CVA: Context-aware Video-text Alignment for Video Temporal Grounding (Supplementary Materials)

Sungho Moon* Seunghun Lee* Jiwan Seo Sunghoon Im†
DGIST
{byeol3325, lsh5688, eccaron, sunghoonim}@dgist.ac.kr

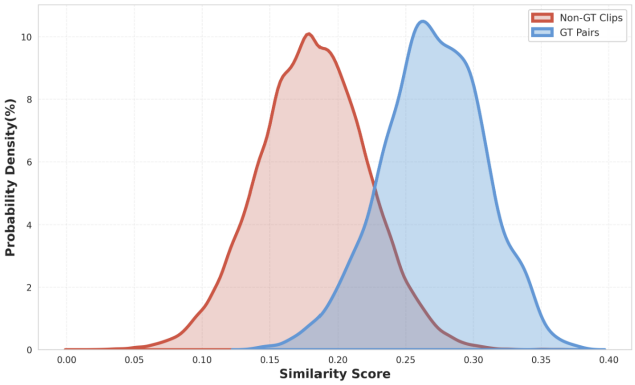


Figure 1. Density distribution of cosine similarity across all clip–text pairs in QVHighlights dataset. The clustered structure motivates selecting an intermediate similarity band for QCD.

Table 1. Ablation on background replacement ratio and context preservation window size in QCD, evaluated on the QVHighlights validation split. **Replace Ratio** denotes the fraction of background clips marked as $M{=}0$ (non-MR region), and **Context Size** indicates the number of adjacent MR-context clips preserved.

Replace Ratio	Context Size	Moment Retrieval			Highlight Detection	
		R1@0.5	R1@0.7	mAP@0.5	mAP	HIT@1
0.2	0	64.12	50.54	64.88	40.78	67.40
0.3	0	64.45	50.92	65.37	41.38	68.02
0.4	0	64.04	51.23	64.58	41.21	67.76
0.3	1	**65.32**	**51.98**	**65.92**	**41.72**	**68.44**
0.3	2	64.88	51.43	65.41	41.33	67.21
0.3	3	64.71	51.20	65.22	41.25	66.98

1. Additional Details about Query-aware Context Diversification (QCD)

Query-aware Context Diversification (QCD) aims to generate synthetic training samples that are both **semantically safe** and **sufficiently informative**. A key challenge is to avoid two failure modes in the video–text similarity space: (1) **low-similarity outliers**, which yield overly trivial negatives, and (2) **high-similarity false negatives**, where clips that are semantically close to the query are mistakenly treated as background. QCD therefore focuses on selecting an *intermediate* similarity band that provides meaningful augmentation without harming alignment.

1.1. Similarity Structure

To characterize the global similarity structure, we compute cosine similarity between each text query and *all* video clips in the QVHighlights dataset [2]. This produces a dense set of clip–text similarity scores. Figure 1 visualizes the resulting distribution, where similarity values exhibit clear clustering rather than a uniform continuum. This landscape provides the basis for determining QCD's safe and informative operating region.

1.2. Avoiding Extremal Similarity Regions

Low-similarity outliers. Clips with very low similarity correspond to unrelated backgrounds. Replacing content with such clips produces synthetic samples that are too easy, offering limited benefit in learning robust temporal discrimination.

High-similarity false negatives. Clips with high similarity often depict activities close to the queried action—even when originating from different videos. Using these clips as negative replacements introduces false-negative supervision, which can degrade retrieval accuracy. Examples in Fig. 2 illustrate such cases.

These observations support restricting QCD to an intermediate similarity band that excludes both extremes. In practice, we determine the intermediate similarity band $[\theta_{\min}, \theta_{\max}]$ based on percentile statistics of the global similarity distribution; the exact values are reported in the main paper.

* These authors contribute equally to this work.
† Corresponding author.

Figure 2. Examples illustrating the risk of using high-similarity clips as negative replacements. Such clips contain semantically relevant actions, and including them introduces false-negative supervision.

1.3. Replacement Ratio and Context Preservation

QCD additionally controls how much of the video is replaced and how much temporal context should be preserved. Table 1 shows that a **moderate replacement ratio** produces the strongest results: too small reduces augmentation diversity, while too large disrupts the video's temporal structure. Similarly, preserving a **small boundary-adjacent context window** yields higher accuracy by maintaining essential temporal cues without overconstraining the augmentation process.

Together, these analyses indicate that QCD is effective when it operates **within an intermediate similarity band that excludes both trivial low-similarity backgrounds and high-similarity false negatives**, while using a moderate replacement ratio and a narrow boundary-adjacent context window. Under this configuration, QCD generates realistic yet alignment-consistent augmented samples, which in turn yields consistent gains in both moment retrieval and highlight detection performance.

2. Additional Details about Context-invariant Boundary Discrimination (CBD)

2.1. Motivation

Temporal boundaries constitute the most ambiguous and error-prone regions in moment retrieval. They occur at the interface between foreground and background, where

Table 2. Ablation on the scope of CBD anchor positions (QVHighlights val split). **w/o CBD** denotes the model using QCD and CTE but without CBD.

Anchor choice	R1@0.5	R1@0.7	mAP@0.5	mAP	HIT@1
w/o CBD	67.62	52.63	67.81	41.89	68.26
All clips	65.22	48.12	63.56	39.78	66.77
Center clips	67.88	52.36	67.98	42.01	68.93
Boundary clips (Ours)	**69.61**	**54.84**	**67.97**	**43.47**	**70.40**

semantic changes are abrupt and clip-level features exhibit significantly higher variance than within-moment interiors. Existing objectives—regression, bipartite matching, IoU-based losses, and rank-aware contrastive formulations—lack dedicated supervision for these transitional regions. This motivates a boundary-focused objective that (1) isolates boundary features, (2) enforces cross-view consistency under augmentation, and (3) strengthens discrimination against both temporally adjacent and semantically similar hard negatives.

2.2. Boundary-focused Anchor Selection

CBD is applied only at the start and end boundaries of each ground-truth moment. Since CBD relies on selecting a small set of temporal **anchor positions** at which contrastive consistency is enforced, we examine whether alternative anchor locations can serve the same role. To evaluate this, we compare three anchor-selection strategies: (1) *All clips*, which treats every temporal index within GT moments as an anchor, (2) *Center clips*, which uses only the central interior clips of the moment as anchors, and (3) our *Boundary clips* formulation, which anchors exclusively at the start and end boundaries.

As shown in Table 2, the **all-clip** variant yields the largest degradation across MR and HD metrics. When every index becomes an anchor, the contrastive objective overconstrains the feature space, suppressing natural temporal variation within the moment and interfering with regression losses that govern span prediction.

The **center-clip** variant softens this effect, but center positions exhibit stable and less ambiguous semantics; consequently, they do not provide the hard contrastive signals needed to correct boundary-related localization errors.

In contrast, the **boundary-only** strategy consistently achieves the best performance. Boundary positions are precisely where temporal uncertainty is highest and where semantic transitions between foreground and background occur. Anchoring CBD at these positions introduces informative positives and challenging negatives, enabling the model to refine its boundary-sensitive representations. These results confirm that effective CBD requires **boundary-centric anchor selection**, and that applying contrastive supervision to interior or global positions is not beneficial.

Table 3. **Boundary-IoU comparison before and after CBD.** Evaluation conducted on QVHighlights val split with boundary width w=2. Scores are computed for samples with whole-window IoU ≥ 0.7.

Method	Start IoU	End IoU	Boundary IoU
w/o CBD	48.97	51.02	50.00
w/ CBD (Ours)	**52.54** (+7.29% ↑)	**55.91** (+9.59% ↑)	**54.26** (+8.52% ↑)

2.3. Boundary-IoU: A Boundary-centric Evaluation Metric

Standard MR metrics evaluate overlap over the full moment span, which may remain high even when boundaries are misaligned. Since CBD explicitly targets boundary fidelity, we adopt **Boundary-IoU**, a metric designed to isolate boundary-localization quality.

Given a ground-truth moment $M_{\text{GT}} = [s, e]$ and boundary width w, where s and e denote the start and end times (in seconds), and in QVHighlights these timestamps correspond to 2-second clip boundaries, we define:

$$B_{\text{start}} = [\, s,\ \min(s + w,\ e)\,], \tag{1}$$

$$B_{\text{end}} = [\, \max(e - w,\ s),\ e\,]. \tag{2}$$

Predicted boundary regions for $M_{\text{pred}} = [s', e']$ are defined analogously. Boundary-IoU is computed as:

$$\text{Boundary-IoU} = \frac{\text{IoU}(B_{\text{start}}, B'_{\text{start}}) + \text{IoU}(B_{\text{end}}, B'_{\text{end}})}{2}. \tag{3}$$

This metric focuses solely on the regions where boundary errors occur, providing a direct and sensitive measure of CBD's impact. Table 3 compares Boundary-IoU scores with and without CBD. CBD consistently improves both start- and end-boundary accuracy, confirming that it effectively models boundary-sensitive representations that are not captured by conventional IoU-based metrics.

3. Additional Ablation Study of Context-enhanced Transformer Encoder (CTE)

The Context-enhanced Transformer Encoder (CTE) is designed **to leverage the inherent continuity of video signals by aggregating information from neighboring clips**. Since adjacent video-clips often share motion cues and local semantics, incorporating multi-scale temporal receptive fields helps the model form more stable and context-aware representations. This is particularly beneficial for moment retrieval, where precise localization requires understanding both short-term transitions (e.g., motion boundaries) and long-range temporal context.

Table 4 presents an expanded ablation study exploring different combinations of receptive field sizes. Each configuration such as {5, 15, 75} denotes the **temporal re-**

Table 4. **Ablation on the Context-enhanced Transformer Encoder (CTE) in QVHighlights val split.** Each CTE variant integrates different temporal receptive fields to capture multi-scale temporal context. CTE1 offers the best balance between accuracy and efficiency.

Variant	Moment Retrieval			Highlight Detection	
	R1@0.5	R1@0.7	mAP@0.5	mAP	HIT@1
Baseline (w/o CTE)	65.32	51.98	65.92	41.72	68.44
CTE1 (Ours, {5, 15, 75})	67.62	52.63	67.81	41.89	68.26
CTE2 ({3, 5, 15})	66.20	50.70	67.50	41.66	68.33
CTE3 ({3, 5, 25})	66.53	51.34	67.75	41.98	**69.81**
CTE4 ({5, 15, 25})	66.84	51.82	67.71	41.81	68.59
CTE5 ({5, 25, 75})	67.17	52.15	67.65	41.65	67.84
CTE6 ({3, 5, 15, 25})	66.59	50.44	**68.17**	**42.02**	68.91
CTE7 ({3, 5, 25, 75})	66.53	50.89	67.91	41.18	67.04
CTE8 ({3, 15, 25, 75})	**67.75**	51.34	67.55	41.40	67.68
CTE9 ({5, 15, 25, 75})	67.30	52.29	67.68	41.61	68.01
CTE10 ({3, 5, 15, 25, 75})	67.56	**52.82**	67.46	41.40	66.20

ceptive field sizes used at each block. Smaller windows (e.g., 3 or 5) capture fine-grained motion patterns, whereas larger windows (e.g., 25 or 75) provide global temporal cues. Combining them yields consistent improvements over the baseline without CTE. Especially, **CTE1 ({5, 15, 75})** demonstrates the best overall balance between retrieval accuracy and highlight detection performance. This configuration effectively integrates short-, mid-, and long-range temporal dependencies while maintaining minimal computational overhead. Consequently, we adopt **CTE1 as the default setting** in our model.

4. Additional Experimental Details

4.1. Implementation Details

Following prior works [1–3], we use pre-extracted multimodal features for all datasets. Video features are obtained from the pre-trained SlowFast network [1] and the CLIP vision encoder [3]. Text queries are encoded using the corresponding CLIP text encoder. All features are provided at the clip level and kept frozen during training. We set the number of learnable queries in CTE to 100. The hyperparameters α and β for QCD are kept identical across all datasets. All experiments are conducted on one NVIDIA A100 GPU (40GB memory, CUDA 11.8, Python 3.8).

4.2. Evaluation Metrics

We evaluate our model on three widely used benchmarks: QVHighlights, Charades-STA, and TACoS. Across these datasets, we follow standard protocols established in prior moment retrieval literature.

For **QVHighlights**, which includes both Moment Retrieval (MR) and Highlight Detection (HD) annotations, we report Recall@1 at IoU thresholds 0.5 and 0.7, mAP@0.5, and the average mAP computed over IoU thresholds from 0.5 to 0.95 with a step size of 0.05. For HD, we additionally report the HIT@1 metric, which measures whether the highest-scoring clip corresponds to a ground-truth highlight. This combination of metrics captures retrieval accuracy, temporal localization precision, and highlight scoring quality.

For **Charades-STA** and **TACoS**, we follow prior work and evaluate performance using Recall@1 at IoU thresholds 0.5 and 0.7. These datasets focus purely on moment retrieval without highlight labels, making R1-based localization accuracy the standard evaluation measure. This consistent metric set provides a comprehensive view of retrieval correctness, boundary alignment quality, and highlight detection performance across the different datasets.

Table 5. Spurious correlation diagnostic on QVHighlights *val* split. **Random** replaces GT clips with noise matching original feature statistics; **Zero** sets GT features to zero. Lower values indicate less reliance on background context.

Mask mode	Method	Spurious R1 ↓		Spurious mAP ↓	
		R1@0.7	R1@0.9	@0.75	Avg.
Random	TD-DETR	2.45	1.03	3.18	3.82
	Ours	**2.39**	**0.84**	**2.70**	**3.17**
Zero	TD-DETR	21.23	14.00	21.35	20.93
	Ours	**7.16**	**5.16**	**7.53**	**7.48**

5. Robustness to Spurious Correlations

A core claim of our framework is that CVA learns context-invariant representations rather than relying on spurious correlations between queries and static backgrounds. To directly validate this, we adopt the *target-masked* diagnostic protocol from TD-DETR [4]: (i) **Random masking** replaces GT-moment clips with noise matching the original feature statistics, and (ii) **Zero masking** removes GT content entirely by setting features to zero. If a model relies on background context rather than the actual target moment, it will still produce high retrieval scores under these masking conditions.

As shown in Table 5, our model consistently achieves lower spurious scores than TD-DETR under both protocols. Notably, under **Zero masking**, CVA reduces spurious R1@0.7 from 21.23 to **7.16** (a **66%** reduction), demonstrating that our model genuinely relies on the target moment content rather than contextual bias. This confirms that the combination of QCD augmentation and CBD loss effectively enforces context-invariant learning.

6. Qualitative Results and Analysis

To further compare temporal grounding behavior across models, we present qualitative examples in Fig. 3. The first example represents a challenging scenario in which the camera frequently focuses on food rather than the person cooking. Because the visual evidence for the target ac-

tion appears only intermittently across clips, accurate localization requires integrating both **short-range temporal transitions** (e.g., brief motion onsets or local dynamics) and **long-range temporal structure** (e.g., scene progression and repeated contextual cues). CG-DETR fails to identify the target moment, and TD-DETR captures only a marginal portion with limited alignment. In contrast, our model closely matches the ground-truth interval. This robustness arises from the combined contributions of **QCD** (which prevents semantic contamination during augmentation), **CTE** (which enhances multi-scale temporal reasoning across clips), and **CBD** (which sharpens boundary discrimination). These components collectively enable accurate grounding even when **clip-level appearance cues are weak, unreliable, or partially missing**.

In the second example, both CG-DETR and TD-DETR activate a number of **false-positive** segments that are not semantically related to the query. Our method suppresses these spurious responses and localizes the intended region more precisely, demonstrating stronger discriminative ability under complex and visually distracting background conditions.

The third example consists of a sequence of short, rapidly transitioning actions. Our model accurately resolves these **fine-grained temporal boundaries**, whereas CG-DETR merges them into a single coarse segment and TD-DETR fails to capture the initial action entirely. This highlights the effectiveness of our boundary-sensitive design in handling dense and fast-changing temporal structures.

Finally, as shown in Fig. 4, the predicted saliency distribution is concentrated sharply within the ground-truth interval. This provides an interpretable visualization of how our model identifies relevant temporal cues while suppressing irrelevant clips.

References

[1] Christoph Feichtenhofer, Hao Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 4

[2] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34:11846–11858, 2021. 1

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4

[4] 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)*, 2025. 4

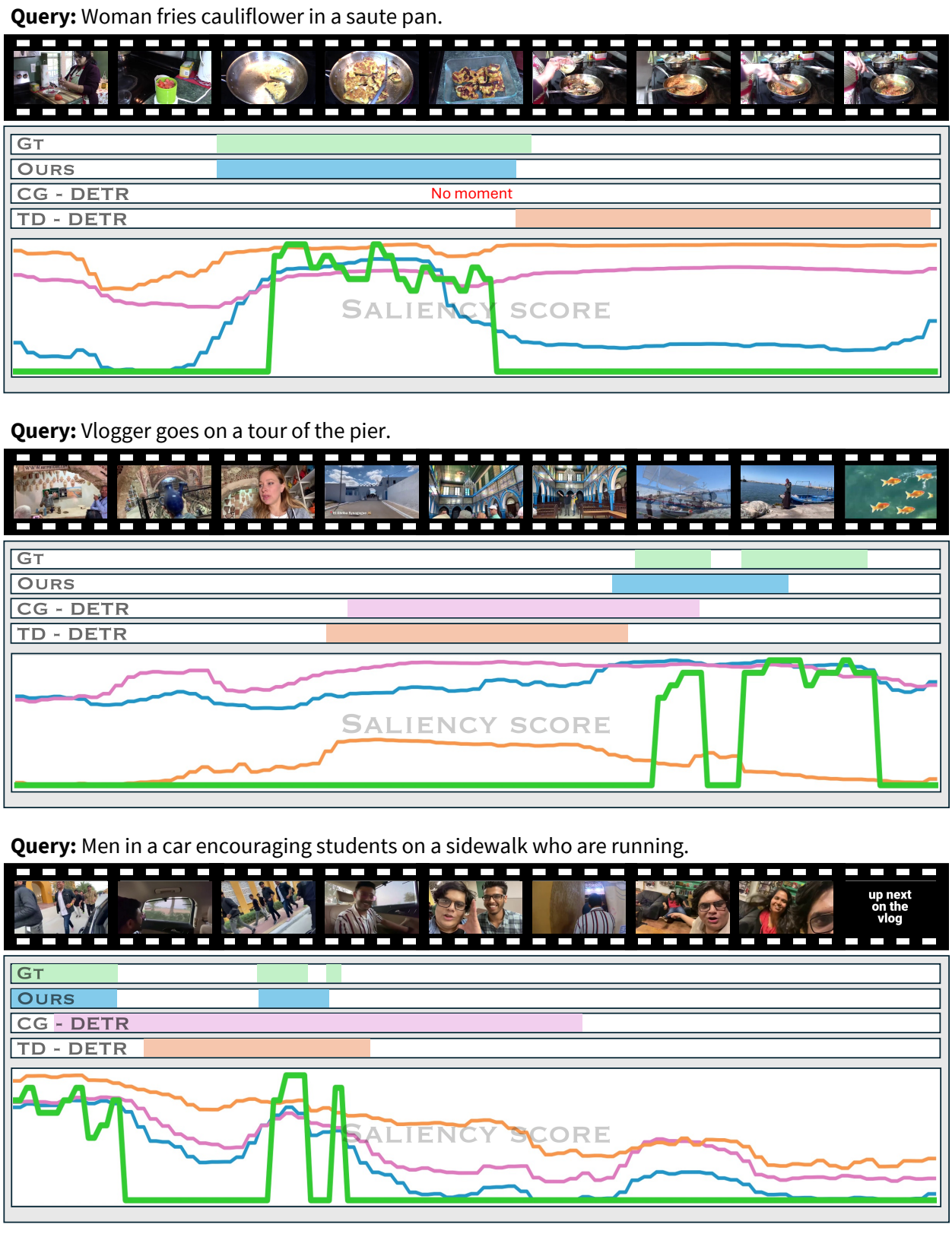
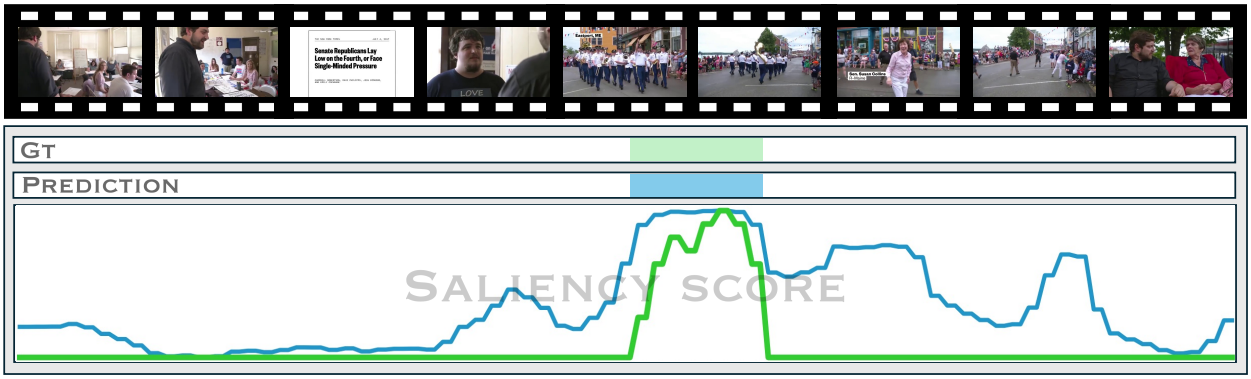


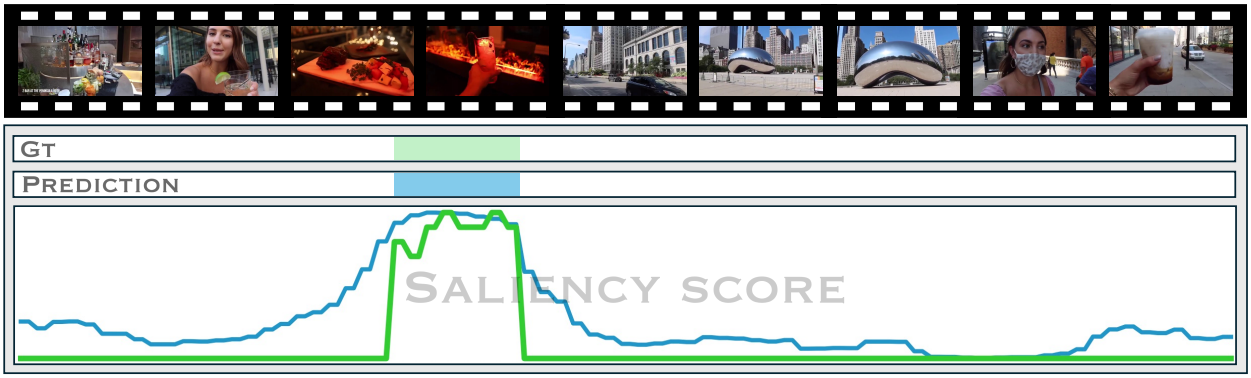
Figure 3. **Qualitative comparison with CG-DETR and TD-DETR on QVHighlights.** Our model aligns more accurately with ground-truth moments, reduces false positives, and resolves fine-grained temporal transitions more effectively. Saliency responses are also better concentrated within ground-truth intervals, reflecting improved moment discrimination.

Query: A marching band marches their way down the street during a parade.



Query: Women are riding on a train together.

Query: A large buffet is on a table.



Query: Three tweets are shown next to each other.

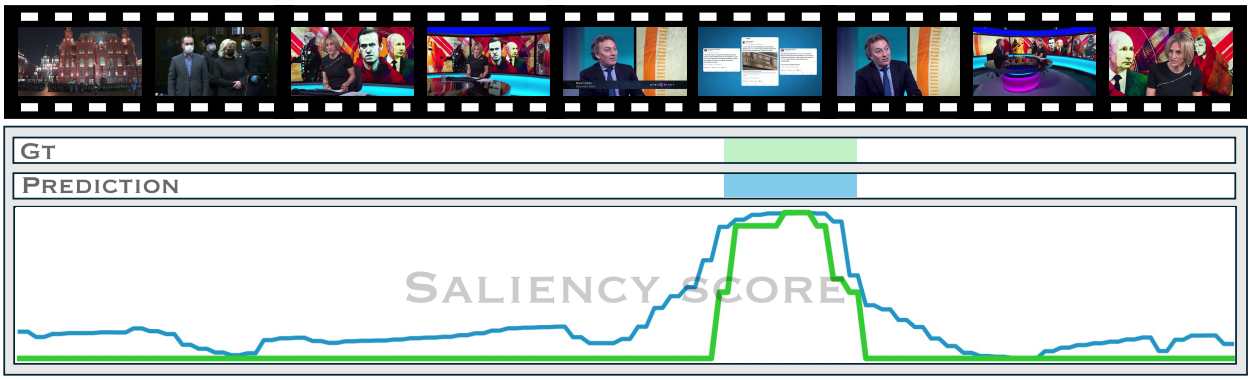


Figure 4. Qualitative results of our method.