DE^{2}TR: Dual Evidence Detection Transformer for Video Temporal Grounding

Supplementary Material(Appendix)

In this supplementary material, Sec. 1 illustrates the detailed architecture of the Modal Fusion and Interaction Pipeline and the Saliency Prediction Head. Sec. 2 describes more details of our proposed Gaussian Labeling Strategy and Boundary Augmentation Strategy. Sec. 3 provides more details on training implementation. Finally, Sec. 4 and Sec. 5 present additional quantitative analyzes and qualitative comparison results, respectively. The code will be released soon.

1 Detailed Network Architecture

1.1 Modal Fusion and Interaction Pipeline

As briefly described in Sec. 3.2 of the main paper, the modal fusion and interaction pipeline aims to align the visual and textual modalities and inject textual contexts into every video clip representation. To reduce the modality gap, we follow TR-DETR [9] to employ a local-global multi-modal alignment mechanism to align the video and text. After alignment, we further incorporate textual context into each video clip representation using cross-attention layers [8–10, 14]. Here, the aligned video features serve as the query (Q), while the aligned text features act as both the key (K) and the value (V). Then, a learnable saliency token is appended to the video sequence, and a standard Transformer encoder is utilized for further feature refinement [8, 11, 13]. Ultimately, this process generates the text-aware video representation $H_m \in \mathbb{R}^{L_v \times d}$ and a context-aware saliency token $H_t \in \mathbb{R}^d$. The representation H_m serves as the foundation for prior maps generation and the memory for the subsequent dual-branch decoding process. The saliency token H_t is used to evaluate the saliency score for each clip.

1.2 Saliency Prediction Head

To accurately predict clip-wise saliency scores, we adopt the input-adaptive saliency predictor used in the previous works [4,8,12,15]. To compute the saliency scores, we project both H_t and the H_m via independent fully-connected layer and calculate their scaled dot-product following [8]. This design effectively adjusts the prediction criteria for each specific video-query pair. Given that the saliency scores directly indicate the semantic correlation, we formulate these scores as our semantic prior map as described in Sec. 3.2 of the main paper.

Algorithm 1 Boundary Augmentation Strategy (BAS)

Input: Training video dataset $\mathcal{D}$, Maximum sequence length T.
Output: Augmented training dataset $\tilde{\mathcal{D}}$.

Initialize: $\tilde{\mathcal{D}} = \emptyset$
Augmentation Process:
for each video sample $(V, [t_s, t_e]) \in \mathcal{D}$ **do**
  $F = V[t_s : t_e],\ d = t_e - t_s$ ▷ Extract foreground & duration
  $P_{\mathrm{L}} \sim \mathrm{Uniform}(0, T - d),\ P_{\mathrm{R}} = (T - d) - P_{\mathrm{L}}$ ▷ Random padding
  $V_{\mathrm{bg}} \sim \mathcal{D}$ ▷ Sample background video
  $B_{\mathrm{L}} = \mathrm{Crop}(V_{\mathrm{bg}}, P_{\mathrm{L}}),\ B_{\mathrm{R}} = \mathrm{Crop}(V_{\mathrm{bg}}, P_{\mathrm{R}})$ ▷ Extract bg slices
  $\tilde{V} = [B_{\mathrm{L}}\ ;\ F\ ;\ B_{\mathrm{R}}]$ ▷ Temporal splicing
  $\tilde{t}_s = P_{\mathrm{L}},\ \tilde{t}_e = P_{\mathrm{L}} + d$ ▷ Update boundaries
  $\tilde{\mathcal{D}} = \tilde{\mathcal{D}} \cup \{(\tilde{V}, [\tilde{t}_s, \tilde{t}_e])\}$
end for
return $\tilde{\mathcal{D}}$

2 Extended Methodology

2.1 Gaussian Labeling Strategy

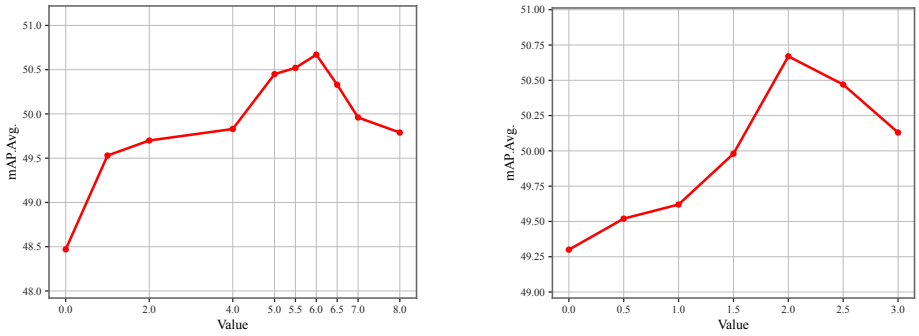
As mentioned in the main paper(Sec. 3.2), we employ an adaptive Gaussian labeling strategy to provide ambiguity-aware soft supervision. In this section, we provide more details about the base scale σ_0 and the ambiguity coefficient δ_k. First, inspired by the IoU-based radius rule in keypoint-based object detection [1, 3], we determine the base scale σ_0 by ensuring a minimum tIoU threshold τ. For a GT moment with width $w = t_k^e - t_k^s$, the maximum allowable boundary shift r that guarantees tIoU $\geq \tau$ can be mathematically bounded by:

$$\frac{w - r}{w + r} \geq \tau \implies r \leq w \frac{1 - \tau}{1 + \tau}. \tag{1}$$

Following the 3σ rule of the Gaussian distribution, we formulate the base scale as $\sigma_0 = r/3$. In the context of VTG, we set $\tau = 0.5$ according to the standard evaluation metric [2,5]. Subsequently, the base scale σ_0 is dynamically modulated by the ambiguity coefficient $\delta_k = \exp(\cos(\mathbf{v}_{in}, \mathbf{v}_{out}))$ to accommodate the specific semantic ambiguity of each boundary. Here, the inner and outer features ($\mathbf{v}_{in}$, $\mathbf{v}_{out}$) are extracted via mean-pooling over the adjacent σ_0-sized windows around the boundary to measure its ambiguity (*e.g.*, $[t_k^s, t_k^s + \sigma_0]$ and $[t_k^s - \sigma_0, t_k^s]$ for the start boundary t_k^s). Ultimately, by scaling σ_0 with δ_k, the final bandwidth σ_k naturally imposes strict supervision ($\sigma_k \downarrow$) on sharp boundaries while relaxing constraints ($\sigma_k \uparrow$) for ambiguous ones.

2.2 Boundary Augmentation Strategy

The core process of our Boundary Augmentation Strategy is summarized in Algorithm 1. This strategy mitigates the inherent boundary ambiguity in real-world videos and provides model with ideal boundary patterns.



(a) Impact of λ_{sea} & λ_{bea} (b) Impact of λ_{bnd}

Fig. 1: Sensitivity analysis of the loss coefficients. We report the Average mAP on the QVHighlights *val* split under different configurations of (a) the dual alignment loss weights λ_{sea} and λ_{bea}, and (b) the boundary regular loss weight λ_{bnd}.

3 More Implementation Details

Following previous work [7,15], we adopt a 3-layer encoder-decoder architecture. The hidden dimension is set to 256, with 8 attention heads and a dropout rate of 0.1. The number of learnable moment queries is default to 10. During training, we employ the Hungarian algorithm to perform bipartite matching between the predicted moment queries and GT targets. The objective coefficients are aligned with the matching costs, where the weights for the temporal span, gIoU, and classification are set to $\lambda_{span} = 1$, $\lambda_{giou} = 10$, and $\lambda_{cls} = 4$, respectively. The coefficient for the saliency loss is set to $\lambda_{sal} = 1$. Moreover, the weights for the local-global multi-modal alignment loss are configured as $\lambda_{\text{local}} = 0.3$ and $\lambda_{\text{global}} = 0.5$, following [9,10]. Finally, the learning rate is decayed by a factor of 10 every 80 epochs [4,15].

4 Additional Experimental Analysis

Effect of Hyper-parameters. To explore the impact of our proposed losses, we evaluate different settings of the dual alignment weights ($\lambda_{sea}, \lambda_{bea}$) and the boundary regular weight (λ_{bnd}), as shown in Fig. 1. The results demonstrate that absent or too small dual alignment weights fail to fully assist queries in capturing global semantics and precise boundary cues. Similarly, insufficient video-level regularization may cause the network to lose a key boundary prior. In contrast, overly large weights dominate the optimization process and degrade the primary grounding performance. Therefore, we empirically set $\lambda_{sea} = \lambda_{bea} = 6$ and $\lambda_{bnd} = 2$ in our model.

High-Ambiguity Splits. To further validate our model's robustness against boundary ambiguity, we construct three new subsets with varying ambiguity

Table 1: Performance on subsets with varying boundary ambiguity. Compared methods are reproduced using official codes. "w/ BAS": Boundary Augmentation Strategy.

Method	Extreme (20%)			Hard (50%)			Moderate (80%)			All (100%)		
	R1@0.5	R1@0.7	mAP	R1@0.5	R1@0.7	mAP	R1@0.5	R1@0.7	mAP	R1@0.5	R1@0.7	mAP
QD-DETR [8]	56.13	39.03	34.80	58.71	40.52	35.48	62.74	45.08	39.52	63.87	47.74	42.45
TR-DETR [9]	51.29	33.55	32.24	56.57	40.77	35.82	62.90	47.50	40.90	65.16	50.39	44.38
BAM-DETR [4]	51.94	38.71	38.10	56.90	42.19	39.98	62.58	47.90	44.83	63.68	51.37	47.84
TD-DETR [15]	54.52	40.65	38.82	58.19	43.61	39.94	63.71	49.03	44.92	65.74	52.97	48.83
Ours	60.00	41.32	40.40	62.06	45.68	42.64	67.26	51.69	48.09	68.13	53.81	50.67
Ours w/ BAS	**61.61**	**42.13**	**41.25**	**63.48**	**47.26**	**43.79**	**68.31**	**52.74**	**49.02**	**69.81**	**55.29**	**52.12**

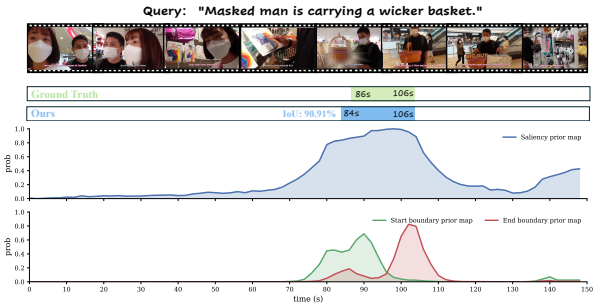


Fig. 2: Visualization Result of Dual Prior Maps

levels based on the QVHighlights *val* split. Specifically, we utilize our proposed ambiguity coefficient δ_k (Sec. 2.1) to filter the top 20% (Extreme), 50% (Hard), and 80% (Moderate) most ambiguous samples. This strict split challenges the models' capability to precisely locate boundaries under severe temporal ambiguity. As shown in Tab. 1, we report the performance of recent advanced methods on each subset in terms of R1@0.5, R1@0.7, and average mAP. Our proposed method achieves state-of-the-art performance across all subsets, exhibiting exceptional robustness to temporal ambiguity. Notably, integrating the BAS yields further consistent gains, confirming that our approach effectively mitigates inherent boundary ambiguities rather than simply overfitting to sharp transitions.

Visualization of Dual Prior Maps. As shown in Fig. 2, our dual prior maps exhibit explicitly complementarity. The semantic prior map (top) has a high response within the foreground span, while the boundary prior maps (bottom) generate distinct Gaussian-like peaks precisely at the event transitions. By providing such crucial structural prior knowledge, these maps can explicitly guide the moment queries to focus on potential regions of interest.

Computational Complexity Analysis. To address potential concerns about computational overhead introduced by the dual-branch design, we evaluated the computational complexity of DE^{2}TR against recent state-of-the-art methods. As shown in Table 2, thanks to our channel-wise query decoupling, which limits

Table 2: Comparison of computational complexity (GFLOPs) among recent DETR-based VTG methods.

Method	GFLOPs
TD-DETR [15]	1.10
MS-DETR [6]	**0.89**
DE^{2}TR (Ours)	1.12

Table 3: Performance of applying the BAS to existing baselines."w/ BAS" stands for using the BAS method.

Method	mAP
QD-DETR [8]	**41.45**
TR-DETR [9]	44.38
QD-DETR w/ BAS	41.17
TR-DETR w/ BAS	**44.85**

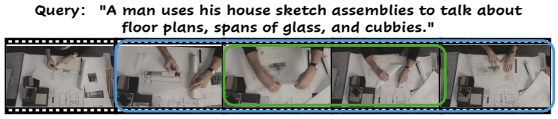


Fig. 3: A failure case of DE^{2}TR handling a visually homogeneous long take. The blue box represents our over-predicted span, while the green box is the Ground Truth.

computations to lower-dimensional sub-spaces, DEFormer introduces minimal computational overhead despite its dual-branch architecture, achieving optimal performance efficiently.

Discussion of diagnostic Experiments and BAS. BAS is proposed as a targeted augmentation strategy to help the dual-branch decoder better learn ideal boundary patterns (as shown in Fig. 4a). Ultimately, BAS succeeds where synthetic-only training fails (Fig. 1d) because existing baselines lack the capability to explicitly model boundaries. As shown in Table 3, applying BAS to existing baselines yields no significant performance gains. This empirical evidence proves that a dedicated boundary modeling architecture is the fundamental prerequisite for unlocking the effectiveness of BAS.

Failure Case Analysis To provide a comprehensive understanding of the limitations of our model, we present a typical failure case in Figure 3. As illustrated, despite significantly mitigating boundary ambiguity, DE^{2}TR struggles with visually homogeneous long takes that lack distinct semantic variance, leading to overly broad spans. Incorporating audio or ASR signals in future multi-modal architectures may present a viable solution.

5 Additional Qualitative Results

To further demonstrate the effectiveness of our method, we provide additional qualitative comparisons with state-of-the-art methods in Fig. 4 and Fig. 5. The visualization results across various challenging scenarios show that our model achieves more precise boundary localization and robust semantic alignment than strong competitors.

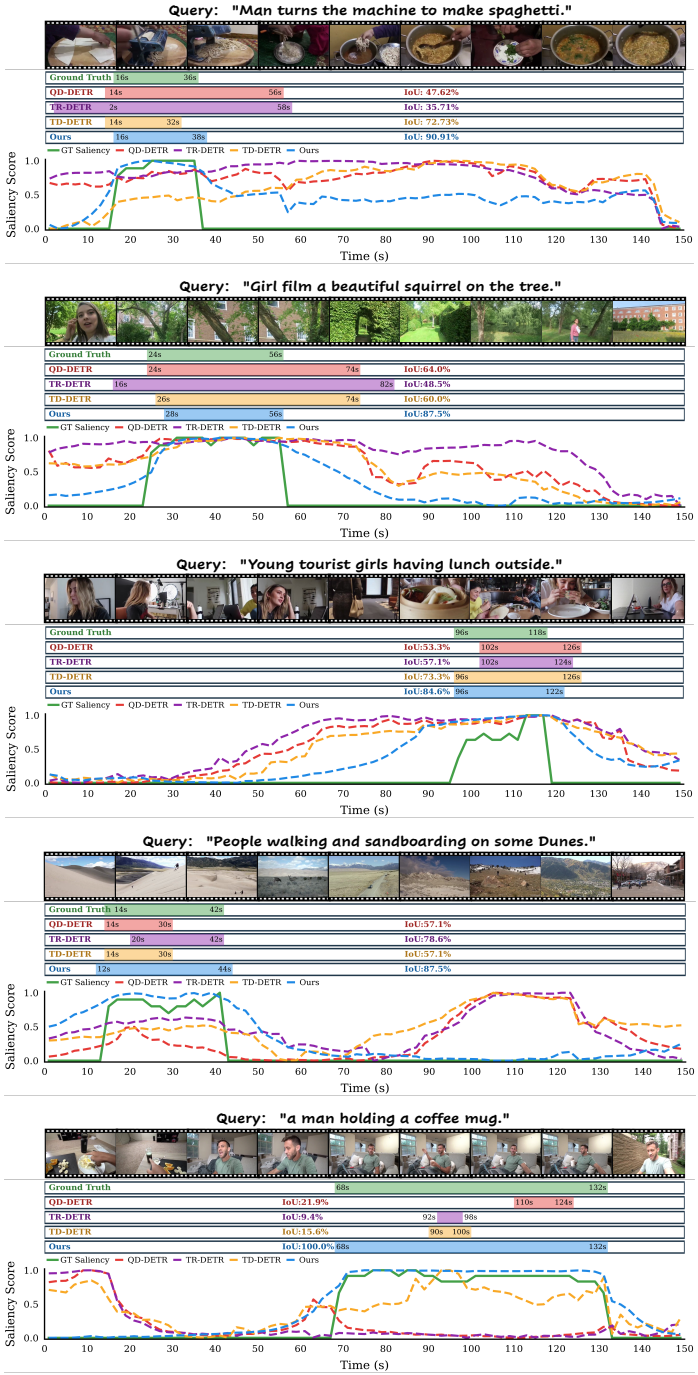


Fig. 4: Visualization Result on the QVHighlights *val* split.

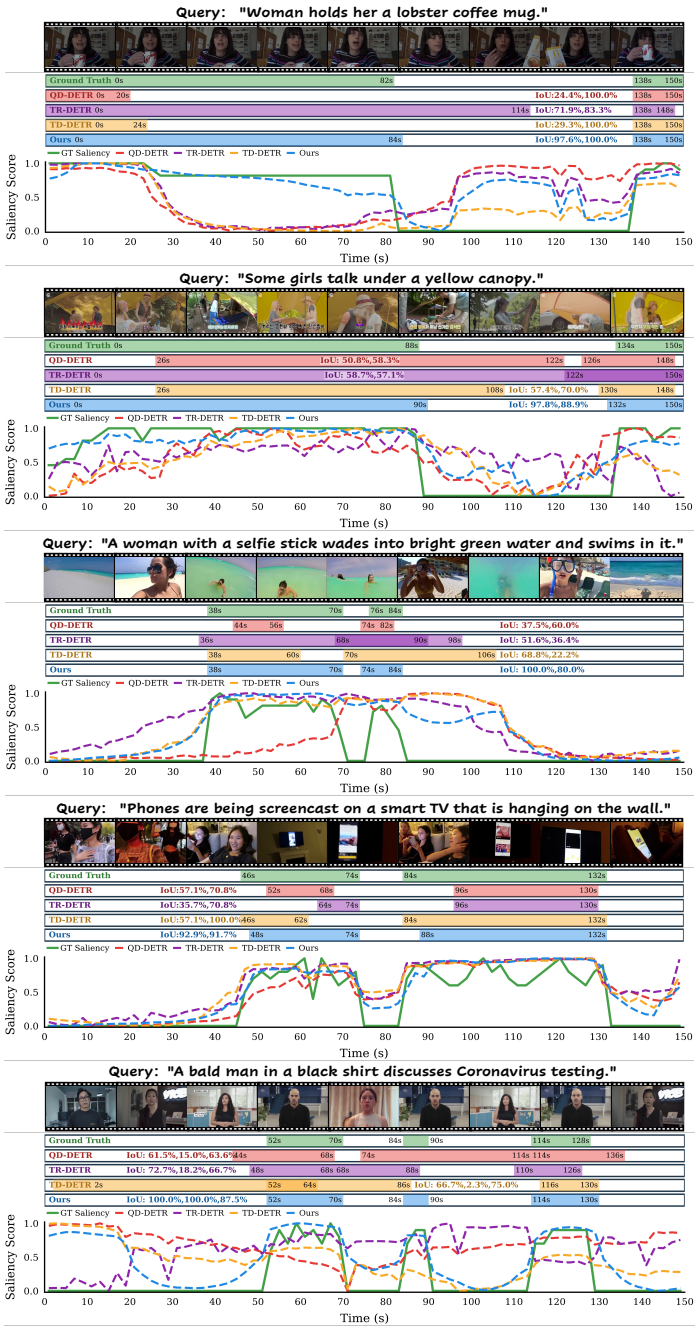


Fig. 5: Visualization Result on the QVHighlights *val* split.

References

1. Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., Tian, Q.: Centernet: Keypoint triplets for object detection. In: Int. Conf. Comput. Vis. pp. 6569–6578 (2019)
2. Gao, J., Sun, C., Yang, Z., Nevatia, R.: Tall: Temporal activity localization via language query. In: Int. Conf. Comput. Vis. pp. 5267–5275 (2017)
3. Law, H., Deng, J.: Cornernet: Detecting objects as paired keypoints. In: Eur. Conf. Comput. Vis. pp. 734–750 (2018)
4. Lee, P., Byun, H.: Bam-detr: Boundary-aligned moment detection transformer for temporal sentence grounding in videos. In: Eur. Conf. Comput. Vis. pp. 220–238. Springer (2024)
5. Lei, J., Berg, T.L., Bansal, M.: Detecting moments and highlights in videos via natural language queries. Adv. Neural Inform. Process. Syst. **34**, 11846–11858 (2021)
6. Ma, H., Wang, G., Yu, F., Jia, Q., Ding, S.: Ms-detr: Towards effective video moment retrieval and highlight detection by joint motion-semantic learning. In: ACM Int. Conf. Multimedia. pp. 4514–4523 (2025)
7. Moon, W., Hyun, S., Lee, S., Heo, J.P.: Correlation-guided query-dependency calibration for video temporal grounding. arXiv preprint arXiv:2311.08835 (2023)
8. Moon, W., Hyun, S., Park, S., Park, D., Heo, J.P.: Query-dependent video representation for moment retrieval and highlight detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 23023–23033 (2023)
9. Sun, H., Zhou, M., Chen, W., Xie, W.: Tr-detr: Task-reciprocal transformer for joint moment retrieval and highlight detection. In: AAAI Conf. Artif. Intell. pp. 4998–5007 (2024)
10. Sun, X., Shi, L., Wang, L., Zhou, S., Xia, K., Wang, Y., Hua, G.: Diversifying query: Region-guided transformer for temporal sentence grounding. In: AAAI Conf. Artif. Intell. pp. 7131–7139 (2025)
11. Sun, X., Wang, L., Zhou, S., Shi, L., Xia, K., Liu, M., Wang, Y., Hua, G.: Moment quantization for video temporal grounding. In: Int. Conf. Comput. Vis. pp. 20137–20146 (October 2025)
12. Um, S.J., Kim, D., Lee, S., Kim, J.U.: Watch video, catch keyword: Context-aware keyword attention for moment retrieval and highlight detection. In: AAAI Conf. Artif. Intell. pp. 7473–7481 (2025)
13. Xiao, Y., Luo, Z., Liu, Y., Ma, Y., Bian, H., Ji, Y., Yang, Y., Li, X.: Bridging the gap: A unified video comprehension framework for moment retrieval and highlight detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 18709–18719 (2024)
14. Yang, J., Wei, P., Li, H., Ren, Z.: Task-driven exploration: Decoupling and inter-task feedback for joint moment retrieval and highlight detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 18308–18318 (2024)
15. Zhou, X., Wei, F., Duan, L., Yao, A., Li, W.: The devil is in the spurious correlations: Boosting moment retrieval with dynamic learning. In: Int. Conf. Comput. Vis. pp. 20981–20990 (October 2025)