a provisional score

$$F_0(i,j) = C(i,j) + E(i,j) + S_0(i,j) + \frac{1}{2}[T_s(i) + T_e(j)]. \tag{15}$$

Greedy temporal NMS on $\mathrm{sg}(F_0)$ selects at most eight reference intervals, where sg denotes stop-gradient. Each candidate is assigned the reference with the largest temporal IoU. Expanding this reference defines a local region R_I, and a larger surrounding window defines an outer ring O_I. Neither reference selection nor region construction uses annotations.

A learned gate produces positive clip masses from Support embeddings:

$$m_t = \exp(\mathrm{clip}(\ell_t - \ell_{t_0}, -8, 8)), \qquad \ell_t = w_m^\top \mathrm{LN}(Z_t^S), \tag{16}$$

where t_0 is the first valid clip. For a set of valid clips X, let $M(X) = \sum_{t\in X} m_t$ and $D(X) = M(X)/|X|$. For candidate I, we compute

$$c_I = \frac{M(I \cap R_I)}{M(R_I)}, \qquad p_I = \frac{D(I)}{D(I) + D(O_I)}, \qquad u_I = \frac{|I \cap R_I|}{2|R_I|}. \tag{17}$$

Here c_I measures coverage of reference-region content and p_I compares candidate density with its outer context. The quantity u_I supplies a uniform-mass reference. The centered descriptor and final correction are

$$\begin{aligned} x_I &= [c_I - 2u_I,\ p_I - \tfrac{1}{2},\ c_I p_I - u_I], \\ \Delta S(I) &= (1 - |S_0(I)|)\tanh(w_r^\top x_I), \\ S(I) &= S_0(I) + \Delta S(I). \end{aligned} \tag{18}$$

Uniform masses make all three descriptor entries zero. Candidates without a sufficiently overlapping reference or valid outer context receive zero correction. The references are chosen from F_0 once, rather than recursively from the corrected final score.

Coverage-based supervision. For each annotated interval G, let I^+ be its highest-IoU valid grid candidate. We construct single-boundary truncations and expansions I^- and measure

$$\delta c = \frac{|I^+ \cap G| - |I^- \cap G|}{|G|}, \qquad \delta s = S(I^+) - S(I^-). \tag{19}$$

High-purity truncations with a sufficient coverage decrease receive $[0.2\,\delta c - \delta s]_+$. For eligible outward expansions that preserve coverage, we instead use $\mathrm{SmoothL1}_{0.05}(\delta s, 0)$. Candidates matching other annotated events, or expansions newly acquiring their content, are excluded. We average within perturbation families, then across available families, annotated intervals, and active queries to obtain $\mathcal{L}_S$. This objective encourages coverage sensitivity without requiring Support alone to penalize every expansion; temporal extent is evaluated through the combined score.

A.1.3 DIRECTIONAL BOUNDARY TRANSITIONS

As illustrated in Figure 2(c), Transition evaluates the two endpoints through different temporal comparisons. From Z^T, define

$$d_t^- = Z_t^T - Z_{t-1}^T, \qquad d_t^+ = Z_t^T - Z_{t+1}^T, \tag{20}$$

with unavailable-neighbor differences set to zero. The input adapter additionally provides start and end cues $U_s(t)$ and $U_e(t)$, as described in Appendix A.2. Separate cue readouts b_s, b_e and difference readouts yield

$$\begin{aligned} T_s(t) &= \tanh\left[b_s(U_s(t)) + w_s^\top d_t^-\right], \\ T_e(t) &= \tanh\left[b_e(U_e(t)) + w_e^\top d_t^+\right]. \end{aligned} \tag{21}$$

The cues and feature differences are read out separately before addition. Candidate I_{ij} receives $\frac{1}{2}[T_s(i) + T_e(j)]$; the field evaluates grid endpoints rather than regressing continuous coordinates.

Position and direction supervision. For each annotated event, its highest-IoU grid candidate provides positive start and end positions. Position negatives shift an endpoint by two clips away from its corresponding annotated boundary; direction negatives use the opposite endpoint of the same event when sufficiently separated. Negatives close to another valid boundary of the same type are excluded. For endpoint type $\sigma \in \{s, e\}$, a positive position b^+ and a position negative b^- receive

$$\ell_{T,\text{pos}} = [m_T - T_\sigma(b^+) + T_\sigma(b^-)]_+, \qquad m_T = 0.2. \tag{22}$$

For an eligible direction pair, let $\bar{b}$ denote the opposite endpoint of the same event and define $D_\sigma(V, Q) = \frac{1}{2}[T_\sigma(b^+ \mid V, Q) - T_\sigma(\bar{b} \mid V, Q)]$. When an eligible alternative query Q^- is available, the direction objective is

$$\ell_{T,\text{dir}} = [m_T - D_\sigma(V, Q) + D_\sigma(V, Q^-)]_+. \tag{23}$$

This objective encourages $D_\sigma(V, Q)$ to exceed $D_\sigma(V, Q^-)$ by m_T; it does not by itself require $D_\sigma(V, Q) > 0$. Without an eligible alternative query, the direction loss is $[m_T - T_\sigma(b^+ \mid V, Q) + T_\sigma(\bar{b} \mid V, Q)]_+ = [m_T - 2D_\sigma(V, Q)]_+$. Available position and direction families are averaged within endpoint type, followed by endpoint types, annotated events, and active queries, giving $\mathcal{L}_T$. These constraints supervise the Transition scores used in F, not a separate boundary predictor.

A.2 BASE PATH AND HIGHLIGHT HEAD

The field formulation is implemented alongside the query-conditioned base scoring path shown in Figure 2. The two paths share pretrained video and text inputs, but do not share a single field representation.

Input adaptation and temporal encoding. The input adapter first aligns each normalized video clip with query tokens and forms channel-wise video–query agreement. It constructs three cue streams: local agreement, boundary-gated propagation of agreement to adjacent clips, and signed left/right agreement differences modulated by changes in query relevance. Separate low-rank encoders and learned gates combine these streams into a video-feature update. We remove its component parallel to the original feature, bound its norm relative to that feature, and restore the original feature norm after addition. The resulting $\widetilde{V}$ retains the input temporal grid; temporal endpoint features are passed through unchanged. The directional inputs U_s, U_e are also exposed to the Transition readouts.

A multimodal temporal encoder maps $(\widetilde{V}, Q)$ to $H \in \mathbb{R}^{N \times 384}$. It combines cross-modal attention, query-conditioned feedback, role-token aggregation, multiscale temporal processing, and temporal self-attention with query-modulated residual blocks. The base scorer reads the temporal features captured after the last temporal layer. This path supplies contextual span features, while Equation 9 uses the original semantic input streams.

Base span scoring. Let $\bar{h}_{ij} = \frac{1}{j-i+1}\sum_{t=i}^{j} H_t$ and let $\bar{q}$ be the masked mean of projected query tokens. The raw base score is

$$\begin{aligned} u_{ij} &= \left[H_i; H_j; \bar{h}_{ij} + \bar{h}_{ij} \odot q_C; \frac{j-i+1}{N_v}\right], \\ c_{ij}^{\text{raw}} &= a_s(H_i) + a_e(H_j) + f_C(u_{ij}). \end{aligned} \tag{24}$$

where q_C is a learned transform of $\bar{q}$, N_v is the valid model-input length, a_s, a_e are endpoint readouts, and f_C is an MLP. For each example, centering and variance scaling over valid candidates gives

$$C(i,j) = \tanh\left(\frac{c_{ij}^{\text{raw}} - \mu_c}{\sqrt{1+\sigma_c^2}}\right). \tag{25}$$

The base endpoint logits a_s, a_e are distinct from the Transition scores in Equation 21.

Highlight output head. For joint retrieval and highlight prediction, we concatenate H_t with the scaled evidence embedding $\widetilde{Z}_t^E$ before its scalar readout:

$$h_t^{\text{HD}} = \text{MLP}_{\text{HD}}\left(\text{LN}([H_t; \widetilde{Z}_t^E])\right). \tag{26}$$

The head maps $448 \to 64 \to 1$ with GELU between the linear layers. It predicts clip-level highlight logits and contributes no term to the candidate ranking score.

A.3 Joint Learning and Inference

Candidate ranking. The joint ranking objective combines distribution matching, candidate comparison, and quality calibration. A training pool $\mathcal{P}_a$ joins high-scoring, NMS-selected, duration-stratified, and GT-neighborhood candidates. For each annotated interval G, we normalize $[\text{IoU}(I,G) - 0.7]_+^2$ over this pool and then average the distributions equally over annotated intervals to obtain $q_a(I)$. When all weights for an interval vanish, equal mass is assigned to its highest-IoU candidates in the pool. The prediction distribution is $p_a(I) = \text{softmax}_{I\in\mathcal{P}_a}(2F_a(I))$. The KL component combines this mixed-pool divergence with a duration-weighted divergence over the 30 highest-scoring candidates; Equation 30 below gives their exact combination.

A candidate-comparison hinge contrasts each sufficiently accurate highest-IoU candidate with the highest-scoring eligible contained and containing alternatives, using their IoU-quality difference as the margin. This yields $\mathcal{L}_{\text{pair}}$. Quality calibration uses binary cross-entropy between $\text{sigmoid}(2F + b_{\text{cal}})$ and maximum annotated IoU, with equal contribution from candidate-pool sources. The learned intercept b_{cal} is used only by this training objective. The resulting ranking loss is

$$\mathcal{L}_{\text{rank}} = \mathcal{L}_{\text{KL}} + \frac{1}{2}\mathcal{L}_{\text{pair}} + \frac{1}{4}\mathcal{L}_{\text{quality}}. \tag{27}$$

The quality term also uses the event-duration weight η_a defined below. Candidate selection is detached from the training graph, while the selected scores remain differentiable.

Joint objective. The base endpoint logits receive a Gaussian-target boundary loss $\mathcal{L}_{\text{endpoint}}$. For highlight prediction, the target is derived from the saliency GT: the fraction of annotators rating a clip "Very Good". The highlight objective $\mathcal{L}_{\text{HD}}$ uses balanced soft-label binary cross-entropy, combining all negative clips with the hardest negatives. Together with the field objectives, training minimizes

$$\begin{aligned}\mathcal{L} = 0.5\mathcal{L}_{\text{rank}} \\ + 0.1\big(\mathcal{L}_E + \mathcal{L}_S + \mathcal{L}_T + \mathcal{L}_{\text{endpoint}} + \mathcal{L}_{\text{HD}}\big).\end{aligned} \tag{28}$$

The HD branch and its loss are omitted in retrieval-only mode. For the Support projection, edge scorer, and local gate/readout, ranking gradients are scaled by 0.1, while auxiliary gradients are unchanged. This modifies the optimization route, not the forward score coefficients.

Inference. We first compute C, E, S_0, T_s, T_e and form F_0. Prediction-derived references then determine the local Support correction, giving $F = F_0 + \Delta S$. All valid grid candidates participate in final ranking and hard temporal NMS, which returns at most 30 moments at IoU threshold 0.5. NMS visits candidates in descending order of F until 30 moments are kept or no candidates remain. For selected indices (i, j), temporal boundaries are recovered from the input clip grid, (τ_i, τ_{j+1}). On QVHighlights the grid step is two seconds; no continuous boundary offset is regressed. Query swapping, annotation pooling, and perturbation-based supervision are training-only operations.

A.4 Reference Geometry and Ranking-Loss Weights

This section specifies the reference geometry and the effective ranking objective used in all experiments.

Reference regions. Reference NMS keeps at most eight intervals and suppresses overlaps strictly greater than 0.5. For a selected reference with half-open clip indices $[p, q)$, let $w = q - p$, $r = \max(\lfloor w/4 \rfloor, 1)$, and $r_o = \max(\lfloor w/2 \rfloor, r + 1)$. The local region and outer ring are

$$R = [p - r, q + r), \qquad W_{\mathrm{ctx}} = [p - r_o, q + r_o), \qquad O = W_{\mathrm{ctx}} \setminus R. \tag{29}$$

Regions are clipped to the input grid, and all statistics exclude invalid clips. The contextual correction requires candidate–reference IoU at least 0.3 and nonempty valid candidate, local-region, and outer-ring sets. References are selected from the detached provisional score before applying the correction in Equation 18.

Event-duration weighting. Let ρ_a be the maximum annotated event duration as a fraction of the valid video duration, and let $\eta_a = 1 + 0.15\rho_a$. The annotations used in this computation already contain normalized temporal coordinates, so no additional division by grid length is applied. An example without valid annotations has $\rho_a = 0$.

Write $K^{(2)}_{\mathrm{mix},a}$ for the KL divergence on the mixed candidate pool, with per-event target weights $[\mathrm{IoU}(I, G) - 0.7]^2_+$, and $K^{(4)}_{\mathrm{top30},a}$ for the raw top-30 divergence with exponent 4. Both use prediction probabilities $\mathrm{softmax}(2F)$ and equal averaging of normalized target distributions over valid annotations. If all target weights vanish for an event, its highest-IoU candidates receive equal target mass. The effective objective is

$$\mathcal{L}_{\mathrm{KL}} = \operatorname*{mean}_{a \in \mathcal{A}_{\mathrm{mix}}} \left[\eta_a K^{(2)}_{\mathrm{mix},a}\right] + \operatorname*{mean}_{a \in \mathcal{A}_{\mathrm{top30}}} \left[(\eta_a - 1) K^{(4)}_{\mathrm{top30},a}\right], \tag{30}$$

where the active sets contain examples with valid targets and candidates for the respective pool. These are ordinary means over active examples, not averages normalized by the sum of duration weights.

Source-balanced quality calibration. For K candidate-pool sources $\mathcal{P}_{a,k}$, define

$$w_{aI} = \frac{1}{K} \sum_{k=1}^{K} \frac{\mathbf{1}[I \in \mathcal{P}_{a,k}]}{\max(1, |\mathcal{P}_{a,k}|)}, \qquad y_{aI} = \max_{G \in \mathcal{G}_a} \mathrm{IoU}(I, G). \tag{31}$$

The quality loss is the mean, over examples with valid annotations, of $\eta_a \sum_I w_{aI}\, \mathrm{BCEWithLogits}(2F_a(I) + b_{\mathrm{cal}}, y_{aI})$. A candidate belonging to several sources receives their summed contributions. The learned intercept is used only for training. Together with the candidate-comparison loss, these terms enter Equation 27; they do not alter the inference score in Equation 1.

B Additional Experimental Details and Results

B.1 Evaluation Protocol and Model Selection

Every run trains for 24 epochs in joint MR/HD mode. The checkpoint is selected by validation average mAP, with ties broken by mAP@0.75, R1@0.7, and the earlier epoch; the selected checkpoint supplies both the MR and HD predictions. Selected epochs lie between 6 and 10 (Table 7). Test MR and HD use the official evaluation.

The Evidence variant is selected by validation average mAP; Table 3 also reports test results for all variants for completeness. The four counterfactual variants lie within 0.16 validation points, and the five variants differ by at most 0.20 test average mAP. The saliency-GT objective considers, within each query, every pair of annotated clips with different saliency scores and applies $[0.2 - (e_{\text{high}} - e_{\text{low}})]_+$ to their Evidence scores, averaging first within and then across queries; clips without a saliency score are not treated as negatives, and its loss weight is 0.1, as for the counterfactual objective. It was trained with the same seeds, architecture, schedule, and remaining objectives, but separately, before the counterfactual variants were trained. Because the counterfactual objective needs Evidence scores for swapped video–query pairs, it computes them in two additional Evidence-only forward passes without dropout, whereas the saliency-GT objective uses the regular forward pass with dropout.

Table 3: Constraints on Evidence on QVHighlights (InternVideo2). The Evidence loss $\mathcal{L}_E$ is built either from the clip-level saliency GT or from counterfactual pairs of moment annotations; all rows keep the same architecture, seeds, checkpoint selection, and other losses, including $\mathcal{L}_{\text{HD}}$. The saliency-GT row was trained separately, before the counterfactual rows (Appendix B.1). Swap: counterfactual pairs swap the query only or both the query and the video. Mean ± SD over five seeds; average mAP unless noted.

				MR Avg mAP		Test	
Evidence objective	$\mathcal{L}_E$ from	Pooling	Swap	Val	Test	mAP@0.75	HD mAP
Saliency GT	Saliency	–	–	57.45 ±0.21	55.64 ±0.34	57.35	43.61
Counterfactual	Moments	mean	query	57.30 ±0.18	55.84 ±0.43	57.50	43.42
Counterfactual	Moments	soft	query	57.34 ±0.15	55.76 ±0.34	57.50	43.48
Counterfactual	Moments	soft	both	57.45 ±0.12	55.66 ±0.34	57.38	43.45
Counterfactual (ours)	Moments	mean	both	57.46 ±0.20	55.84 ±0.49	57.51	43.41

B.2 Retrained Ablations

The base-only model in Table 2 disables the three fields from initialization and is trained with the remaining objectives. The loss ablations L1–L4 set only the named loss weights to zero from initialization and keep all three fields, their parameters, and the other objectives; the ranking-loss weight is not renormalized. L4 removes $\mathcal{L}_E$, $\mathcal{L}_S$, and $\mathcal{L}_T$ jointly and keeps the endpoint and HD losses, so it is a joint field-loss ablation rather than the removal of every auxiliary objective. Standard deviations in Table 2 are taken over seeds of each metric, whereas those in Table 4 are taken over paired full-minus-deleted differences.

B.3 Input Features

InternVideo2. We use the InternVideo2-1B features released by SG-DETR: 512-dimensional clip features with two temporal-endpoint channels on a 2-second grid (at most 75 clips before padding) and 512-dimensional token features for queries of up to 40 tokens. Following the SG-DETR loader, video sequences are zero-padded to a multiple of four positions, giving 76 positions for a full 150-second clip.

CLIP+SlowFast. The base path receives the concatenated 2,818-dimensional input (512 CLIP, 2,304 SlowFast, and two temporal-endpoint channels), while the Evidence, Support, and Transition interactions read the 512-dimensional CLIP stream. For 99 videos whose CLIP sequence is one clip

shorter than the SlowFast sequence, the last CLIP feature is repeated. For this setting we also stored the checkpoint with the best validation HD mAP. On test it reaches HD mAP 40.16 and MR average mAP 46.47, against 40.20 and 49.38 for the MR-selected checkpoint, so selecting by MR costs no HD accuracy.

B.4 Field Removal

For each checkpoint, all full-model tensors and Support references are computed once; the reduced score is $F - E$, $F - S$, or $F - \frac{1}{2}(T_s + T_e)$, followed by the same NMS and evaluation on the 1,550 validation queries. The intervention changes no parameters and recomputes no references, so it measures how much the trained score relies on each term rather than the accuracy of a model retrained without it.

Table 4 gives average mAP overall and by moment length; with CLIP+SlowFast features, the mean long-moment effects are -0.19, 0.23, and 1.09 for Evidence, Support, and Transition. Table 5 and Figure 4 break the effects down by IoU threshold for both feature sets and both metrics. With either feature set, the Transition effect is small at loose IoU and largest at IoU 0.9–0.95. Without Transition, R1 at loose thresholds changes little or even rises slightly, whereas R1 at IoU 0.8 and above falls in every seed. With InternVideo2 features, removing Evidence lowers mAP and R1 at every threshold. With CLIP+SlowFast features, its effect is positive only below IoU 0.8, consistent with its small overall effect (0.13, three of five seeds). The Support effect on mAP is largest at IoU 0.75 for both feature sets.

Table 4: Removing one field from trained models (validation, five seeds). The removed term is subtracted from the final score; every other tensor, including the Support references, is kept. Δ: full minus reduced average mAP (mean $\pm$ SD of paired differences); n_+: seeds with $\Delta > 0$. Length groups follow the official evaluator (short $<$10 s, middle 10–30 s, long $>$30 s).

	InternVideo2						CLIP + SlowFast	
	Avg mAP	Δ All	Δ Short	Δ Middle	Δ Long	n_+	Avg mAP	Δ All (n_+)
Full model	57.46	–	–	–	–	–	49.89	–
− Evidence	56.71	$0.76_{\pm 0.43}$	$1.38_{\pm 0.42}$	$0.99_{\pm 0.69}$	$-0.09_{\pm 0.53}$	5/5	49.75	$0.13_{\pm 0.38}$ (3/5)
− Support	57.15	$0.31_{\pm 0.21}$	$0.63_{\pm 0.24}$	$0.28_{\pm 0.16}$	$0.01_{\pm 0.43}$	5/5	49.47	$0.42_{\pm 0.18}$ (5/5)
− Transition	55.96	$1.51_{\pm 0.37}$	$0.75_{\pm 0.23}$	$1.38_{\pm 0.45}$	$1.22_{\pm 0.42}$	5/5	49.08	$0.81_{\pm 0.35}$ (5/5)

B.5 Charades-STA and TACoS

Charades-STA (Gao et al., 2017) and TACoS (Regneri et al., 2013) annotate moments but no clip-level saliency. Because every field objective is built from moment annotations, the same training procedure applies to them in retrieval-only mode, without the HD branch and $\mathcal{L}_{\text{HD}}$ (Table 6). On Charades-STA we use a 1-second clip grid and train for 24 epochs on 11,203 training queries, holding out 1,201 video-disjoint queries for checkpoint selection (R1@0.5, then R1@0.7, then mIoU). On TACoS we use a 2-second grid, train for 12 epochs, and evaluate on the standard test split.

B.6 Per-seed Results

Table 7 lists the selected epoch, validation average mAP, and test results of every seed for both feature sets; all five InternVideo2 seeds exceed the published SG-DETR test average mAP.

B.7 Complete QVHighlights Comparison

Table 9 lists every published QVHighlights test result we verified in the methods' own papers, grouped by input features and additional training data. Table 8 adds the InternVideo2 validation results that Gordeev et al. (2026) report for earlier methods.

Table 5: Field removal by IoU threshold (validation, five seeds). Full-model rows give mAP or R1 at each threshold; Δ rows give full minus reduced, averaged over seeds. **Bold**: the removal lowers the metric in all five seeds. Standard deviations are shown as bands in Figure 4.

	IoU threshold										
	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	Avg
mAP, InternVideo2											
Full model	74.64	71.70	69.47	66.43	63.19	59.08	53.33	47.45	39.74	29.62	57.46
Δ − Evidence	**0.43**	**0.66**	**0.70**	**0.85**	**0.96**	**1.03**	0.62	**0.65**	**0.77**	**0.88**	**0.76**
Δ − Support	0.19	0.28	**0.39**	**0.37**	**0.39**	**0.49**	0.23	0.09	0.28	**0.44**	**0.31**
Δ − Transition	0.25	**0.43**	**0.42**	**0.53**	**0.79**	**0.87**	**1.58**	**1.86**	**3.85**	**4.49**	**1.51**
mAP, CLIP + SlowFast											
Full model	69.06	65.31	63.01	59.60	55.74	51.32	45.26	38.76	30.02	20.81	49.89
Δ − Evidence	**0.45**	**0.40**	0.41	0.34	0.34	0.26	−0.06	−0.09	−0.39	−0.33	0.13
Δ − Support	**0.41**	**0.59**	**0.59**	**0.69**	**0.51**	**0.74**	**0.51**	0.21	0.00	−0.02	**0.42**
Δ − Transition	**0.41**	**0.55**	**0.58**	**0.62**	0.62	0.37	**0.85**	**1.26**	**1.22**	**1.57**	**0.81**
R1, InternVideo2											
Full model	74.14	70.03	66.18	63.42	60.37	56.94	50.75	46.68	39.51	29.79	–
Δ − Evidence	**0.46**	**0.85**	**0.80**	**1.01**	**1.15**	**1.14**	0.81	**0.83**	**1.05**	**1.11**	–
Δ − Support	0.08	0.23	0.49	0.36	0.46	0.40	0.32	0.15	0.37	0.55	–
Δ − Transition	−0.23	−0.09	−0.22	−0.03	0.30	0.22	**1.24**	**1.59**	**3.59**	**4.19**	–
R1, CLIP + SlowFast											
Full model	67.82	62.92	58.93	56.05	52.41	48.76	42.03	37.73	29.88	21.15	–
Δ − Evidence	0.44	0.50	0.43	0.44	0.41	0.34	−0.10	−0.17	−0.44	−0.32	–
Δ − Support	0.27	0.75	**0.79**	**0.86**	**0.66**	**0.90**	**0.54**	0.17	0.14	0.03	–
Δ − Transition	0.28	0.49	**0.53**	**0.45**	0.46	0.21	**1.15**	**1.05**	**1.15**	**1.52**	–

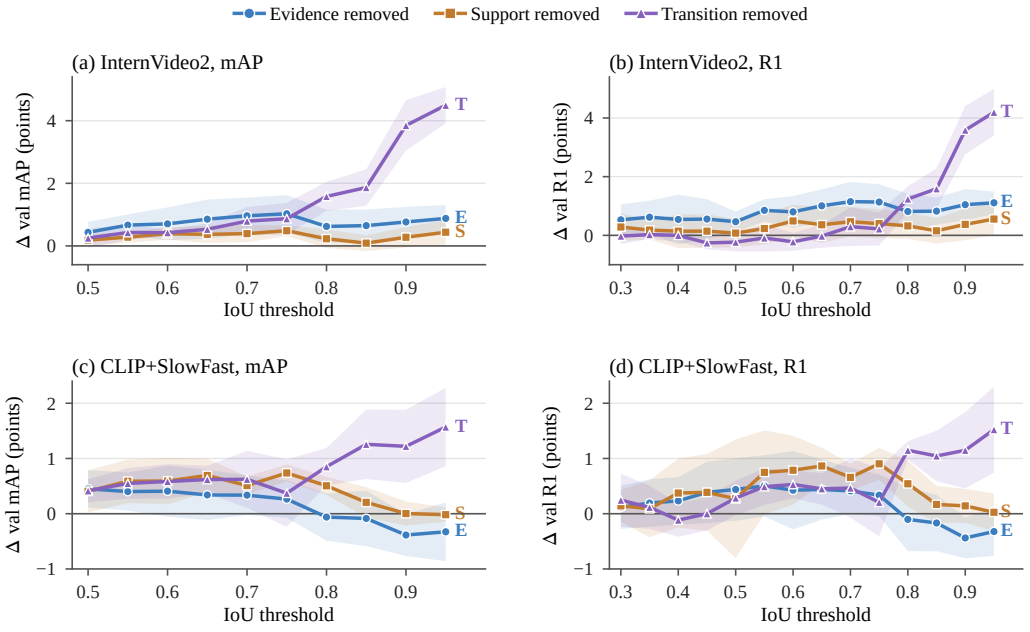

Figure 4: Field removal by IoU threshold for both feature sets, measured by mAP (left) and R1 (right) on validation. Lines: mean over five seeds; bands: ±1 SD. Values below zero mean that removing the field raises the metric, as for Transition and R1 at loose thresholds.

Table 6: Moment retrieval on Charades-STA and TACoS (R1 at IoU thresholds). Published rows use the standard splits and are shown at their published precision. **Bold**: best in the InternVideo2 group, excluding task pretraining (gray). Our Charades-STA model is trained on 11,203 training queries, with 1,201 video-disjoint training queries held out for checkpoint selection, and evaluated on the standard 3,720-query test split.

	Charades-STA test		TACoS test		
Method	R1@0.5	R1@0.7	R1@0.3	R1@0.5	R1@0.7
CLIP + SlowFast features					
BAM-DETR (2024)	59.95	39.38	56.69	41.54	26.77
FlashVTG (2025)	60.11	38.01	53.71	41.76	24.74
AMR (2025)	62.02	40.67	55.16	40.91	26.07
CVA (2026a)	62.61	40.78	58.8	43.21	27.73
DE^{2}TR (2026)	60.39	39.87	56.29	42.52	27.14
InternVideo2 features					
FlashVTG (6B) (2025)	70.32	49.87	–	–	–
SDST* (2025)	**72**	**52.6**	–	44.5	**32.3**
SG-DETR (2026)	70.2	49.5	–	44.7	29.9
SG-DETR + PT (2026)	71.1	52.8	–	46.4	33.9
EventFieldNet (ours)	70.59	48.80	–	**45.81**	30.63

Table 7: Per-seed QVHighlights results. Each checkpoint is selected by validation average mAP and supplies both MR and HD predictions. Val: validation average mAP of the selected checkpoint; the remaining columns are test.

Seed	Epoch	Val	R1@0.5	R1@0.7	mAP@0.5	mAP@0.75	Avg	HD mAP	HIT@1
InternVideo2									
2041	6	57.45	72.89	56.61	73.71	57.64	55.77	43.44	71.21
2042	6	57.19	73.15	55.32	73.88	56.34	55.05	43.39	71.66
2043	8	57.37	72.24	57.52	74.15	57.82	56.25	43.57	71.27
2044	8	57.64	72.63	57.13	74.06	58.26	56.26	43.30	70.69
2045	7	57.67	72.50	56.29	74.43	57.47	55.89	43.34	70.36
Mean		57.46	72.68	56.57	74.05	57.51	55.84	43.41	71.04
SD		0.20	0.35	0.85	0.27	0.72	0.49	0.10	0.51
CLIP + SlowFast									
2041	8	50.42	67.83	50.32	69.18	50.33	49.06	39.97	64.66
2042	9	49.30	67.19	50.52	68.77	50.66	49.43	40.66	67.38
2043	9	49.51	67.57	49.81	68.46	49.83	48.91	40.13	65.56
2044	8	49.98	67.57	50.32	69.37	50.84	49.34	40.06	66.15
2045	10	50.23	68.55	51.30	70.12	51.13	50.14	40.17	65.69
Mean		49.89	67.74	50.45	69.18	50.56	49.38	40.20	65.89
SD		0.47	0.51	0.54	0.63	0.50	0.48	0.27	0.99

Table 8: QVHighlights validation with InternVideo2 features. Baseline rows are the reproductions reported by Gordeev et al. (2026); ours is the five-seed mean.

Method	mAP@0.5	mAP@0.75	Avg
Moment-DETR	60.20	34.43	35.40
UniVTG	63.51	38.83	37.78
QD-DETR	67.78	46.40	45.52
CG-DETR	69.86	49.35	48.69
TR-DETR	70.08	49.20	47.99
SG-DETR	73.52	57.91	55.64
EventFieldNet (ours)	**74.64**	**59.08**	**57.46**

Table 9: Complete QVHighlights test comparison. Published values are copied from each method's own paper; the last column lists deviations from the standard input and training setting. Our rows give five-seed means (SDs in Table 7). **Bold**: best within the two groups whose methods use only QVHighlights training data with the stated features.

	MR R1		MR mAP			HD		
Method	@0.5	@0.7	@0.5	@0.75	Avg	mAP	HIT@1	Note
CLIP + SlowFast, QVHighlights training data only								
Moment-DETR (2021)	52.89	33.02	54.82	29.40	30.73	35.69	55.60	
QD-DETR (2023)	62.40	44.98	62.52	39.88	39.86	38.94	62.40	
UniVTG (2023)	58.86	40.86	57.60	35.59	35.47	38.20	60.96	
MomentDiff (2023)	57.42	39.66	54.02	35.73	35.95	–	–	
CG-DETR (2026b)	65.43	48.38	64.51	42.77	42.86	40.33	66.21	
TR-DETR (2024)	64.66	48.96	63.98	43.73	42.62	39.91	63.42	
UVCOM (2024)	63.55	47.47	63.37	42.67	43.18	39.74	64.20	
BAM-DETR (2024)	62.71	48.64	64.57	46.33	45.36	–	–	
MESM (2024b)	62.78	45.20	62.64	41.45	40.68	–	–	
SFABD (2024b)	–	–	62.38	44.39	43.79	–	–	BERT text
BM-DETR (2025)	60.12	43.05	63.08	40.18	40.08	–	–	
CDTR (2025a)	65.79	49.60	66.44	45.96	44.37	–	–	
RGTR (2025a)	65.50	49.22	67.12	45.77	45.53	–	–	
FlashVTG (2025)	66.08	50.00	67.99	48.70	47.59	–	–	
Sim-DETR (2025)	67.64	50.91	67.81	47.59	46.93	–	–	
DualGround (2025)	**68.20**	**51.72**	**69.23**	47.71	49.02	**41.15**	66.30	
DE^{2}TR (2026)	67.24	51.39	67.92	49.83	48.74	41.08	**66.34**	
EventFieldNet (ours)	67.74	50.45	69.18	**50.56**	**49.38**	40.20	65.89	
CLIP + SlowFast with audio, augmentation, or generated training data								
UMT (2022)	56.23	41.18	53.83	37.01	36.12	38.18	59.99	audio
AMR (2025)	68.22	51.88	68.53	49.82	48.43	–	–	aug.
TD-DETR (2025)	64.53	50.37	66.21	47.32	46.69	–	–	aug.
KDA (2025b)	66.70	50.88	67.57	46.31	45.67	–	–	gen. moments
MS-DETR (2025)	64.72	48.77	66.41	44.91	44.89	40.45	65.95	gen. text
MomentMix+LAD$_{\mathrm{QD}}$ (2026)	63.62	48.77	65.50	47.78	46.61	40.35	64.46	aug.
MomentMix+LAD$_{\mathrm{UVCOM}}$ (2026)	65.37	50.71	66.65	49.22	48.21	40.91	66.54	aug.
CVA (2026a)	70.05	55.32	69.49	48.45	47.49	44.43	66.01	aug.
DE^{2}TR + BAS (2026)	67.98	52.46	68.36	51.24	49.87	40.51	66.75	aug.
InternVideo2 features								
FlashVTG (2025)	70.69	53.96	72.33	53.85	52.00	–	–	
DualGround (2025)	71.87	**56.94**	72.41	54.38	52.73	**44.02**	70.80	
SDST* (2025)	70.82	56.23	71.31	54.99	53.31	43.40	69.13	IV2 layers, tuned
SG-DETR (2026)	72.2	56.6	73.2	55.8	54.1	43.8	69.1	
EventFieldNet (ours)	**72.68**	56.57	**74.05**	**57.51**	**55.84**	43.41	**71.04**	
InternVideo2 features with task pretraining								
SG-DETR + PT (2026)	74.2	60.4	76.2	60.8	58.8	44.7	71.0	InterVid-MR
Other visual inputs								
R^2-Tuning (2024a)	68.03	49.35	69.04	47.56	46.17	40.75	64.20	CLIP, tuned
MQVTG (2025b)	66.28	50.00	67.98	48.55	47.08	39.49	64.07	CLIP patches
CoSTL (2026)	75.43	54.24	71.48	48.45	46.68	41.36	71.59	BLIP-2
A^{2}FA+F^{2}SGD (2025)	80.8	64.3	–	64.3	63.4	–	–	CLIP