REFOCUS-DETR: REVISITING SALIENCY AND MEMORY FOR MOMENT RETRIEVAL AND HIGHLIGHT DETECTION

Anonymous authors
Paper under double-blind review

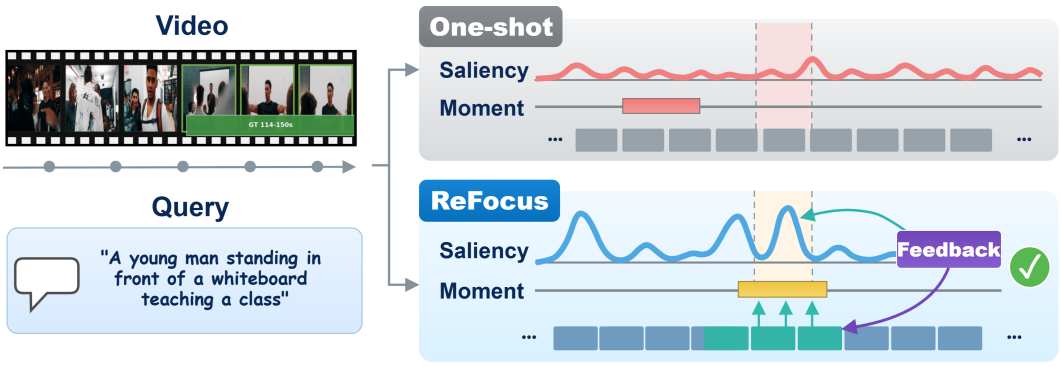

Figure 1: **ReFocus-DETR** turns the usual one-way saliency-to-localization pipeline into a closed loop: query-aware saliency proposes candidate evidence, span semantics refine clip memory, and progressively localized moments revise both saliency and memory.

ABSTRACT

Moment Retrieval (MR) and Highlight Detection (HD) provide complementary views of query-conditioned video relevance, yet most DETR-based systems connect them only once: an early saliency map reweights encoder memory, after which a decoder localizes moments from a fixed representation. Errors in this map therefore persist, while increasingly precise spans cannot repair the features that produced them. We introduce ReFocus-DETR, a closed-loop detector with two complementary feedback paths. Bidirectional Saliency–Span Interaction (BSSI) converts saliency peaks into parameter-free temporal proposals and writes their span semantics back to clip memory through position-Gaussian attention. Iterative Memory Refinement (IMR) then uses the shared decoder's current spans to revise saliency and update memory along position, channel, and intra-moment coherence axes. Near-identity gates and stop-gradient span conditioning keep the recurrent updates stable. Across QVHighlights, TACoS, Charades-STA, and TVSum, ReFocus-DETR consistently improves both localization and saliency metrics. The gains persist across feature regimes and exceed parameter-matched and training-only controls, while adding 2.6% parameters to the detector.

1 INTRODUCTION

The ability to locate semantically relevant intervals in untrimmed video supports video search and editing Lei et al. (2021), surveillance Gordeev et al. (2026), sports analysis Song et al. (2015), and personalized recommendation Liu et al. (2022). Moment Retrieval (MR) predicts the center and duration of query-relevant intervals Gao et al. (2017); Highlight Detection (HD) scores the relevance of every clip Sun et al. (2014). Their outputs are naturally complementary: salient clips indicate where a moment may occur, whereas a localized moment provides higher-level evidence about which clips should be salient. QVHighlights and Moment-DETR first formalized this joint setting Lei et al. (2021) using DETR-style set prediction Carion et al. (2020).

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.

Subsequent systems improve query encoding, task coupling, or feature quality, including QD-DETR Moon et al. (2023b), CG-DETR Moon et al. (2023a), TR-DETR Sun et al. (2024), UVCOM Xiao et al. (2024), Keyword-DETR Um et al. (2025), and MS-DETR Ma et al. (2025). A parallel line uses multimodal large language models (MLLMs) to generate timestamps: Mr. BLIP adapts an image–text model to MR Meinardus et al. (2024), while UniTime introduces explicit timestamp tokens Li et al. (2025c). Reinforcement-finetuned models such as VideoChat-R1 Li et al. (2025a), TempSamp-R1 Li et al. (2025b), and TAR-TVG Guo et al. (2025) further optimize temporal reasoning. These approaches broaden semantic reasoning, and TempSamp-R1 also predicts highlight scores. Their autoregressive inference and large-model training have a different cost profile, however, and they do not use localized predictions to recurrently revise the clip memory consumed by a detector.

Despite their different components, DETR-based joint models retain a common one-shot dependency: local query–clip saliency reweights encoder memory before decoding, and later localization evidence cannot revise that upstream state. This creates a specific error-amplification path. An underestimated relevant clip is weakened before the decoder reads it; the decoder may subsequently identify the correct neighborhood Jang et al. (2023), but its improved span cannot repair either the saliency map or the memory used by the next prediction. The problem is therefore not insufficient network capacity alone, but the absence of a state update conditioned on the model's current localization. A training-only auxiliary saliency loss can improve the initial estimate, but it cannot use the current test video's evolving spans to repair its state during inference. Likewise, simply deepening the decoder refines queries against the same memory; it does not correct the upstream evidence that caused the error.

We introduce **ReFocus-DETR**, which makes this dependency bidirectional. **BSSI** forms the saliency-to-span path: it extracts parameter-free proposals from saliency peaks, pools span semantics, and selectively writes them back to clip memory with position-Gaussian attention. **IMR** forms the span-to-saliency/memory path: across weight-shared decoder passes, predicted moments recalibrate saliency and update memory along temporal, channel, and intra-moment coherence axes. Stop-gradient span conditioning and near-identity gates prevent unstable recurrent credit assignment. This organization is deliberate: BSSI supplies structured span evidence before decoding, while IMR uses progressively improved decoder evidence to correct the next state.

Our contributions are:

- We formulate joint MR/HD as closed-loop state refinement and isolate the failure mode of one-shot saliency amplification through quality-stratified and parameter-matched controls.
- We propose IMR, a weight-shared decoder that uses current spans to co-refine saliency and encoder memory through prediction-aware temporal, channel, and coherence updates.
- We propose BSSI, a complementary saliency-to-span pathway with parameter-free proposals, position-Gaussian clip–span attention, gated memory writing, and cycle consistency. The complete model improves both tasks across four benchmarks and two feature regimes.

2 Related Work

2.1 Moment Retrieval

Early methods decompose moment retrieval into proposal generation and ranking via sliding windows Hendricks et al. (2017); Ge et al. (2019) or anchor networks Zhang et al. (2020b; 2019), while proposal-free approaches directly regress temporal boundaries through cross-modal interaction Yuan et al. (2019); Mun et al. (2020); Zhang et al. (2020a); Rodriguez et al. (2020). The DETR paradigm Carion et al. (2020) unified these threads: Moment-DETR Lei et al. (2021) adapts set-prediction matching to 1-D video spans, spawning a line of successors that exploit richer query representations Moon et al. (2023b;a), event-aware attention Jang et al. (2023), boundary-aligned predictions Lee & Byun (2024), unified grounding pre-training Lin et al. (2023), and disentangled motion–semantic encoding Ma et al. (2025). Sim-DETR Tang et al. (2025) diagnoses query conflicts in DETR-based temporal sentence grounding, showing that self-attention adjustments and global-local bridging unlock substantial gains. Iterative refinement has also been explored in object detection through cascaded heads Cai & Vasconcelos (2018), multi-scale deformable attention Zhu et al. (2021),

and denoising queries Li et al. (2022); Zhang et al. (2023); in temporal grounding, MomentDiff Li et al. (2023) iteratively denoises moment predictions via diffusion. These methods refine *predictions or queries*; they do not revisit the encoder representation in light of localization outcomes. Our SRM and MRM modules operate on the *encoder memory* rather than on decoder queries, and plug into any DETR-style backbone.

Generative MLLMs provide another direction. Mr. BLIP Meinardus et al. (2024) adapts an image–text MLLM to output temporal moments, VTG-LLM and UniTime Guo et al. (2024); Li et al. (2025c) encode timestamps explicitly, and VideoChat-R1 Li et al. (2025a) applies reinforcement fine-tuning to spatio-temporal perception. TempSamp-R1 Li et al. (2025b) addresses sparse on-policy exploration with off-policy temporal solutions, whereas TAR-TVG Guo et al. (2025) constrains intermediate reasoning with auditable timestamp anchors. These models improve temporal reasoning, and TempSamp-R1 also evaluates highlight detection. Their generative prediction path differs from ours: ReFocus-DETR uses localized predictions to revise clip-level saliency and memory during inference.

2.2 Highlight Detection and Joint MR/HD Frameworks

Early highlight detection assigns salience without textual guidance via unsupervised Mahasseni et al. (2017) or supervised Gygli et al. (2016); Xu et al. (2021) objectives, with TVSum Song et al. (2015) and category-specific methods Sun et al. (2014); Xiong et al. (2019) driving early progress. Text-guided highlight detection emerged alongside MR Lei et al. (2021): UMT Liu et al. (2022) integrates audio cues, QD-DETR Moon et al. (2023b) sharpens saliency via negative-pair contrastive learning, and SG-DETR Gordeev et al. (2026) introduces saliency-guided cross-attention with a local-to-global amplifier, demonstrating that pre-aligned features from InternVideo2 Wang et al. (2024) yield large gains across benchmarks. Keyword-DETR Um et al. (2025) further improves saliency through temporally-weighted keyword clustering. CoSTL Dong et al. (2026) complements this direction by learning fine-grained spatial and multi-scale temporal representations from raw frames. On the joint front, TR-DETR Sun et al. (2024) first explicitly couples MR and HD by using saliency to guide query generation and boundaries to refine scores within a single forward pass; UVCOM Xiao et al. (2024) and TaskWeave Yang et al. (2024) introduce progressive inter-modality interaction and decoupled task heads; MS-DETR Ma et al. (2025) addresses data sparsity via motion-semantics contrastive denoising. Despite architectural diversity, their principal saliency-to-decoder path remains one-shot: saliency is computed before localization and is not recurrently revised from the decoder's current spans. TR-DETR's one-shot coupling is the closest in spirit to our approach, but it does not iterate and does not refine encoder memory. Our method generalises this to T refinement steps and, crucially, corrects saliency errors that originate upstream of the decoder entirely—something no prior joint framework is designed to do. Feature quality is equally critical: CLIP Radford et al. (2021) and SlowFast Feichtenhofer et al. (2019) remain standard baselines, while InternVideo Wang et al. (2022) and InternVideo2 Wang et al. (2024) advance video-language alignment at scale. ReFocus-DETR is feature-agnostic and yields consistent gains under both feature regimes.

3 Proposed Method

3.1 Preliminaries

Given L uniformly sampled video clips and a query of N tokens, MR predicts normalized center–width spans $\{(\hat{c}_i, \hat{w}_i)\}$ and HD predicts saliency logits $\mathbf{s} \in \mathbb{R}^L$. Frozen InternVideo2-1B encoders produce projected video and text features $\mathbf{V} \in \mathbb{R}^{L\times d}$ and $\mathbf{Q} \in \mathbb{R}^{N\times d}$, where $d = 256$. Following SG-DETR Gordeev et al. (2026), a cross-modal encoder and lightweight global–local bridge (GLB) Tang et al. (2025) produce clip memory $\mathbf{M}_0 \in \mathbb{R}^{L\times d}$ and initial saliency $\mathbf{s}_0$. We index clips by $\ell = 0, \ldots, L-1$ and set $u_\ell = (\ell + 0.5)/L$. The standard amplifier is $\mathcal{A}(\mathbf{M}, \mathbf{s}) = \mathbf{M} \odot (1 + \sigma(\mathbf{s}))$, with broadcasting over channels.

ReFocus-DETR adds two complementary state transitions. BSSI first maps $(\mathbf{M}_0, \mathbf{s}_0)$ to saliency-aware clip memory $(\mathbf{M}_B, \mathbf{s}_B)$ (§3.3); without BSSI these reduce to $(\mathbf{M}_0, \mathbf{s}_0)$. IMR then initializes $\mathbf{s}^{(0)} = \mathbf{s}_B$ and $\mathbf{M}^{(0)} = \mathcal{A}(\mathbf{M}_B, \mathbf{s}_B)$, then invokes one weight-shared decoder (§3.2). This ordering makes the two paths explicit: BSSI writes saliency-derived span evidence into clip memory, whereas IMR writes decoder-derived localization evidence back into saliency and memory.

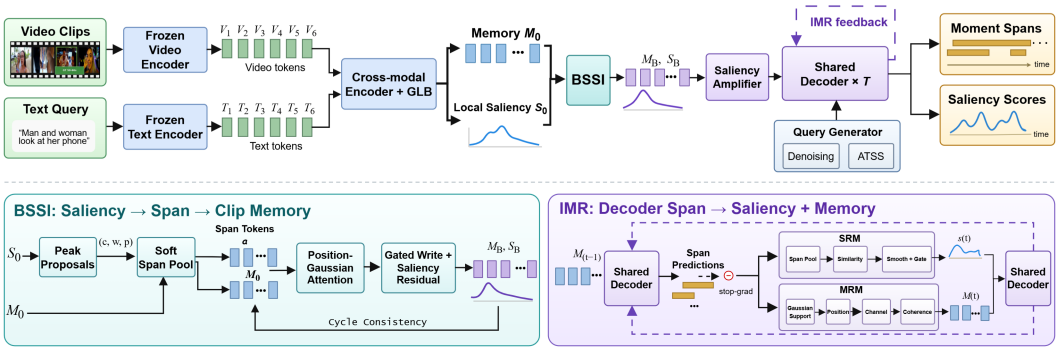

Figure 2: **Architecture of our proposed ReFocus-DETR**. The upper path maps frozen video–text features to query-conditioned memory, applies BSSI once, and performs T weight-shared decoder passes. The lower panels expand BSSI (saliency peaks → span pooling → gated clip-memory writing) and IMR (decoded spans → saliency and memory refinement → the next pass). Dashed arrows summarize feedback between shared passes. SRM saliency re-amplifies $\mathbf{M}_B$ before MRM constructs the next memory; the final HD output is SRM saliency. The cycle arrow denotes a training loss, not a recurrent BSSI update.

3.2 Iterative Memory Refinement Decoder

IMR uses decoded moments to update the evidence available to the next pass. We count T shared-decoder evaluations. For the full model ($T > 1$),

$$\begin{aligned} \hat{\mathcal{Y}}^{(t)} &= \mathrm{Dec}(\mathbf{M}^{(t-1)}), \qquad 1 \le t \le T, \\ \mathbf{s}^{(t)} &= \mathcal{R}_S\big(\mathbf{V}, \mathrm{sg}(\hat{\mathcal{Y}}^{(t)}); \mathrm{sg}(\mathbf{s}^{(t-1)})\big), \\ \mathbf{M}^{(t)} &= \mathcal{R}_M\big(\mathcal{A}(\mathbf{M}_B, \mathbf{s}^{(t)}), \mathrm{sg}(\hat{\mathcal{Y}}^{(t)})\big), \qquad t < T. \end{aligned} \tag{1}$$

Here $\hat{\mathcal{Y}}^{(t)}$ contains spans and foreground confidences, Dec shares parameters across passes, and sg stops gradients. Equation 1 alternates prediction with state refinement; the last pass returns $(\hat{\mathcal{Y}}^{(T)}, \mathbf{s}^{(T)})$ without another memory update. The $T = 1$ control bypasses both refiners and returns $(\mathrm{Dec}(\mathbf{M}^{(0)}), \mathbf{s}_B)$. Every decoder pass is supervised (Eq. 9). MRM rebuilds memory from $\mathbf{M}_B$, rather than accumulating residuals on previous refined states. Both refiners use the same top-K_{pred} predictions; Appendix E evaluates this choice.

Saliency refinement (SRM). Let the top $K_{\text{pred}} = 5$ detached predictions be $(\hat{c}_i, \hat{w}_i, \hat{p}_i)$. A learned query token attention-pools the video features inside each interval into $\mathbf{f}_i^{(t)}$. Unlike uniform pooling, the token can emphasize the clips that best summarize the predicted event. SRM then constructs one confidence-weighted moment prototype and compares it with every clip:

$$\begin{aligned} \boldsymbol{\pi}^{(t)} &= \mathrm{softmax}(\hat{\mathbf{p}}^{(t)}/\tau_s), & \bar{\mathbf{f}}^{(t)} &= \sum_i \pi_i^{(t)} \mathbf{f}_i^{(t)}, \\ q_\ell^{(t)} &= \cos_\epsilon(\mathbf{V}_\ell, \bar{\mathbf{f}}^{(t)}), & \tilde{\mathbf{q}}^{(t)} &= \mathrm{AvgPool}_3(\mathbf{q}^{(t)}), \\ \mathbf{s}^{(t)} &= (1 - \alpha_s)\,\mathrm{sg}(\mathbf{s}^{(t-1)}) + \alpha_s \tilde{\mathbf{q}}^{(t)}, & \alpha_s &= \sigma(\gamma_s). \end{aligned} \tag{2}$$

Here $\cos_\epsilon(\mathbf{a}, \mathbf{b}) = \langle \mathbf{a}, \mathbf{b} \rangle / (\|\mathbf{a}\|_2 \|\mathbf{b}\|_2 + \epsilon)$, with $\epsilon > 0$ for numerical stability. We learn $\tau_s > 0$ and use padded three-clip averaging ($w_s = 1$). Initializing $\gamma_s = -4$ gives $\alpha_s \approx 0.018$, so the correction begins small. Equation 2 converts sparse moment evidence into a dense HD signal: confidence weighting favors reliable spans, cosine similarity reduces feature-scale sensitivity, and averaging smooths isolated fluctuations. Softmax normalizes confidence contributions; it does not explicitly remove duplicate spans.

Memory refinement (MRM). MRM uses the same predictions to localize updates. With normalized clip coordinate u_ℓ, its confidence-weighted support is

$$h_\ell^{(t)} = \max_{i \le K_{\text{pred}}} \hat{p}_i^{(t)} \exp\left[-\frac{(u_\ell - \hat{c}_i^{(t)})^2}{2(\sigma_g \hat{w}_i^{(t)})^2 + \epsilon} \right]. \tag{3}$$

Equation 3 concentrates refinement around confident predictions; $\hat{p}_i \in [0,1]$ and $\sigma_g > 0$ control strength and bandwidth. Let $\mathbf{X} = \mathcal{A}(\mathbf{M}_B, \mathbf{s}^{(t)})$ and abbreviate $\mathbf{h} = \mathbf{h}^{(t)}$. Define weighted pooling as $P_{\mathbf{h}}(\mathbf{Z}) = \sum_\ell h_\ell \mathbf{Z}_\ell / (\sum_\ell h_\ell + \epsilon)$. MRM applies

$$\begin{aligned} \mathbf{X}_{\text{pos}} &= \mathbf{X} + \boldsymbol{\alpha}_{\text{pos}} \odot \mathbf{h} \odot [\mathcal{C}(\mathbf{X}) - \mathbf{X}], \\ \mathbf{X}_{\text{ch}} &= \mathbf{X}_{\text{pos}} + \alpha_h \mathbf{h} \odot \phi_{\text{ch}}(P_{\mathbf{h}}(\mathbf{X}_{\text{pos}})), \\ \mathbf{M}^{(t)} &= \text{LN}(\mathbf{X}_{\text{ch}} + \alpha_c \mathbf{h} \odot [P_{\mathbf{h}}(\mathbf{X}_{\text{ch}}) - \mathbf{X}_{\text{ch}}]). \end{aligned} \tag{4}$$

The position gate is $\boldsymbol{\alpha}_{\text{pos}} = \sigma(\mathbf{X}\mathbf{W}_p + \mathbf{b}_p)$; $\mathbf{W}_p \in \mathbb{R}^{d \times 1}$ maps each clip to a scalar gate. $\mathcal{C}$ is a Conv1D block and ϕ_{ch} is a two-layer $d/4$-bottleneck MLP. Scalar gates satisfy $\alpha_h = \sigma(\gamma_h)$ and $\alpha_c = \sigma(\gamma_c)$. We initialize $\mathbf{b}_p, \gamma_h, \gamma_c$ to -4 and the MLP output layer to zero, making residual corrections initially small before normalization. Vectors are broadcast over clip or channel dimensions as appropriate. The three lines of Eq. 4 respectively correct temporal neighborhoods, add content-dependent channel offsets, and pull supported features toward a shared centroid. This centroid aggregates the selected spans; it is not a separate centroid for each moment. The gated pull promotes coherence while retaining local variation. Gaussian tails provide soft attenuation outside predicted spans, rather than an exact foreground mask.

3.3 Bidirectional Saliency-Span Interaction

BSSI turns the initial saliency map into structured span evidence before IMR. Its inputs are clip memory $\mathbf{M}_0 \in \mathbb{R}^{L \times d}$ and local saliency logits $\mathbf{s}_0 \in \mathbb{R}^L$—never decoder-query features. Let $\mathbf{p} = \sigma(\mathbf{s}_0)$. We select up to $K_{\text{peak}} = 5$ strongest local maxima above a fixed threshold. A peak at ℓ_q gives confidence $\hat{p}_q = p_{\ell_q}$, center $\hat{c}_q = (\ell_q + 0.5)/L$, and width

$$\hat{w}_q = \text{clamp}\Big(\tfrac{1}{2} \, \text{AvgPool}_{2w_p+1}(\mathbf{p})_{\ell_q}, 0.05, 0.5 \Big). \tag{5}$$

We use half-window $w_p = 1$ (three clips); Appendix E reports its sensitivity. Equation 5 maps local saliency magnitude to a bounded proposal width; it does not directly estimate action boundaries. Proposal extraction itself has no learned parameters.

For proposal q, differentiable interval weights $\omega_{\ell q}$ softly select and normalize weights over clips within $[\hat{c}_q - \hat{w}_q/2, \hat{c}_q + \hat{w}_q/2]$, yielding $\mathbf{a}_q = \sum_\ell \omega_{\ell q} \mathbf{M}_{0,\ell}$. Each clip attends to these span tokens. Coordinate projection $\phi(\hat{c}_q, \hat{w}_q)$ augments the keys, while a confidence-scaled Gaussian bias favors temporally compatible clip–span pairs:

$$\begin{aligned} B_{\ell q} &= \hat{p}_q e^{\tau_b} \exp\left[-\frac{(u_\ell - \hat{c}_q)^2}{2(\hat{w}_q/2)^2 + \epsilon} \right], \\ \mathbf{z}_\ell &= \text{MHA}(\mathbf{M}_{0,\ell}, \{\mathbf{a}_q + \phi(\hat{c}_q, \hat{w}_q)\}_q, \{\mathbf{a}_q\}_q; \mathbf{B}_{\ell:}). \end{aligned} \tag{6}$$

In Eq. 6, τ_b is a learned log-scale and ϕ projects span coordinates into the key space. $B_{\ell q}$ is an additive attention-logit boost, not a hard mask. A per-dimension gate then controls how much span evidence is written to each clip, and a zero-initialized MLP produces the corresponding saliency residual:

$$\begin{aligned} \mathbf{r}_\ell &= \sigma(\mathbf{W}_r[\mathbf{z}_\ell; \mathbf{M}_{0,\ell}] + \mathbf{b}_r), \qquad \mathbf{M}_{B,\ell} = \text{LN}(\mathbf{M}_{0,\ell} + \mathbf{r}_\ell \odot \mathbf{z}_\ell), \\ s_{B,\ell} &= s_{0,\ell} + \text{MLP}(\mathbf{z}_\ell). \end{aligned} \tag{7}$$

We initialize $\mathbf{b}_r = -4$ and the MLP output layer to zero, giving a small gated memory residual before layer normalization and an exactly zero initial saliency residual. Equation 7 supplies the state used to initialize IMR; layer normalization itself is not an identity map.

Finally, we re-pool $\tilde{\mathbf{a}}_q$ from $\mathbf{M}_B$ over the same soft interval and preserve its span semantics with

$$\mathcal{L}_{\text{cycle}} = \frac{1}{K_B} \sum_{q=1}^{K_B} \|\tilde{\mathbf{a}}_q - \text{sg}(\mathbf{a}_q)\|_2^2. \tag{8}$$

Table 1: Performance comparison on QVHighlights val split.

Method	Features Extractor	MR					HD	
		mAP			R1		≥ Very Good	
		@0.5	@0.75	Avg	@0.5	@0.7	mAP	HIT@1
Moment-DETR Lei et al. (2021)	CLIP+SlowFast	–	–	32.20	53.94	34.84	35.65	55.55
QD-DETR Moon et al. (2023b)	CLIP+SlowFast	62.23	41.82	41.22	62.68	46.66	39.13	63.03
CG-DETR Moon et al. (2023a)	CLIP+SlowFast	64.07	42.81	42.33	64.13	48.06	39.91	64.84
TR-DETR Sun et al. (2024)	CLIP+SlowFast	66.27	46.42	45.09	–	–	40.55	64.77
Sim-DETR Tang et al. (2025)	CLIP+SlowFast	69.70	51.11	49.50	69.48	54.06	–	–
Keyword-DETR Um et al. (2025)	CLIP+SlowFast	–	–	47.69	68.97	53.35	41.67	67.03
MS-DETR Ma et al. (2025)	CLIP+SlowFast	67.19	46.05	46.00	66.90	51.68	40.57	66.58
TempSamp-R1 Li et al. (2025b)	Qwen2.5-VL-7B	–	–	–	–	–	28.30	54.90
SG-DETR Gordeev et al. (2026)	InternVideo2-1b	73.52	57.91	55.64	72.80	59.50	43.91	71.47
Mr.BLIP Meinardus et al. (2024)	BLIP-2	69.39	55.78	–	**76.13**	**63.35**	–	–
ReFocus-DETR	InternVideo2-1b	**74.87**	**59.21**	**56.94**	75.94	63.07	**44.62**	**73.25**

Here K_B is the number of retained proposals. With no valid proposal, BSSI returns its input state and the cycle loss is zero. Equation 8 penalizes changes in pooled span semantics while permitting clipwise updates; IMR consumes the resulting clip memory, not the proposal tokens directly.

3.4 Training Objective

ReFocus-DETR is trained with deep supervision applied at every decoder pass. For pass t, $\mathcal{L}_{\text{det}}^{(t)}$ contains Hungarian-matched span regression and foreground classification, together with quality/centerness, saliency contrastive, and ATSS auxiliary losses. The matching applies to the DETR predictions; the auxiliary head retains its own assignment rule. Only losses with available task annotations are applied. The objective is

$$\mathcal{L} = \mathcal{L}_{\text{det}}^{(T)} + \lambda_{\text{iter}} \sum_{t=1}^{T-1} \mathcal{L}_{\text{det}}^{(t)} + \lambda_{\text{cyc}} \mathcal{L}_{\text{cycle}}, \qquad \lambda_{\text{iter}} = 0.5, \quad \lambda_{\text{cyc}} = 0.1. \tag{9}$$

Prediction tensors supplied to SRM and MRM are detached, so gradients train each pass and all shared parameters but do not backpropagate through the discrete top-K selection or through spans from an earlier pass.

4 Experiments

4.1 Datasets and Implementation Details

Datasets and metrics. We evaluate joint MR/HD on QVHighlights Lei et al. (2021), MR on Charades-STA Gao et al. (2017) and TACoS Regneri et al. (2013), and HD on TVSum Song et al. (2015). MR metrics include top-1 recall at the reported IoU thresholds (R1, abbreviated R in Tables 3–4), mean IoU, and QVHighlights mAP averaged over thresholds $0.50{:}0.05{:}0.95$. QVHighlights HD uses mAP and HIT@1 with the "Very Good" relevance threshold; TVSum reports category-wise top-5 mAP and its mean. Each model is trained on the corresponding dataset; the comparison is not zero-shot transfer.

Implementation Details. Visual and textual features are pre-extracted using InternVideo2-1B encoder across all four benchmarks and kept frozen throughout training. The model builds on SG-DETR and is optimized with AdamW (lr $= 1 \times 10^{-4}$, weight decay $= 1 \times 10^{-4}$) for 80 epochs with cosine annealing and batch size 16. We use $\lambda_{L1} = 10$, $\lambda_{\text{gIoU}} = 1$, $\lambda_{\text{CE}} = 5$, $\lambda_{\text{cent}} = 1$, and unit weights for the DETR, ATSS, and saliency losses. Unless stated otherwise, $T = 3$, $K_{\text{pred}} = K_{\text{peak}} = 5$, and $w_s = w_p = 1$. All hyperparameters are fixed across datasets. Training uses one NVIDIA RTX 4090 GPU; Appendix C reports detector profiling; Appendix D reports multi-seed variability.

Table 2: Highlight detection results on TVsum. † denotes the methods that utilize the audio modality.

Method	VT	VU	GA	MS	PK	PR	FM	BK	BT	DS	Avg
QD-DETR Moon et al. (2023b)	88.2	87.4	85.6	85.0	85.8	86.9	76.4	91.3	89.2	73.7	85.0
CG-DETR Moon et al. (2023a)	86.9	88.8	**94.8**	87.7	86.7	**89.6**	74.8	93.3	89.2	75.9	86.8
TR-DETR Sun et al. (2024)	89.3	93.0	94.3	85.1	88.0	88.6	80.4	91.3	89.5	**81.6**	88.1
SG-DETR Gordeev et al. (2026)	89.9	90.5	88.0	87.4	90.0	84.5	77.9	93.2	89.9	80.3	87.1
MS-DETR Ma et al. (2025)	89.8	**93.3**	94.7	88.7	88.5	89.0	**80.5**	93.9	88.7	76.8	88.4
UMT† Liu et al. (2022)	87.5	81.5	88.2	78.8	81.4	87.0	76.0	86.9	84.4	79.6	83.1
QD-DETR† Moon et al. (2023b)	87.6	91.7	90.2	88.3	84.1	88.3	78.7	91.2	87.8	77.7	86.6
ReFocus-DETR	**90.6**	92.3	91.1	**89.5**	**90.8**	87.1	79.4	**94.2**	**90.6**	81.5	**88.7**

Table 3: Results on MR datasets TACoS test and val splits.

Split	Method	R@0.3	R@0.5	R@0.7	mIoU
val	Sim-DETR Tang et al. (2025)	57.06	42.79	26.82	39.44
	CG-DETR Moon et al. (2023a)	52.23	39.61	22.23	36.48
	MS-DETR Ma et al. (2025)	53.16	39.65	23.42	37.01
	ReFocus-DETR	**59.09**	**43.13**	**27.46**	**41.05**
test	SG-DETR Gordeev et al. (2026)	–	44.70	29.91	40.90
	MS-DETR Ma et al. (2025)	56.51	43.00	25.37	39.23
	ReFocus-DETR	**61.19**	**46.10**	**30.22**	**42.82**

Table 4: Charades-STA test. †: 7B generative MLLM with RL.

Method	R@0.3	R@0.5	R@0.7	mIoU
TAR-TVG† Guo et al. (2025)	**83.60**	71.40	50.20	61.10
TempSamp-R1† Li et al. (2025b)	**83.60**	**74.10**	**52.90**	**62.10**
QD-DETR Moon et al. (2023b)	–	57.31	32.55	–
UniVTG Lin et al. (2023)	70.81	58.01	35.65	50.10
CG-DETR Moon et al. (2023a)	70.43	58.44	36.34	50.13
SG-DETR Gordeev et al. (2026)	–	70.20	49.50	59.10
Mr.BLIP Meinardus et al. (2024)	–	69.31	49.29	58.63
MS-DETR Ma et al. (2025)	71.34	59.62	36.48	50.59
ReFocus-DETR	75.08	71.79	50.81	59.91

4.2 Experimental Results

Joint MR/HD. On QVHighlights validation (Table 1), ReFocus-DETR improves feature-matched SG-DETR by 1.35 mAP@0.5, 1.30 average mAP, 3.14 R1@0.5, and 3.57 R1@0.7. HD improves simultaneously by 0.71 mAP and 1.78 HIT@1, indicating that the feedback loop does not trade one task against the other. The gains also transfer to the held-out test split: ReFocus-DETR obtains 74.7/58.1/56.0 for mAP@0.5/mAP@0.75/average mAP and 44.4/71.7 for HD-mAP/HIT@1, improving SG-DETR on all seven metrics (Appendix A). Over five seeds, average MR is 56.88 ± 0.14 versus 53.00 ± 0.18 for the +FE ablation baseline (Appendix D).

Evaluation across datasets. The same configuration is used without dataset-specific module changes. On TACoS test (Table 3), ReFocus-DETR reaches 46.10 R@0.5 and 42.82 mIoU, gains of 1.40 and 1.92 over SG-DETR. On Charades-STA (Table 4), it improves the same compact baseline by 1.59 R@0.5 and 1.31 R@0.7. Recent 7B RL-based MLLMs obtain higher absolute scores on this single-task benchmark, but require autoregressive decoding and multi-GPU fine-tuning; ReFocus-DETR occupies a complementary 39.7M-parameter detector regime, excluding its frozen feature encoders. On TVSum (Table 2), the average rises to 88.7, with gains distributed across categories rather than dominated by one video type.

What causes the gain? Appendix B separates feedback from capacity and supervision. A parameter-matched self-attention/MLP refiner reaches 53.92 average MR, whereas decoder-conditioned IMR+BSSI reaches 56.94; an additional saliency pseudo-loss without inference-time feedback reaches 54.34. Porting the same modules to Moment-DETR with CLIP+SlowFast improves average MR from 32.20 to 36.23 and HD-mAP from 35.65 to 38.06. Thus the effect survives a weaker feature regime and cannot be explained by the added parameter budget alone. The full detector adds 0.99M parameters (2.6%) and increases detector latency from 21.6 to 30.4 ms at batch size one; we report this non-negligible trade-off and full profiling in Appendix C.

4.3 Ablation Studies

Overall component contribution. Table 5 quantifies the cumulative effect of the proposed design on QVHighlights. Starting from the baseline, feature extraction (+FE) model InternVideo2 provides a strong first gain, improving moment retrieval from 67.41/49.65/48.87 to 71.23/54.42/53.03 in

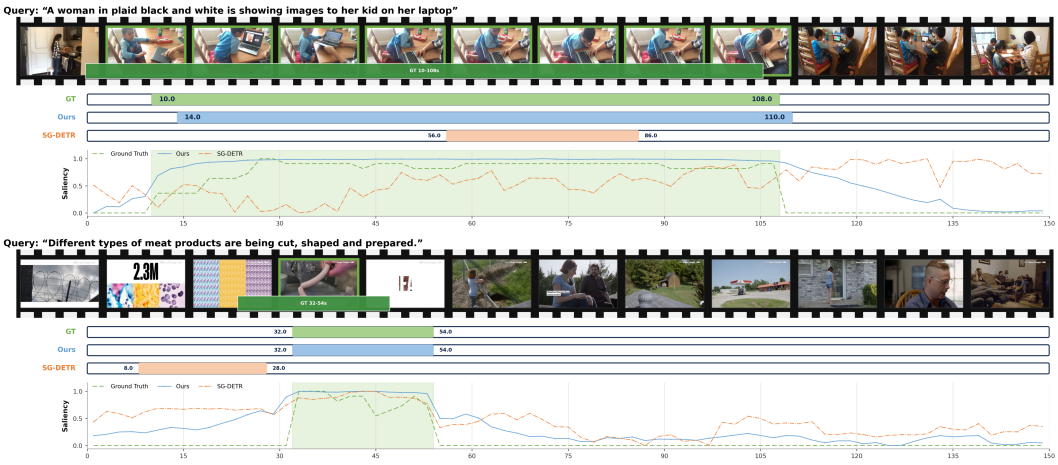

Figure 3: **Qualitative QVHighlights comparison**. ReFocus-DETR retrieves a near-complete interval in the upper example and matches the annotated interval in the lower example. Curves show final saliency predictions, not intermediate decoder passes. Additional examples appear in Appendix H.

mAP@0.5/mAP@0.75/Avg, while also increasing highlight detection from 40.02 to 42.54 HD-mAP and from 67.62 to 70.37 HIT@1. Adding IMR further strengthens both tasks, reaching 72.91/57.03/55.99 for MR and 43.21/71.33 for HD, which indicates that iterative feedback is not merely a training stabilizer but a meaningful refinement mechanism.

The full model with BSSI achieves the best results on every metric, 74.87/59.21/56.94 for MR and 44.62/73.25 for HD, confirming that saliency-peak proposal generation and bidirectional span interaction contribute complementary gains beyond IMR alone. The cumulative gains show that BSSI remains useful after IMR; the controlled ablations below further separate their effects.

Table 5: Effects of different components on QVHighlights val split.

Method	mAP			R1		HD	
	@0.5	@0.75	Avg	@0.5	@0.7	mAP	HIT@1
Baseline	67.41	49.65	48.87	67.48	53.76	40.02	67.62
+FE	71.23	54.42	53.03	71.55	58.78	42.54	70.37
+IMR	72.91	57.03	55.99	74.21	60.42	43.21	71.33
+BSSI	74.87	59.21	56.94	75.94	63.07	44.62	73.25

Effect of iterative depth. Here $T = 1$ is the one-pass detector with no feedback update. Adding a second pass improves every metric, and $T = 3$ gives the best overall trade-off: 72.91 mAP@0.5, 57.03 mAP@0.75, 55.99 average MR, 43.21 HD-mAP, and 71.33 HIT@1. A fourth decoder call adds cost but slightly reduces MR, indicating saturation rather than gains from unrestricted depth.

Table 6: Effect of the number of shared-decoder passes T.

T	mAP@0.5	mAP@0.75	Avg	HD-mAP	HIT@1
1	71.23	54.42	53.03	42.54	70.37
2	72.18	55.61	54.21	42.98	70.91
3	**72.91**	**57.03**	**55.99**	43.21	**71.33**
4	72.52	57.00	55.85	**43.25**	71.30

SRM and MRM contributions. Table 7 isolates the two IMR pathways with the +FE features fixed and BSSI disabled. SRM alone increases average MR from 53.03 to 54.96 and HD-mAP from 42.54 to 42.87. MRM alone reaches 54.88 average MR and 42.96 HD-mAP, with higher mAP@0.75 than SRM (55.74 versus 55.18). This pattern is consistent with their different interfaces: SRM reconstructs dense relevance from predicted moments, whereas MRM changes the features read by the next decoder pass. Combining them yields 55.99 average MR, 57.03 mAP@0.75, and 43.21 HD-mAP, exceeding either isolated pathway on every reported metric. In particular, the 1.29-point mAP@0.75 gain over MRM alone shows that memory refinement still benefits from revised saliency. The gains are not additive, as both modules use the same predictions; their combination supports coordinated state refinement rather than two independent sources of supervision applied to fixed initial features.

Table 7: Contribution of SRM and MRM.

SRM	MRM	mAP			HD	
		@0.5	@0.75	Avg	mAP	HIT@1
×	×	71.23	54.42	53.03	42.54	70.37
✓	×	72.02	55.18	54.96	42.87	70.92
×	✓	72.31	55.74	54.88	42.96	71.05
✓	✓	72.91	57.03	55.99	43.21	71.33

Table 8: Ablation on MRM refinement stages.

Spatial	Channel	Coherence	mAP		HD	
			@0.5	@0.75	mAP	HIT@1
×	×	×	71.23	54.42	42.54	70.37
✓	×	×	71.88	55.02	42.79	70.71
✓	✓	×	72.14	55.23	42.81	70.91
✓	✓	✓	72.31	55.74	42.96	71.05

Table 9: Detector cost at batch size one.

Model	Params (M)	FLOPs (G)	Latency (ms)
SG-DETR	38.72	16.3	21.6
+IMR	39.31	19.4	27.9
+IMR+BSSI	39.71	20.1	30.4

Table 10: Five-seed QVHighlights validation.

Model	Avg. MR	HD-mAP	HIT@1
+FE	53.00±0.18	42.51±0.13	70.31±0.25
+IMR	55.91±0.16	43.17±0.11	71.27±0.20
+IMR+BSSI	**56.88±0.14**	**44.57±0.10**	**73.17±0.20**

Ablation of MRM refinement stages. Table 8 adds the MRM stages sequentially with SRM and BSSI disabled. Position refinement alone improves mAP@0.5 by 0.65 points and mAP@0.75 by 0.60, while increasing HD-mAP and HIT@1 by 0.25 and 0.34. This provides evidence that local temporal correction is useful before any channel or centroid update. Adding channel modulation raises mAP@0.5 from 71.88 to 72.14 and mAP@0.75 from 55.02 to 55.23, indicating that a span-conditioned content summary contributes beyond temporal support. The coherence stage supplies a further 0.51 points at mAP@0.75, compared with 0.17 at mAP@0.5, and improves both HD metrics. Its larger strict-overlap gain is consistent with more precise localization, although this table does not isolate boundary error directly. The final row exactly matches MRM-only in Table 7. These are conditional, sequential gains: they support the proposed ordering, but do not establish the standalone importance of channel or coherence refinement. The results also support retaining the small channel increment, because coherence is evaluated on the already modulated representation consumed downstream.

4.4 Efficiency and Statistical Reliability

Table 9 makes the accuracy–cost trade-off explicit. Because decoder parameters are shared, IMR+BSSI increases the detector from 38.72M to 39.71M parameters, but three decoder calls raise latency from 21.6 to 30.4 ms. The $T = 4$ ablation adds still more computation without improving MR, so the gain is not monotonic in compute. Across five seeds (Table 10), the full model improves average MR by 3.88 points and HD-mAP by 2.06 points over +FE, while the corresponding standard deviations are at most 0.18 and 0.13. Appendix C provides the profiling protocol and peak memory; Appendix D reports all five metrics.

5 Conclusion

ReFocus-DETR formulates joint moment retrieval and highlight detection as reciprocal evidence exchange between clip relevance and temporal localization. BSSI converts early saliency into span context before decoding; IMR uses localized predictions to revise saliency and the memory read by the next decoder pass. This coupling addresses an upstream error that refining decoder queries alone cannot resolve: a relevant clip weakened by the initial saliency estimate can regain influence after the model finds a plausible moment. Controlled feature, capacity, supervision, and module ablations support the value of this information flow across both tasks. The gains carry measurable latency, and saliency-derived proposals become less reliable when many weakly separated peaks compete. Future work will explore vision–language models to provide semantic evidence for ambiguous events and reinforcement learning to decide when further refinement is warranted. Adapting the number of passes to each video and query could preserve the benefits of feedback while reducing its computational cost.

AI USE STATEMENT

In this work, we used generative AI tools for coding tasks, specifically Claude Code and Codex coding agents. We have not used generative AI tools for other tasks requiring disclosure, such as research ideation, data generation, experimental design, or generation of scientific claims; the remaining required-disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for language polishing and editing of the manuscript, which is a recommended-disclosure task. We have reviewed all AI-assisted work. In particular, AI-generated code was inspected, adapted, and tested for correctness; AI-polished text was checked for accuracy, originality, and consistency with our intended meaning.

ETHICS STATEMENT

This work does not involve human subjects, personally identifiable information, or sensitive data. The experiments are conducted on publicly available benchmark datasets and do not involve applications that we believe raise ethical concerns. We do not foresee any negative societal impacts specific to this work. To the best of our knowledge, this manuscript does not cover any topics or experiments that require ethical review under the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, this manuscript provides detailed implementation details, model configurations, hyperparameter settings, evaluation metrics, and experimental protocols. We also specify the computational resources and random seeds used where applicable. The code will be publicly released upon acceptance.

REFERENCES

Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: Delving into high quality object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6154–6162, 2018.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pp. 213–229. Springer, 2020.

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.

Runzhou Ge, Jiyang Gao, Kan Chen, and Ram Nevatia. MAC: Mining activity concepts for language-based temporal localization. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pp. 245–253, 2019.

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

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

Yongxin Guo, Jingyu Liu, Mingda Li, Dingxin Cheng, Xiaoying Tang, Dianbo Sui, Qingbin Liu, Xi Chen, and Kevin Zhao. VTG-LLM: Integrating timestamp knowledge into video LLMs for enhanced video temporal grounding. *arXiv preprint arXiv:2405.13382*, 2024.

Michael Gygli, Yale Song, and Liangliang Cao. Video2GIF: Automatic generation of animated GIFs from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1001–1009, 2016.

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.

Jinhwa 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.

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. Springer, 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, pp. 11846–11858, 2021.

Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M. Ni, and Lei Zhang. DN-DETR: Accelerate DETR training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13619–13627, 2022.

Pan 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, 2023.

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

Yunheng Li, Jing Cheng, Shaoyong Jia, Hangyi Kuang, Shaohui Jiao, Qibin Hou, and Ming-Ming Cheng. TempSamp-R1: Effective temporal sampling with reinforcement fine-tuning for video LLMs. In *Advances in Neural Information Processing Systems*, 2025b.

Zeqian Li, Shangzhe Di, Zhonghua Zhai, Weilin Huang, Yanfeng Wang, and Weidi Xie. Universal video temporal grounding with generative multi-modal large language models. *arXiv preprint arXiv:2506.18883*, 2025c.

Kevin Qinghong Lin, Pengchuan Zhang, Jin 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.

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.

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*, 2025.

Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. Unsupervised video summarization with adversarial LSTM networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 202–211, 2017.

Boris Meinardus, Anil Batra, Anna Rohrbach, and Marcus Rohrbach. The surprising effectiveness of multimodal large language models for video moment retrieval. *arXiv preprint arXiv:2406.18113*, 2024.

WonJun Moon, Sangeek Hyun, Sangmin Lee, and Jae-Pil Heo. Correlation-guided query-dependency calibration in video representation learning for temporal grounding. *arXiv preprint arXiv:2311.08835*, 2023a.

WonJun Moon, Sangeek Hyun, SuBeen 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, 2023b.

Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10810–10819, 2020.

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*, 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.

Cristian Rodriguez, Edison Marrese-Taylor, Fatemeh Sadat Saleh, Hongdong Li, and Stephen Gould. Proposal-free temporal moment localization of a natural-language query in video using guided attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2464–2473, 2020.

Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. TVSum: Summarizing web videos using titles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5179–5187, 2015.

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.

Min Sun, Ali Farhadi, and Steve Seitz. Ranking domain-specific highlights by analyzing edited videos. In *European Conference on Computer Vision*, pp. 787–802. Springer, 2014.

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

Sung Jin Um, Dongjin Kim, Sangmin Lee, and Jung Uk Kim. Watch video, catch keyword: Context-aware keyword attention for moment retrieval and highlight detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. InternVideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. InternVideo2: Scaling video foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*, 2024.

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*, pp. 18709–18719, 2024.

Bo Xiong, Yannis Kalantidis, Deepti Ghadiyaram, and Kristen Grauman. Less is more: Learning highlight detection from video duration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1258–1267, 2019.

Minghao Xu, Hang Wang, Bingbing Ni, Riheng Zhu, Zhenbang Sun, and Changhu Wang. Cross-category video highlight detection via set-based learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7970–7979, 2021.

Jing 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.

Yitian Yuan, Tao Mei, and Wenwu Zhu. To find where you talk: Temporal sentence localization in video with attention-based location regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 9159–9166, 2019.

Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S. Davis. MAN: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1247–1257, 2019.

Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. *arXiv preprint arXiv:2004.13931*, 2020a.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *International Conference on Learning Representations*, 2023.

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, 2020b.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021.

A HELD-OUT QVHIGHLIGHTS EVALUATION

Table 11 evaluates whether the validation improvements persist on the held-out QVHighlights test split. Relative to feature-matched SG-DETR, ReFocus-DETR improves R1@0.5/R1@0.7 by 1.9/4.1 points and mAP@0.5/mAP@0.75/average mAP by 1.5/2.3/1.9 points. The larger gains at stricter overlap thresholds are consistent with more precise localization, rather than improvements confined to coarse event detection. HD-mAP and HIT@1 also rise by 0.7 and 2.6 points, so the retrieval gains accompany improved clip ranking. ReFocus-DETR leads six of the seven reported columns; CoSTL has the highest R1@0.5. Mr. BLIP and CoSTL use different feature and training regimes, so these rows provide broader context rather than controlled module comparisons. The SG-DETR comparison holds the feature family fixed; Appendix B provides the closer ablation controls.

Table 11: QVHighlights test split. Best values are bold. A dash indicates that the source does not report the metric.

Method	R1@0.5	R1@0.7	mAP@0.5	mAP@0.75	Avg. mAP	HD-mAP	HIT@1
Moment-DETR Lei et al. (2021)	52.9	33.0	54.8	29.4	30.7	35.7	55.6
QD-DETR Moon et al. (2023b)	62.4	45.0	62.6	39.9	39.9	38.6	62.4
CG-DETR Moon et al. (2023a)	65.4	48.4	64.5	42.8	42.9	40.3	66.2
TR-DETR Sun et al. (2024)	64.7	49.0	64.0	43.7	42.6	39.9	63.4
Mr. BLIP Meinardus et al. (2024)	74.7	60.5	68.1	53.4	51.4	–	–
CoSTL Dong et al. (2026)	**75.4**	54.2	71.5	48.5	46.7	41.4	71.6
SG-DETR Gordeev et al. (2026)	72.2	56.6	73.2	55.8	54.1	43.7	69.1
ReFocus-DETR	74.1	**60.7**	**74.7**	**58.1**	**56.0**	**44.4**	**71.7**

B CONTROLLED ALTERNATIVES AND FEATURE PORTABILITY

Matched-capacity and training-only controls. The blind refiner substitutes a parameter-matched self-attention/MLP block for decoder-conditioned refinement. The pseudo-loss variant adds saliency correction supervision while keeping inference one-shot. Table 12 distinguishes additional capacity from additional training supervision: blind refinement improves average MR by 0.89 points over +FE, while the pseudo-loss improves it by 1.31. Both effects are positive, but the full model gains 3.91 points. Relative to blind refinement, IMR+BSSI adds 3.02 average MR and 1.77 HD-mAP; relative to the pseudo-loss, the gains are 2.60 and 1.83. The corresponding mAP@0.75 margins are 4.11 and 3.90, showing that the difference extends to strict localization. These controls support the value of prediction-conditioned inference updates beyond either intervention alone. Parameter matching does not imply matched runtime: the repeated decoder evaluations remain an additional cost.

Table 12: Controls on QVHighlights validation. Added parameters are relative to the +FE baseline.

Configuration	Added params	mAP@0.5	mAP@0.75	Avg. MR	HD-mAP	HIT@1
+FE baseline	–	71.23	54.42	53.03	42.54	70.37
Blind refinement	1.0M	72.04	55.10	53.92	42.85	70.86
+ saliency pseudo-loss	–	72.08	55.31	54.34	42.79	70.94
IMR+BSSI	1.0M	**74.87**	**59.21**	**56.94**	**44.62**	**73.25**

Different architecture and features. We insert the modules into Moment-DETR using CLIP+SlowFast features (Table 13). IMR alone increases average MR by 2.61 points, R1@0.7 by 3.27, and HD-mAP by 1.27. Adding BSSI contributes a further 1.42, 2.16, and 1.14 points on these metrics. The complete model therefore gains 4.03 average MR, 5.54 R1@0.5, 5.43 R1@0.7, 2.41 HD-mAP, and 4.57 HIT@1 over Moment-DETR. Improvements on every reported metric show that feedback remains useful with another detector and feature extractor. This experiment establishes portability across the two evaluated settings; it does not require the absolute scores of those settings to be comparable. The baseline and final rows also agree with the corresponding results in the main text.

Table 13: Moment-DETR with CLIP+SlowFast on QVHighlights.

Configuration	Avg. MR	R1@0.5	R1@0.7	HD-mAP	HIT@1
Moment-DETR	32.20	53.94	34.84	35.65	55.55
+IMR	34.81	57.62	38.11	36.92	58.17
+IMR+BSSI	**36.23**	**59.48**	**40.27**	**38.06**	**60.12**

C Efficiency and the Iteration Trade-Off

Table 14 profiles the detector with pre-extracted features on an RTX 4090. Latency uses batch size one and is averaged after warm-up; FLOPs cover the detector forward pass. Peak memory is measured during training at batch size 16, so it should not be read as inference memory. The frozen InternVideo2-1B encoder is excluded from these measurements.

Relative to SG-DETR, IMR adds 0.59M parameters, 3.1 GFLOPs, and 6.3 ms. BSSI adds a further 0.40M parameters, 0.7 GFLOPs, and 2.5 ms. The full increase is 0.99M parameters (2.6%), 3.8 GFLOPs (23.3%), 1.60 GB training memory (8.7%), and 8.8 ms latency (40.7%). Shared decoder weights explain the small parameter increase but do not eliminate repeated computation or activation storage. These figures describe detector overhead, not end-to-end video processing cost.

The depth ablation in Table 6 complements this profile. Moving from $T = 2$ to $T = 3$ improves average MR by 1.78 points; moving to $T = 4$ reduces it by 0.14 and mAP@0.5 by 0.39. HD-mAP increases only 0.04 at $T = 4$, while HIT@1 declines 0.03. Thus $T = 3$ is the adopted compromise across both tasks, rather than the best setting for every individual metric.

Table 14: Detector complexity and measured runtime.

Configuration	Params (M)	FLOPs (G)	Peak memory (GB)	Latency (ms)	Decoder calls
SG-DETR	38.72	16.3	18.42	21.6	1
+IMR ($T = 3$)	39.31	19.4	19.61	27.9	3
+IMR+BSSI	39.71	20.1	20.02	30.4	3

D Multi-Seed Robustness

Table 15 reports the mean and standard deviation over five training seeds with the same split and hyperparameters. The +FE row is the controlled ablation baseline, distinct from the published SG-DETR row in Table 1. IMR improves mean average MR from 53.00 to 55.91 and HD-mAP from 42.51 to 43.17; BSSI further raises them to 56.88 and 44.57. The full-model gains over +FE are 3.60, 4.75, 3.88, 2.06, and 2.86 points for mAP@0.5, mAP@0.75, average MR, HD-mAP, and HIT@1.

Each mean gain is larger than the reported standard deviations for the corresponding configurations. The five-seed full-model means are also within 0.09 points of the single-run ablation results on all five metrics, preserving the same ordering. This supports robustness to training randomness in this setting. Standard deviations summarize dispersion across runs; they are not confidence intervals or a paired significance test.

Table 15: Five-seed mean $\pm$ standard deviation on QVHighlights validation.

Configuration	mAP@0.5	mAP@0.75	Avg. MR	HD-mAP	HIT@1
+FE	71.19±0.24	54.37±0.22	53.00±0.18	42.51±0.13	70.31±0.25
+IMR	72.84±0.20	56.95±0.18	55.91±0.16	43.17±0.11	71.27±0.20
+IMR+BSSI	**74.79±0.18**	**59.12±0.16**	**56.88±0.14**	**44.57±0.10**	**73.17±0.20**

E HYPERPARAMETER SENSITIVITY

Proposal half-window. The half-window w_p controls the local average used to estimate BSSI proposal widths, independently of the SRM smoothing window w_s. Table 16 shows that $w_p = 1$ performs best on every reported metric. Relative to using the peak value alone ($w_p = 0$), the three-clip neighborhood improves average MR by 0.46 and HD-mAP by 0.41. Expanding to five clips ($w_p = 2$) reduces these metrics by 0.22 and 0.15. Local averaging therefore offers a useful middle setting, while the modest range suggests that performance is not sharply tuned to a single window. This operation smooths the value used for width estimation; it does not remove the selected peak or directly change its center.

Table 16: Sensitivity to BSSI proposal half-window w_p.

w_p	mAP@0.5	mAP@0.75	Avg. MR	HD-mAP	HIT@1
0	74.35	58.66	56.48	44.21	72.81
1	**74.87**	**59.21**	**56.94**	**44.62**	**73.25**
2	74.61	58.93	56.72	44.47	73.08

Numbers of predicted and saliency-derived spans. Table 17 varies the number of decoder predictions used by SRM/MRM and the maximum number of BSSI proposals separately, keeping the other setting at its default. For $K_{\text{pred}} \in \{3, 5, 7, 10\}$, average MR varies by at most 0.23 points and HD-mAP by 0.31. Five predictions give the best values, while seven remain close: 56.88 versus 56.94 average MR and 44.55 versus 44.62 HD-mAP. Adding more candidates is therefore not uniformly beneficial.

The BSSI count follows the same trend. Increasing K_{peak} from three to five improves average MR by 0.32 and HD-mAP by 0.34, whereas seven proposals slightly lower both scores. These results support a compact evidence set: too few spans may omit useful context, while additional candidates can include redundant or less reliable evidence. This is a mechanism-consistent interpretation, rather than a direct measurement of candidate quality. The broader peak-density analysis in Appendix F tests a related limitation.

Table 17: Sensitivity to K_{pred} in SRM/MRM and K_{peak} in BSSI.

Setting	mAP@0.5	Avg. MR	HD-mAP	HIT@1
$K_{\text{pred}} = 3$	74.59	56.71	44.31	72.96
$K_{\text{pred}} = 5$	**74.87**	**56.94**	**44.62**	**73.25**
$K_{\text{pred}} = 7$	74.80	56.88	44.55	73.17
$K_{\text{pred}} = 10$	74.68	56.79	44.43	73.04
$K_{\text{peak}} = 3$	74.44	56.62	44.28	72.89
$K_{\text{peak}} = 5$	**74.87**	**56.94**	**44.62**	**73.25**
$K_{\text{peak}} = 7$	74.70	56.76	44.49	73.08

F FAILURE-MODE DIAGNOSTICS

Does one-shot saliency actually fail? Table 18 partitions QVHighlights examples by the AP of their initial saliency map. The bottom third gains 6.12 mAP@0.5 and 3.27 HD-mAP, compared with 1.24 and 0.72 in the top third. Thus the improvement is largest where the initial relevance estimate is least reliable, while remaining positive in all three groups. This association supports the proposed motivation: inaccurate initial saliency leaves more room for correction from localized moment evidence. It does not imply that every low-quality example is repaired.

Figure 3 gives two complementary examples. In the upper case, ReFocus-DETR predicts $[14, 110]$ seconds for a $[10, 108]$ annotation, whereas SG-DETR retrieves only $[56, 86]$. In the lower case, ReFocus-DETR matches $[32, 54]$, while SG-DETR selects the earlier $[8, 28]$ interval. The final saliency curves provide context for these localization differences; they do not display the trajectory of the iterative updates.

Table 18: Gain from +FE to the full model, stratified by initial-saliency AP.

Initial-saliency quality	Examples	Δ mAP@0.5	Δ HD-mAP
Bottom third	517	+6.12	+3.27
Middle third	516	+3.58	+1.91
Top third	517	+1.24	+0.72

When do saliency-peak proposals degrade? Table 19 groups examples by the number of valid peaks before selecting at most K_{peak} proposals. Mean proposal IoU falls from 0.68 for one or two peaks to 0.57 for three to five and 0.43 for six or more. The HD-mAP gain from BSSI declines in parallel, from 1.62 to 1.37 and 0.81 points, but remains positive in every group. The six-or-more category counts available peaks, not the number retained by BSSI.

The pattern identifies a limitation of saliency-derived proposals: dense candidate peaks can make temporal support less discriminative. Peak count is only a proxy for this ambiguity, so the table alone cannot separate weak saliency, repeated events, and nearby distractors. The positive gains indicate that BSSI remains useful in the evaluated groups, while motivating adaptive proposal selection for difficult videos.

Table 19: BSSI proposal fragility versus valid saliency-peak count.

Valid peaks per video	Mean proposal IoU	HD-mAP gain from BSSI
1–2	0.68	+1.62
3–5	0.57	+1.37
6+	0.43	+0.81

G Implementation Clarifications

State and gradient flow. BSSI takes the initial clip memory $\mathbf{M}_0$ and saliency $\mathbf{s}_0$; its outputs initialize $\mathbf{s}^{(0)} = \mathbf{s}_B$ and $\mathbf{M}^{(0)} = \mathcal{A}(\mathbf{M}_B, \mathbf{s}_B)$. In Fig. 2, S_0 and S_B denote these saliency states. The IMR panel summarizes the loop: the SRM output enters the amplifier, and MRM operates on the re-amplified base memory before the next decoder call. The final MR output comes from the decoder and the final HD output from SRM. The ATSS/denoising query-generation blocks belong to the underlying detector.

Predicted spans and confidences are detached before refinement, as is the previous saliency state in SRM. This blocks gradients through the prediction conditioning path without freezing the pooling, gates, or shared decoder. The loss can still train refiners through their outputs and later decoder calls. The cycle term is a training constraint on pooled BSSI features, not an additional inference loop.

Dimensions and normalization. Both $\mathbf{M}$ and $\mathbf{V}$ contain L clip vectors of dimension d; decoder queries are separate. Clip indices are zero-based, so $u_\ell = (\ell + 0.5)/L$ stays inside $(0, 1)$. Gaussian support and position gates broadcast across channels, while pooled channel vectors broadcast across clips. Positive denominator offsets prevent undefined normalization. Small gate initialization limits residual updates, but the subsequent layer normalization means the complete memory map is not exactly the identity.

Scope of the controls. The $T = 1$ row is the non-iterative +FE control, with SRM and MRM disabled. The depth, SRM/MRM, and MRM-stage ablations exclude BSSI; the cumulative component study adds BSSI after IMR. These scopes explain why the MRM-only score, the IMR score, and the full-model score differ. The +FE baseline and the published SG-DETR comparison share the InternVideo2 feature family but are separate result rows; their numerical gains are reported separately.

H Additional Prediction Results

Figures 4–7 illustrate different temporal extents and positions within the video. The full-duration example in Fig. 4 shows that a moment can cover the entire clip sequence. The short late event in Fig. 5 and the early event in Fig. 6 have predicted intervals matching their annotations. Figure 7 retains a two-second end-boundary error, despite correctly locating the event. These selected examples describe behavior rather than estimate dataset-level accuracy; aggregate performance is reported in the quantitative tables. Time labels are in seconds, and saliency curves are the final predictions.

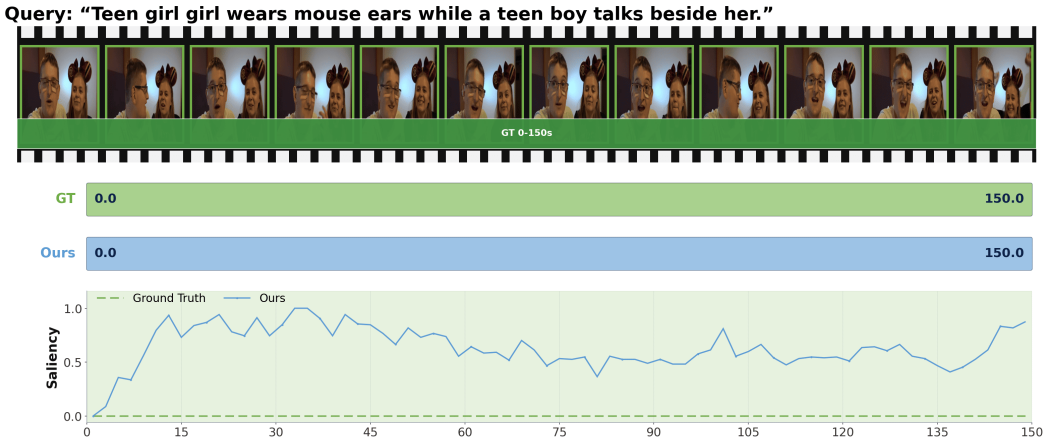

Figure 4: Full-duration localization. Both the annotated and predicted intervals span $[0, 150]$ seconds.

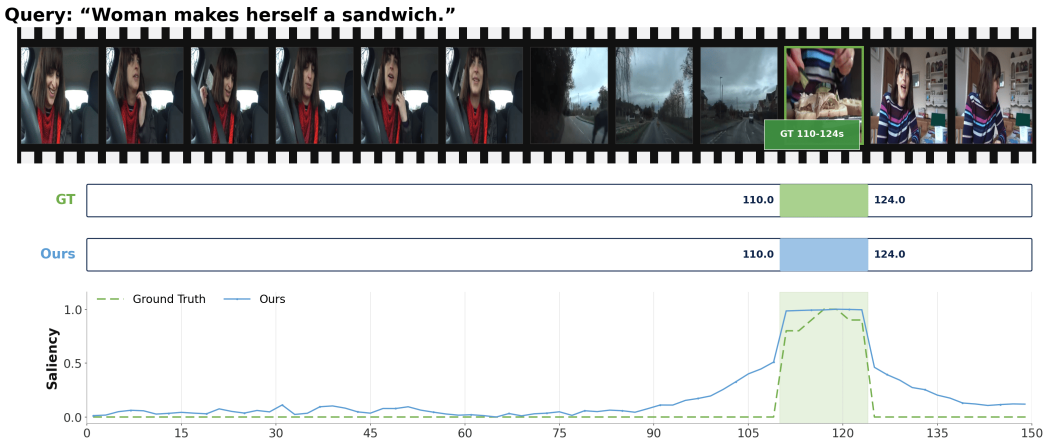

Figure 5: Short late event: making a sandwich. The predicted and annotated intervals both cover $[110, 124]$ seconds; saliency is concentrated around the event, with weaker responses nearby.



Figure 6: Early event: holding up a device. The predicted interval matches the $[4, 18]$-second annotation. Some saliency persists after the event, illustrating that correct boundaries need not imply perfect clip ranking.



Figure 7: Residual boundary error: eating at a table. ReFocus-DETR predicts $[114, 136]$ seconds against the $[114, 138]$ annotation. Elevated responses after the annotated end show a remaining saliency error.