SA-DETR:Span Aware Detection Transformer for Moment Retrieval

Tianheng Xiong[1,2] , **Wei Wei**[1,2*] , **Kaihe Xu**[2,3], **Dangyang Chen**[2,3]
[1] Cognitive Computing and Intelligent Information Processing (CCIIP) Laboratory,
School of Computer Science and Technology, Huazhong University of Science and Technology
[2] Joint Laboratory of HUST and Pingan Property & Casualty Research (HPL),
[3] Ping An Property & Casualty Insurance company of China, Ltd.,
xiongtianheng52@gmail.com, weiw@hust.edu.cn
xukaihenupt@gmail.com, chendangyang273@pingan.com.cn

Abstract

Moment Retrieval aims to locate specific video segments related to the given text. Recently, DETR-based methods, originating from Object Detection, have emerged as effective solutions for Moment Retrieval. These approaches focus on multimodal feature fusion and refining Queries composed of span anchor and content embedding. Despite the success, they often overlook the video-text instance related information in Query Initialization and the crucial guidance role of span anchors in Query Refinement, leading to inaccurate predictions. To address this, we propose a novel **S**pan **A**ware **DE**tection **TR**ansformer (SA-DETR) that leverages the importance of instance related span anchors. To fully leverage the instance related information, we generate span anchors based on video-text pair rather than using learnable parameters, as is common in conventional DETR-based methods, and supervise them with GT labels. To effectively exploit the correspondence between span anchors and video clips, we enhance content embedding guided by textual features and generate Gaussian mask to modulate the interaction between content embedding and fusion features. Furthermore, we explore the feature alignment across various stages and granularities and apply denoise learning to boost the span awareness of the model. Extensive experiments on QVHighlights, Charades-STA, and TACoS demonstrate the effectiveness of our approach.

1 Introduce

Video has emerged as a leading form of media with the advancement of the Internet. The pressing need to extract valuable content from videos has driven the development of video understanding and retrieval tasks, including Video Action Recognition(Xu et al., 2020 Zhang et al., 2022a), Video Retrieval(Miech et al., 2019; Xue et al., 2022), and Video Question Answering(Yu et al., 2019; Yang et al., 2021). These methods enhance the retrieval and understanding of videos, but the fundamental task of locating relevant video segments based on specific description remains a challenge. For this, the task of Moment Retrieval(Gao et al., 2017; Anne Hendricks et al., 2017) has gradually developed in recent years.

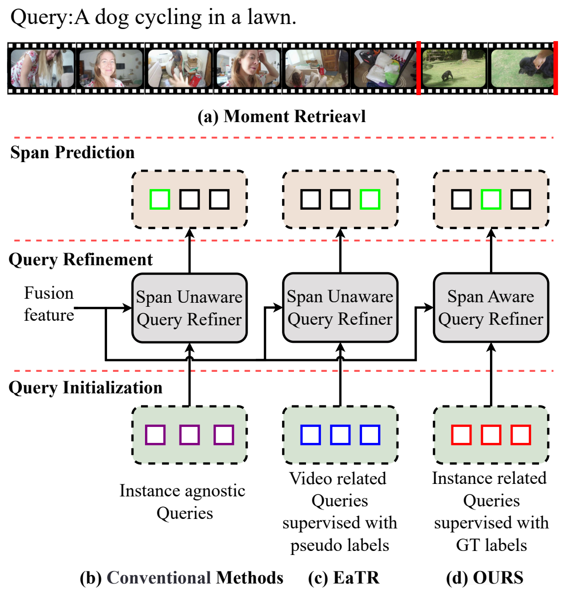

Figure 1: (a)Moment Retrieval. (b)(c)(d)Differences in Query Initialization and Query Refinement among various methods.

As illustrated in Figure 1(a), the goal of Moment Retrieval is to identify relevant video segments based on textual description. The key of Moment Retrieval hinges on achieving robust alignment and fusion between different modalities, as well as utilizing fused features to accurately locate segment boundaries. Previous works can be divided into proposal-based methods(Gao et al., 2017; Zhang et al., 2020b; Qu et al., 2020) and proposal-free methods(Yuan et al., 2019; Zhang et al., 2020a; Liu et al., 2021). While the former typically obtains

[*] Corresponding author.

Proceedings of the 31st International Conference on Computational Linguistics, pages 7634–7647
January 19–24, 2025. ©2025 Association for Computational Linguistics

localization results by ranking numerous carefully designed proposals, leading to higher precision but causing redundant computations, the latter directly predicts moments with fusion features, achieving higher efficiency but lacking boundary perception. The advent of Detection Transformer(Carion et al., 2020) balanced the precision and efficiency. Its Queries operate like proposals but without the complexity, and Hungarian matching supplants the burdensome Non-Maximum Suppression(NMS) post-processing. Consequently, it was rapidly adopted for Moment Retrieval, inspiring a range of DETR-based methods.

In DETR-based methods, Query typically consists of span anchor and content embedding. The former provides positional guidance, while the latter carries semantic information. In Query Initialization, conventional methods(Figure 1 b) overlook the instance related information by initializing span anchors as learnable parameters. Unlike Object Detection, which uses numerous anchor boxes to match varied object sizes within a single image, Moment Retrieval involves span anchors that are closely tied to video-text pairs. Learnable parameters in this context fail to provide sufficient prior knowledge. EaTR (Jang et al., 2023) (Figure 1 c) addresses the initialization issue by recognizing events in video using learnable event slots with slot attention. They generate span anchors based on these detected events and employ a Temporal Self-similarity Matrix(TSM) to construct pseudo labels for supervision. However, they assume multiple events in the video, and the pseudo labels generated by TSM are not accurate. In Query Refinement, previous methods do not fully leverage the guiding role of span anchors. They primarily utilize span anchors only as positional encoding to guide the refiner, overlooking the strong correspondence between span anchors and the video clip feature in Moment Retrieval.

In this paper, we propose **S**pan **A**ware **DE**tection **TR**ansformer(SA-DETR), which emphasizes the crucial role of span anchors in Moment Retrieval. Our method focuses on instance related Query Initialization and span aware Query Refinement. In Multi Modal Align Encoder, we align the visual and textual features in different granularities at multi fusion stages. In Dual Path Query Initializer, we initialize span anchors in direct Query group with instance related fusion token and supervise them with GT labels. Furthermore, we incorporate denoise learning to generate span anchors in noise Query group to simulate inaccurate initialization spans and provide additional supervision information. In Span Aware Refine Decoder, we introduce a span based enhance block to ease the semantic mismatch between content embedding and fusion feature. Additionally, span anchors are used to generate Gaussian mask to modulate the interaction between them directly in cross attention layers.

We have validated SA-DETR on several Moment Retrieval benchmarks, surpassing all previous methods and achieving competitive results. In summary, our contributions can be summarized as follows:

- We propose a novel SA-DETR that emphasizes the important role of instance related span anchors in Moment Retrieval.
- We explore the impact of feature alignment at different stages and granularities, and enhance the span awareness of the model with denoise learning.
- Experiments on QVHighlights(Lei et al., 2021), Charades-STA(Gao et al., 2017) and TACoS(Regneri et al., 2013) have demonstrated the effectiveness of our method.

2 Related Work

2.1 Moment Retrieval with DETR

Detection Transformer(DETR) was initially proposed for Object Detection, featuring a simple Encoder-Decoder architecture that eliminates the need for manually designed anchor boxes and complex NMS post-processing. Due to its high compatibility with Moment Retrieval, Moment-DETR(Lei et al., 2021) first introduced it to solve Moment Retrieval and Highlight Detection concurrently. Subsequently, a series of DETR-based Moment Retrieval methods were developed, among them, BM-DETR (Jung et al., 2023) enhanced background awareness and temporal sensitivity in videos, QD-DERT(Moon et al., 2023b) explored the significant role of textual queries in Moment Retrieval and Highlight Detection tasks, TR-DETR(Sun et al., 2024) and UVCOM(Xiao et al., 2024) discussed the differences and relations between Moment Retrieval and Highlight Detection tasks, EaTR(Jang et al., 2023) concentrated on the events occurring in the video, CG-DETR(Moon et al., 2023a) tried to guide multi modal interaction with their correlation, BAM-DETR(Lee and Byun, 2023) explored the different representation of span anchors. Our method

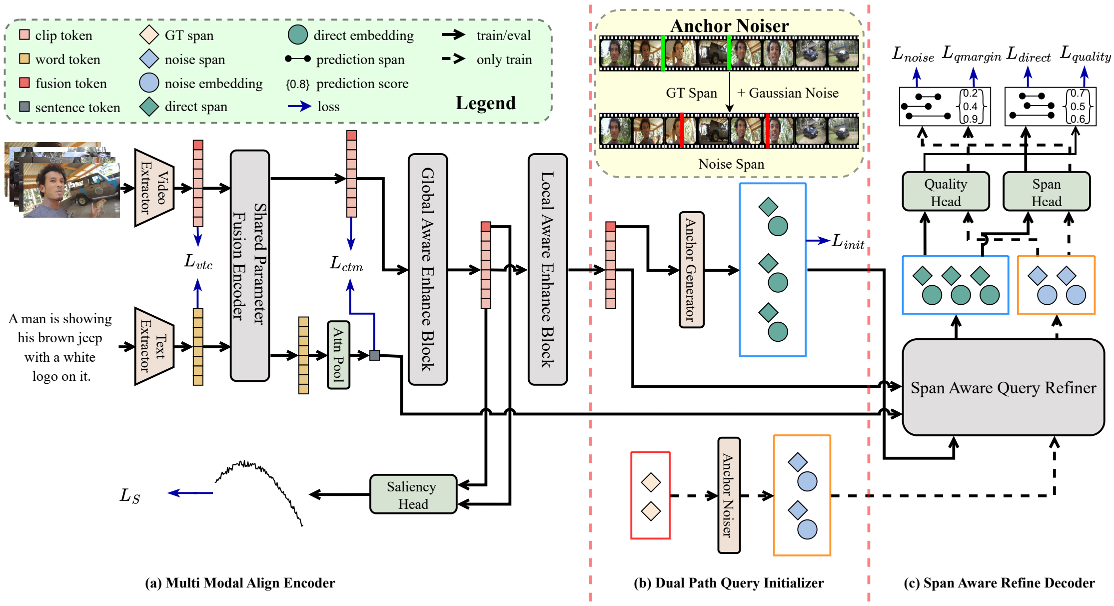

Figure 2: Overall of SA-DETR. For the given video-text pair, in the Multi Modal Align Encoder, we first extract features using frozen backbones, then align the visual and textual features at both video-text and clip-text levels before and after modal fusion. Additionally, we enhance the fusion visual feature from the different perspectives of Moment Retrieval and Highlight Detection tasks. In the Dual Path Query Initializer, we initialize span anchors with fusion token in the direct Query group and introduce noise to GT spans to generate span anchors in the noise Query group. In the Span Aware Refine Decoder, we refine the Queries using fusion feature with the guidance of corresponding span anchors and get the final prediction spans and quality scores. Specifically, noise Queries are used only at the training stage.

adopts DETR-based architecture, but unlike the above methods, we focus on instance related Query Initialization and span aware Query Refinement in Moment Retrieval.

2.2 Denoise Learning

DN-DETR(Li et al., 2022) first introduced denoise learning to address the slow convergence issue in DETR-based methods. This approach involves adding minor perturbations to GT bounding boxes as anchor boxes, providing a bypass for model convergence. DINO(Zhang et al., 2022b) expanded denoise learning into the contrastive setting, using varying degrees of noise as positive and negative groups. MomentDiff(Li et al., 2024) leveraged the generative diffusion model to recover video moments from noise, mitigating dataset biases and enhancing retrieval accuracy. DenoiseLoc(Xu et al., 2023) applied denoise learning to video activity localization tasks to mitigate boundary ambiguity. Similar to the above methods, we employ denoise learning with a contrastive setting. In addition to accelerating model convergence, span anchors generated with various noise scales in the noise Query group can effectively simulate the less precise span anchors initialized in the direct Query group, which can enhance the model's ability to refine accurate predictions from span anchors with various initial quality.

3 Method

3.1 Objective and Overall

For a given pair of video and text, we represent the video as L_v clips $\{C_1, C_2, ..., C_{L_v}\}$, and the text as L_t word tokens $\{W_1, W_2, ..., W_{L_t}\}$. The objective of Moment Retrieval is to locate the spans described in the text, denoted as $\{(c_i, w_i)_{i=1}^{N}\}$(c_i, w_i means the center and width of the span individually, and N is the count of spans related to the text). The goal of Highlight Detection is to compute the correlation scores $\{s_i\}_{i=1}^{L_v}$ of each video clip with text description. The overall of SA-DETR is illustrated in Figure 2.

3.2 Multi Modal Align Encoder

Feature Extractor. For the given video, we divide it into non-overlapping clips and employ a frozen video extractor to extract feature at the clip level to get the visual feature $F_v \in R^{L_v \times d_v}$. For the given text, we leverage a frozen text extractor to extract word-level textual feature as $F_t \in R^{L_t \times d_t}$.

Multi Stage Modal Aligner. The alignment and fusion of video and text are essential for the model to perceive their relationship. Previous methods(Lei et al., 2021; Moon et al., 2023b) directly merged visual and textual feature, neglecting their important connection. TR-DETR(Sun et al., 2024) aligned video and text feature at multiple levels, but overlooked the influence of the fusion stage. To this end, we developed a Multi Stage Modal Aligner that aligns features at the video-text level and clip-text level before and after the modal fusion, respectively. The former ensures that semantically related video and text are similar in semantic space, while the latter allows the model to recognize clip feature strongly associated with semantics. This alignment order helps the model understand the relationship between video and text in a coarse-to-fine path.

For visual feature $F_v \in R^{L_v \times d_v}$ and textual feature $F_t \in R^{L_t \times d_t}$, we first use two separate MLPs to project them onto the same dimension d, resulting in $F_v \in R^{L_v \times d}$ and $F_t \in R^{L_t \times d}$.

To apply video-text alignment, we use mean pooling to pool the video feature and text feature, then adopt the contrastive loss from CLIP(Radford et al., 2021) to obtain video-text contrastive loss L_{vtc}.

Subsequently, we concatenate the visual feature F_v with a learnable fusion token $g \in R^{1 \times d}$, then employ cross-attention layers with shared parameters to fuse visual and textual feature, resulting in text-related visual feature $\widehat{F_v} \in R^{(L_v+1) \times d}$ and video-related textual feature $\widehat{F_t} \in R^{L_t \times d}$. Specifically, we project visual feature F_v as Q_v, and textual feature F_t as K_t and V_t for text-related visual feature. The process for video-related textual feature is the reverse. Notably, we add positional embeddings to Q_v.

To perform clip-text alignment, we use attention pooling on video-related textual feature $\widehat{F_t}$ to derive sentence token $M_t \in R^d$. Then, we calculate the cosine similarity $S \in R^{L_v}$ between the visual feature $\widehat{F_v}$ without fusion token and M_s, then we employ L_{local} from TR-DETR and L_{intra} from UniVTG(Lin et al., 2023) for fine-grained alignment. Besides aligning clips with corresponding text, the model also needs to learn the non-corresponding between clips and unrelated texts. To achieve this, we incorporate L_{inter} from UniVTG. The clip-text matching loss is composed of three parts: $L_{ctm} = L_{local} + L_{intra} + L_{inter}$.

Local and Global Enhance Block. Both Moment Retrieval and Highlight Detection require video-text understanding but from different perspectives. Highlight Detection emphasizes the relevance differences between various clips and text, requiring global awareness. In contrast, Moment Retrieval focuses on locating segments of consecutive clips, necessitating local awareness. For this, we devise the local/global enhance block to enhance features according to the specific tasks.

For global awareness, we employ a standard Transformer Encoder as the global enhance block, resulting in $\widehat{M_v} \in R^{(L_v+1) \times d}$. Following QD-DETR(Moon et al., 2023b), we use $\widehat{M_v}$ to generate HD scores and saliency loss L_S.

For local awareness, we draw inspiration from UVCOM(Xiao et al., 2024) and apply a simple three-layer stacked 1D convolution with strides of 1, 3, and 1 as local enhancement block, resulting in $\overline{M_v}$. Finally, we split $\overline{M_v}$ into fusion feature M_v and fusion token M_g.

3.3 Dual Path Query Initializer

Direct Query group Initializer. To obtain video-text instance related initialization span anchors, we employ a straightforward method. With fusion token M_g and a simple three-layer MLP, we generate $S_d = MLP(M_g) \in \{(c_i, w_i)\}_{i=1}^{N_q} \in R^{N_q \times 2}$, where N_q is the number of direct Queries. These span anchors will be matched with GT spans through Hungarian matching, producing the initialization moment loss L_{init}. Additionally, the content embedding of direct Query group $C_d \in R^{N_q \times d}$ is initialized as learnable parameters of all zeros.

Noise Query group Initializer. We construct noise span anchors in noise Query group by perturbing the boundaries of GT spans. Specifically, for a given GT span(c, w) and a noise scale $\sigma \in (0, 1)$, we introduce random noise to generate noised span anchor$(c + \Delta c, w + \Delta w)$, ensuring that $|\Delta c| \leq \frac{\sigma c}{2}, |\Delta w| \leq \sigma w$, and that the noise span anchor remains valid. We use a contrastive learning approach to create positive and negative groups, simulating high-quality and low-quality span anchors separately. The noise scale of the negative noise group is a constant larger than that of the positive group $\sigma_p = \sigma_n + \delta$. For each GT, we generate N_d positive and negative noise span anchors. Additionally, the content embedding of the noise Query group is initialized as learnable parameter of all ones to distinguish from the direct Query group.

3.4 Span Aware Refine Decoder

Span Aware Query Refiner. To fully leverage the guidance of span anchors, we introduce the Span Aware Query Refiner, as depicted in Figure 3. We take the i-th refine process of direct Query group as an example. The input includes i-th span anchors $S_d^i \in R^{N_q \times 2}$, i-th content embedding $C_d^i \in R^{N_q \times d}$, fusion feature $M_v \in R^{L_v \times d}$ and sentence token $M_t \in R^d$.

Following previous methods, we use self-attention layers to exchange information between Queries and eliminate redundancy. Specifically, C_d^i is projected as $Q_{C_d^i}$, $K_{C_d^i}$ and $V_{C_d^i}$. Additionally, we convert S_d^i into positional embedding $P_{S_d^i} = MLP(PE(S_d^i)) \in R^{N_q \times d}$.The specific process is as follows:

$$\widehat{C_d^i} = softmax(\frac{(Q_{C_d^i} + P_{S_d^i})(K_{C_d^i} + P_{S_d^i})^T}{\sqrt{d}})V_{C_d^i} + C_d^i \tag{1}$$

We introduce the Span Based Enhance Block to enhance each content embedding with video clips from the corresponding span anchor, guided by textual memory. The goal is to mitigate the mismatch between content embedding and fusion feature in the cross-attention layers. First, we sample the fusion feature M_v based on span anchors S_d^i using Temporal Align(Xu et al., 2020), obtaining the sample feature $M_s = TemporalAlign(M_v, S_d^i) \in R^{N_q \times N_s \times d}$, where N_s is the number of clips sampled. Next, we modulate M_s with the sentence token M_t to enhance the M_s relevant to the text:

$$\begin{aligned} s &= \frac{W_v M_s * (W_t M_t)^T}{\sqrt{d}} \\ \widehat{M_s} &= mean(M_s \odot s) \end{aligned} \tag{2}$$

where $s \in R^{N_q \times N_s}$, W_s, W_t are learnable parameters, and $\odot$ represents element-wise multiplication. After obtaining the text-related sample feature $\widehat{M_s} \in R^{N_q \times d}$, we use gate fusion(Jang et al., 2023) to fuse it with $\widehat{C_d^i}$:

$$\begin{aligned} \widehat{g} &= diag(sigmoid(\widehat{C_d^i} * \widehat{M_s})) \\ \overline{C_d^i} &= W_f((\widehat{M_s} + \widehat{C_d^i}) \odot \widehat{g}) + \widehat{C_d^i} \end{aligned} \tag{3}$$

where $\widehat{g} \in R^{N_q}$, W_f is learnable parameters.

Next, we use cross-attention layers to fuse the content embedding and fusion feature M_v. We project $\overline{C_d^i}$ as $Q_{\overline{C_d^i}}$, and M_v as K_{M_v} and V_{M_v}, then apply positional encoding $P_{M_v} = PE(M_v) \in R^{L_v \times d}$ to M_v. We directly concatenate the feature and positional encoding instead of adding them to decouple the interaction of position and content(Liu et al., 2022), we get the attention map as:

$$map = \frac{(Q_{\overline{C_d^i}} || P_{S_d^i})(K_{M_v} || P_{M_v})^T}{\sqrt{2d}} \tag{4}$$

After obtaining the attention map $map \in R^{N_q \times L_v}$, inspired by CNM(Zheng et al., 2022), we use span anchors to generate Gaussian masks. i.e. for a span anchor (c, w):

$$mask = exp(-\frac{\alpha(i/L_v - c)^2}{w^2}), i = 1, ..., L_v \tag{5}$$

where α is a hyperparameter to control the scale of the Gaussian mask. These masks are used to modulate the attention map:

$$\widetilde{C_d^i} = softmax(Map \odot mask)V_{M_l} + Q_{\overline{C_d^i}} \tag{6}$$

Finally, we obtain the refined content embedding $C_d^{i+1} = FFN(\widetilde{C_d^i}) + \widetilde{C_d^i}$ with a simple feed-forward network.

Prediction Head. We use a simple three-layer MLP to predict the offset of the span anchor $\Delta_{S_d^{i+1}} = MLP(C_d^{i+1}) \in R^{N_q \times 2}$ and obtain the refined span anchor $S_d^{i+1} = S_d^i + \Delta_{S_d^{i+1}}$. Following BAM-DETR(Lee and Byun, 2023), we use a single-layer Linear to predict the Query quality $Q_{S_d^{i+1}} = sigmoid(Linear(C_d^{i+1})) \in R^{N_q}$.

3.5 Matching, Objective and Inference

Matching. For initialized spans and prediction spans in direct Query group, as there is no one-to-one correspondence with GT spans, we employ

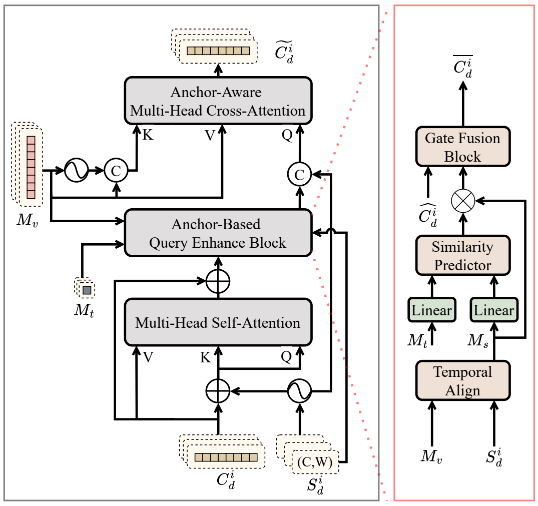

Figure 3: The structure of Span Aware Query Refiner

Hungarian matching to match them with GT spans. For prediction spans in noise Query group, we directly match prediction spans with their corresponding GT spans.

Moment Loss. Taking Direct query group as an example, for a real span m and its matched prediction span $\widehat{m}$, we use L1 loss and giou loss(Rezatofighi et al., 2019) to measure their difference:

$$L_{direct} = \sum_{j=1}^{N}(\lambda_{l1} L_{l1}(m_j, \widehat{m_j}) + \lambda_{giou} L_{giou}(m_j, \widehat{m_j})) \tag{7}$$

where N is the number of GT spans, $\lambda_{l1}, \lambda_{giou}$ are balance parameters for L_{l1} and L_{giou}. In addition, the L_{init} and the L_{noise} can be obtained in the same way. Note that only the prediction spans in the positive noise Query group produce moment loss. The total moment loss is $L_M = L_{direct} + L_{noise} + L_{init}$.

Quality Loss. The quality scores measure the quality of predictions directly. For the direct Query group, following BAM-DETR, we compute the maximum intersection ratio between each prediction span with all GT spans to determine the quality score. Additionally, to emphasize matched pairs, we assign a higher weight to those spans:

$$L_{quality} = \sum_{j=1}^{M} c_j |q_j - \max_{\forall n} |\frac{m_j \cap m_n}{m_j \cup m_n}|| \tag{8}$$

If m_j matches any GT spans, $c_j = w_q$, otherwise $c_j = 1$.

For the quality scores $Q_{S_n^P}$ and $Q_{S_n^N}$ generated by the positive noise Query group and corresponding negative noise Query group, we use the margin loss to enhance the model's ability to perceive the quality of Queries:

$$L_{qmargin} = \frac{1}{N_{gt}} \sum_{j=1}^{N_{gt}} max(q_j^n - q_j^p + \delta_q, 0) \tag{9}$$

where N_{gt} is the count of GT spans in a batch, δ_q is the margin between positive quality and negative quality, the total quality loss is $L_Q = L_{quality} + L_{qmargin}$.

Total Loss. Total loss of the model is composed of the following four parts: Moment loss L_M, Quality loss L_Q, Align loss $L_A = L_{vtc} + L_{ctm}$ and Saliency loss L_S:

$$L_{TOTAL} = \lambda_A L_A + \lambda_S L_S + \lambda_M L_M + \lambda_Q L_Q \tag{10}$$

whereλ_* is the balance weights.

Inference. Noise Query group is only enabled at the training stage. During the inference stage, we take the span with highest quality score as the final prediction.

4 Experiments

4.1 Datasets and Metrics

Datasets. We conduct experiments on QVHighlights, TACoS, and Charades-STA. Due to the space limitation, more details related to the datasets can be found in the Appendix A.1.

Metrics. We evaluate the model following previous works (Lei et al., 2021, Moon et al., 2023b). For Moment Retrieval, we default to reporting Recall@1 at IOU thresholds of 0.5 and 0.7, for QVHighlights with multiple GT spans, we record the mAP at IOU thresholds of 0.5 and 0.75, and also report the average mAP at IOU thresholds of [0.5:0.05:0.95], for TACoS, we also report the mIOU of the Top-1 Prediction. For Highlight Detection, we report the mAP and HIT@1 on the QVHighlights dataset.

4.2 Implement Details

Frozen Backbone. For a fair comparison, we choose pre-trained SlowFast(Feichtenhofer et al., 2019), CLIP(Radford et al., 2021), and VGG(Simonyan and Zisserman, 2014) as video extractor, and CLIP, Glove(Pennington et al., 2014) as text extractor. Specifically, for QVHighlights and TACoS, we cut the videos into 2-second clips then extract video feature using CLIP+SlowFast, and extract word tokens with CLIP. For Charades-STA, when using CLIP+SlowFast visual backbone, we cut the videos into 1-second clips and use CLIP for word tokens extraction. When utilizing the VGG feature, we divide the video into 1/8-second clips and encode the text using GloVe to obtain word tokens.

Training Settings. Among all experiments, we configure the Shared Parameter Fusion Encoder, Global Aware Enhance Block, and Span Aware Query Refiner with 2 layers each. We set the model dimension to 256 and the heads to 8 for all Transformer-like structures. We use AdamW(Loshchilov and Hutter, 2017) as the optimizer. All experiments were conducted on a single RTX3090 with torch2.2.1+cu118. Due to space limitation, more hyperparameters and loss settings will be found in Appendix A.2.

Method	MR					HD	
	R1		mAP			$\geq VeryGood$	
	@0.5	@0.7	@0.5	@0.75	Avg.	mAP	HIT@1
M-DETR	52.89	33.02	54.82	29.40	30.73	35.69	55.60
UniVTG	58.86	40.86	57.60	35.59	35.47	38.20	60.69
MH-DETR	60.05	42.28	60.75	38.13	38.38	38.22	60.51
QD-DETR	62.40	44.98	62.52	39.88	39.86	38.94	62.40
EaTR	61.36	45.79	61.86	41.91	41.74	37.15	58.65
TR-DETR	64.66	48.96	63.98	43.73	42.62	39.91	63.42
CG-DETR	**65.43**	48.38	64.51	42.77	42.86	**40.33**	**66.21**
UVCOM	63.55	47.47	63.37	42.67	43.18	39.74	64.20
BAM-DETR	62.71	48.64	64.57	46.33	45.36	-	-
SA-DETR	64.96	**49.09**	**65.30**	**47.80**	**47.40**	40.02	65.69

Table 1: Joint results of Moment Retrieval and Highlight Detection on QVHighlights online test split [1]

Method	feat	R1@0.5	R1@0.7
2D-TAN	VGG	40.94	22.85
QD-DETR	VGG	52.77	31.13
TR-DETR	VGG	53.47	30.81
MH-DETR	VGG	55.47	32.41
SA-DETR	VGG	**55.59**	**37.1**
2D-TAN	SF+C	46.02	27.40
M-DETR	SF+C	52.07	30.59
QD-DETR	SF+C	57.31	32.55
TR-DETR	SF+C	57.61	33.52
UniVTG	SF+C	58.01	35.65
CG-DETR	SF+C	58.44	36.34
UVCOM	SF+C	59.25	36.64
BAM-DETR	SF+C	59.95	39.38
SA-DETR	SF+C	**61.16**	**41.51**

Table 2: results on Charades-STA test split, SF denotes SlowFast, C denotes CLIP.

Method	R1@0.3	R1@0.5	R1@0.7	mIOU
2D-TAN	40.01	27.99	12.92	27.22
VSLNet	35.54	23.54	13.15	24.99
M-DETR	37.97	24.67	11.97	25.49
UniVTG	51.44	34.97	17.35	33.60
CG-DETR	52.23	39.61	22.23	36.48
UVCOM	-	36.39	23.32	-
BAM-DETR	56.69	41.54	26.77	39.31
SA-DETR	**58.16**	**42.56**	**27.87**	**40.03**

Table 3: results on TACoS test split

4.3 Main Results

Results on QVHighlights. As shown in Table 1, we compare the Moment Retrieval and Highlight Detection performance of SA-DETR with other DTER-based methods on the test split of QVHighlights. For the fair comparison, all models are trained from scratch with only video and text pairs without any pre-training. For Moment Retrieval, SA-DETR significantly outperforms previous methods on almost all metrics, which highlights the importance of the awareness of instance realted span guidance. Although the HD task is not the main focus of our method, the multi-stage feature alignment and fusion enables the model to achieve competitive results.

Results on Charades-STA & TACoS. We test the MR performance of our model on the test splits of Charades-STA and TACoS datasets. As shown in Table 2, on Charades-STA, regardless of whether we use VGG or SlowFast+CLIP backbone, our model achieves better performance. Particularly at a high IOU of R1@0.7, we surpass MH-DETR(Xu et al., 2024) by 4.69% and BAM-DETR by 1.13%. As shown in Table 3, on the TACoS dataset, our model outperforms all previous methods by a significant margin.

4.4 Ablation Studies

Main Components Ablation. As shown in Table 4, we conduct ablation experiments on QVHighlights and report the results on val split. Feature Align(FA) represents the multi-stage feature alignment, Query Initialization(QI) denotes the instance related Query initialization, Span Aware(SA) indicates the span aware Decoder, Denoise Learning(DN) signifies the contrastive denoise learning. Setting(a) serves as the baseline, consisting of a fusion Encoder with shared parameters and local/global enhance blocks, along with a decoder similar to DAB-DETR(Liu et al., 2022). In contrast, setting(j) represents the complete model with all components. The experiment results are as follows: 1) For settings (b) to (e), we verified that each components have a positive effect on model

[1] CodaLab online test server

settings	FA	QI	AD	DN	MR R1 @0.5	MR R1 @0.7	MR mAP Avg.	HD mAP	HD HIT@1
(a)					62.39	46.77	40.71	39.33	62.13
(b)	✓				65.16	49.23	44.26	40.24	67.29
(c)		✓			63.74	50.32	45.39	39.42	62.90
(d)			✓		63.35	48.19	42.85	39.45	62.65
(e)				✓	65.03	48.39	43.69	39.50	62.39
(f)		✓		✓	63.35	50.06	46.27	39.53	63.03
(g)		✓	✓		63.87	50.13	45.96	39.63	64.13
(h)		✓	✓	✓	63.74	50.52	47.01	39.75	63.55
(i)	✓	✓		✓	65.29	51.35	47.54	40.59	66.32
(j)	✓	✓	✓	✓	**67.03**	**52.52**	**48.84**	**40.81**	**67.61**

Table 4: Components ablation on QVHighlights val split.

Method	R1@0.5	R1@0.7	mAP
baseline	64.58	49.61	44.69
+Dynamic Anchor	64.97	51.16	47.19
+Init Loss	**67.03**	**52.52**	**48.84**

Table 5: Ablation on Query Initializer

modulate	enhance	R1@0.5	R1@0.7	mAP
		65.29	51.35	47.54
✓		66.84	52.77	48.14
	✓	66.58	**52.58**	47.97
✓	✓	**67.03**	52.52	**48.84**

Table 6: Ablation on Span Aware Query Refiner

L_{vtc}	L_{ctm}	MR R1 @0.5	MR R1 @0.7	MR mAP Avg.	HD mAP	HD HIT@1
before	before	66.71	52.71	48.24	**40.89**	66.58
after	after	65.42	51.74	47.56	40.20	65.94
after	before	65.55	51.87	47.85	40.65	65.81
before	after	**67.03**	**52.52**	**48.84**	40.81	**67.61**

Table 7: Ablation on Align Stage, before denotes before modal fusion, and after denotes after modal fusion

performance. 2) Compared to setting(c), setting(f) introduces DN, the noise Query group simulates the inaccurate span anchors in the direct Query group, and the combination of the two modules achieves a boost effect. Compared to setting(d), setting(g) adds QI , compared to the learnable instance unrelated span anchor, the instance related span anchor provided by QI plays a better guiding role in the Span Aware Decoder. 3) Compared to setting(j), setting(h) removes FA, and both Moment Retrieval and Highlight Detection performance significantly decreased, indicating that well-aligned feature plays an important role in both tasks. Compared to setting(j), setting(i) removes the AD. Without guidance in the refinement process of the span anchor, the model cannot locate accurate results, leading to a decline in MR performance.

Ablation on Query Initializer. As shown in Table 5, we set up ablation experiments on QVHighlights to verify the important role of Query Initialization. We replace the instance related span anchors with learnable parameters and remove the L_{init} as the baseline, the model's performance significantly decreased. After adding dynamic span anchors, the performance improved. Subsequently, by adding L_{init} to supervise the initialization of span anchors, the performance further enhanced.

Ablation on Span Aware Query Refiner. We conduct ablation experiments on the components of the Span Aware Query Refiner, as shown in Table 6, the result indicate that utilizing span anchors to enhance content embedding and modulating the interaction both have positive impacts on the model performance. Notably, when both techniques are used together, they lead to the highest performance improvement.

Ablation on Align Stage. As shown in Table 7, we investigated the impact of video-text level(L_{vtc}) and clip-text level(L_{ctm}) alignment on model performance during different stages of modal fusion. The experiments indicate that performing video-text level feature alignment before modal fusion, specifically before the share-parameter encoder, allows us to project paired video-text pairs into closer semantic space from a global perspective. This

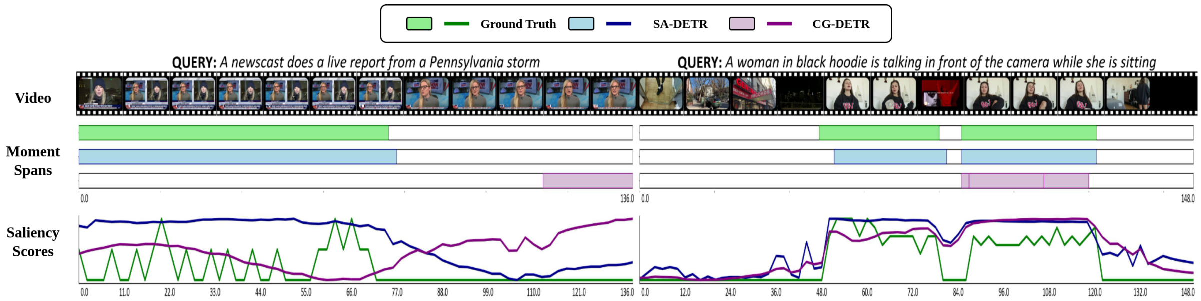

Figure 4: Qualitative Results.

Method	R1@0.5	R1@0.7	mAP
w/o contrastive groups	66.52	52.32	47.88
course learning	**67.55**	52.19	48.36
fixed margin	67.03	**52.52**	**48.84**

Table 8: Ablation on Contrastive Denoise Learning

facilitates their fusion and subsequent local alignment of clip-text level.

Ablation on Contrastive Denoise Learning. Negative noise Query group provides span anchors with high noise, which helps the model better evaluate the quality of different Queries. We conducted experiments, as depicted in Table 8, to ascertain the efficacy of contrastive setting. The model's performance significantly deteriorates when negative groups are not utilized. However, by implementing a coarse learning strategy that progressively diminishes the noise scale margin between negative groups and positive groups throughout the training process, the model's performance is enhanced. Furthermore, by maintaining a constant noise margin between negative and positive groups, the model can consistently discern the differences between them, leading to the most substantial performance improvement.

Convergence Speed. We compare the convergence speed and quality with other methods on QVHighlights and report the mAP on val split. As shown in Figure 5, when denoise learning is not used, although the performance of the model surpasses other methods, the early convergence of SA-DETR is comparable to other models. When denoise learning is added, the convergence speed and quality of the model significantly improve.

Due to space constraints, more ablation experiments can be found in Appendix A.3 .

4.5 Qualitative Results

As shown in Figure 4, we compare our prediction results with CG-DETR. In the left case, our method can precisely determine the segment related to the text and obtain correct saliency scores. In the right case, our method successfully locates the repetitive and complex segments without overlap.

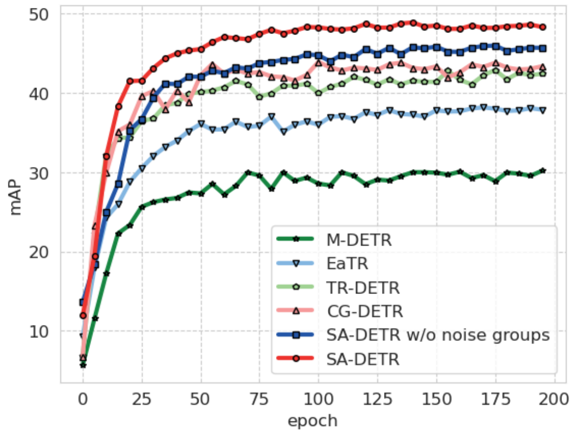

Figure 5: Comparison with other methods on convergence speed and quality, all models are trained from scratch with the official code on QVHighlights, we report mAP on val split here.

5 Conclusion

We propose the **S**pan **A**ware **DE**tection **TR**ansformer(SA-DETR), an effective method to address Moment Retrieval. In SA-DETR, we explore the importance of span anchors during the Query Initialization and Refinement. Specifically, we initialize span anchors using instance related fuse token and supervise them with GT labels. Additionally, we guide the Query refinement with span anchors to achieve more accurate localization. Furthermore, we investigate the impact of feature alignment at different granularities and stages on model performance and verify the boost effectiveness of denoise learning in the model's span awareness. Our approach achieves competitive results on QVHighlights, Charades-STA and TACoS, demonstrating its effectiveness.

Limitation

Although our method effectively addresses the moment retrieval task with awareness of span anchors, there are still certain limitations in the following aspects:

- We use Hungarian matching for both initialized spans and refined spans. However, we do not consider the stability of matching between different layers of span anchors and the GT labels. Consequently, there may be cases where a GT label matches different span anchors from different layers, leading to a decrease in model performance.
- While our method addresses Modal Fusion and Align, Highlight Detection, and Moment Retrieval within a unified framework, these three problems have distinct emphases and optimization goals. We simply optimize them simultaneously without considering their differences or the order of optimization.
- Our experiments only involve video and text modalities. We have not designed a general multimodal fusion structure to incorporate other modalities, such as audio.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62276110, No. 62172039 and in part by the fund of Joint Laboratory of HUST and Pingan Property Casualty Research (HPL). The authors would also like to thank the anonymous reviewers for their comments on improving the quality of this paper.

References

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer.

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

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275.

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

Minjoon Jung, Youwon Jang, Seongho Choi, Joochan Kim, Jin-Hwa Kim, and Byoung-Tak Zhang. 2023. Overcoming weak visual-textual alignment for video moment retrieval. *arXiv preprint arXiv:2306.02728*.

Pilhyeon Lee and Hyeran Byun. 2023. Bam-detr: Boundary-aligned moment detection transformer for temporal sentence grounding in videos. *arXiv preprint arXiv:2312.00083*.

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

Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. 2022. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13619–13627.

Pandeng Li, Chen-Wei Xie, Hongtao Xie, Liming Zhao, Lei Zhang, Yun Zheng, Deli Zhao, and Yongdong Zhang. 2024. Momentdiff: Generative video moment retrieval from random to real. *Advances in neural information processing systems*, 36.

Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. 2023. Univtg: Towards unified video-language temporal grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2794–2804.

Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. 2021. Context-aware biaffine localizing network for temporal sentence grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11235–11244.

Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. 2022. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640.

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

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

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Xiaoye Qu, Pengwei Tang, Zhikang Zou, Yu Cheng, Jianfeng Dong, Pan Zhou, and Zichuan Xu. 2020. Fine-grained iterative attention network for temporal language localization in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4280–4288.

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

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

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666.

Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. 2012. A database for fine grained activity detection of cooking activities. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1194–1201. IEEE.

Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 510–526. Springer.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Hao Sun, Mingyao Zhou, Wenjing Chen, and Wei Xie. 2024. Tr-detr: Task-reciprocal transformer for joint moment retrieval and highlight detection. *arXiv preprint arXiv:2401.02309*.

Yicheng Xiao, Zhuoyan Luo, Yong Liu, Yue Ma, Hengwei Bian, Yatai Ji, Yujiu Yang, and Xiu Li. 2024. 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*, pages 18709–18719.

Mengmeng Xu, Mattia Soldan, Jialin Gao, Shuming Liu, Juan-Manuel Pérez-Rúa, and Bernard Ghanem. 2023. Boundary-denoising for video activity localization. *arXiv preprint arXiv:2304.02934*.

Mengmeng Xu, Chen Zhao, David S Rojas, Ali Thabet, and Bernard Ghanem. 2020. G-tad: Sub-graph localization for temporal action detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10156–10165.

Yifang Xu, Yunzhuo Sun, Benxiang Zhai, Youyao Jia, and Sidan Du. 2024. Mh-detr: Video moment and highlight detection with cross-modal transformer. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. 2022. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. *arXiv preprint arXiv:2209.06430*.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2021. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1686–1697.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134.

Yitian Yuan, Tao Mei, and Wenwu Zhu. 2019. 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, pages 9159–9166.

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

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. 2022b. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*.

Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. 2020a. Span-based localizing network for natural language video localization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6543–6554.

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

Minghang Zheng, Yanjie Huang, Qingchao Chen, and Yang Liu. 2022. Weakly supervised video moment localization with contrastive negative sample mining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3517–3525.

Datasets	vid feat.	txt feat.	hyperparameters										loss							
			bs	epoch	lr	N_q	N_n	N_s	α	σ_p	σ_n	λ_{l1}	λ_{giou}	w_q	δ_q	λ_A	λ_S	λ_M	λ_Q
QVHighlights	SF+C	CLIP	32	200	1e-4	10	5	10	2	0.2	0.6	10	1	1	0.4	1	1	1	1
TACoS	SF+C	CLIP	32	200	2e-4	10	5	15	2	0.2	0.6	10	1	10	0.4	0	4	1	1
Charades-STA	SF+C	CLIP	32	200	2e-4	10	5	15	2	0.2	0.6	10	1	10	0.4	0	4	1	1
Charades-STA	VGG	GloVe	16	100	1e-4	10	5	10	2	0.2	0.6	10	1	10	0.4	0	4	1	1

Table 9: Implementation details. From left to right: bs denotes batch size,lr denotes learning rate,N_q denotes the number of Queries,N_n denotes the number of noise groups,N_s denotes the sample frames in Temporal Align,α denotes the hyperparameter of Gaussian mask generation,σ_p and σ_n denote the noise scale of positive and negative noise groups,w_q denotes the weight of matched spans in $L_{quality}$,δ_q denotes the margin in $L_{qmargin}$,λ_A, λ_S, λ_M and λ_Q denote the weight of L_A, L_S, L_M and L_Q separately.

A Appendix

A.1 Details of Datasets

QVHighlights. QVHighlights dataset was constructed to address both Moment Retrieval and Highlight Detection tasks simultaneously. It covers a range of content including daily activity vlogs and news reports. In this dataset, a single query may correspond to multiple moments, comprising 10148 videos, 10310 queries and their associated 18367 moments.

Charades-STA. Charades-STA was derived from the Charades(Sigurdsson et al., 2016) dataset, Charades-STA focuses on indoor activities, encompassing 6672 videos and 16124 video-query pairs.

TACoS. TACoS was built from the MPII Cooking Composite Activities dataset(Rohrbach et al., 2012), TACoS captures human activities in the kitchen, featuring 127 videos and 18818 video-query pairs.

A.2 More Implementation Details

To ensure stable convergence, we gradually decay the learning rate to 0 after 40 epochs for QVHighlights. For additional hyperparameter and loss settings, please refer to table 9.

A.3 More Ablation Studies

Ablation on Noise Group Nums. We investigate the impact of the number of noise groups in denoise learning on QVHighlights. As shown in Figure 6. When the number of noise groups is small, the model cannot obtain sufficient additional supervisory information. Conversely, when the number of noise groups is large, the persistent noise affects the model‘s convergence and disrupts its original learning path. Empirical evidence shows that the model performs optimally when the number of noise groups is set to 5.

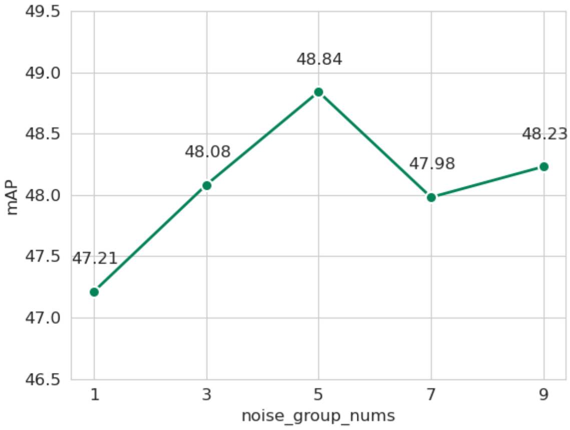

Figure 6: Ablation on noise group nums

L_{vtc}	L_{ctm}	MR			HD	
		R1 @0.5	R1 @0.7	mAP Avg.	mAP	HIT@1
		63.74	50.52	47.01	39.75	63.55
✓		65.68	50.77	47.19	40.03	64.39
	✓	65.29	**52.58**	47.77	39.97	63.42
✓	✓	**67.03**	52.52	**48.84**	**40.81**	**67.61**

Table 10: Ablation on Feature Align Loss

Method	R1@0.5	R1@0.7	mAP
QD-DETR	**62.52**	46.84	41.35
+dynamic anchor	62.06	47.42	42.09
+Gaussian mask	62.0	**48.13**	**42.84**

Table 11: Ablation on Component Generalizability.

Method	R1@0.5	R1@0.7	mAP
M-DETR	53.33 ± 1.4	34.16 ± 1.4	31.18 ± 1.1
QD-DETR	61.94 ± 0.4	47.02 ± 1.0	41.13 ± 0.5
SA-DETR	66.02 ± 0.8	51.72 ± 0.8	47.87 ± 0.6

Table 12: Performance Statistical Analysis.

Ablation on Feature Align Loss. As shown in Table 10, we investigated the impact of Feature Align Loss L_{vtc} and L_{ctm} on QVHighlights. The experiments show that L_{vtc} aligns the video with the text in global aware, significantly improving the performance of Highlight Detection. L_{ctm} aligns the clips and the text in local aware, improving the performance of Moment Retrieval. The combination of the two produces a boost effect.

Ablation on Component Generalizability. As shown in Table 11, we investigated the generalizability of our instance related span anchor and Gaussian mask modulate. We add them to QD-DETR and report the results on QVHighlights val split. These results demonstrate that our components can be effectively integrated into existing models and improve performance, confirming their generalizability.

Performance Statistical Analysis. we conducted experiments to verify the robustness and statistical significance of our results. Specifically, we repeat experiments on QVHighlights val set using seeds 0, 1, 2, 3 and 2018. The mean and standard deviation of Moment Retrieval metrics are shown in the table 12.