International Journal on Semantic Web and Information Systems
Volume 19 • Issue 1

Query-Guided Refinement and Dynamic Spans Network for Video Highlight Detection and Temporal Grounding in Online Information Systems

Yifang Xu, School of Electronic Science and Engineering, Nanjing University, China*
https://orcid.org/0009-0009-8332-4791

Yunzhuo Sun, School of Physics and Electronics, Hubei Normal University, China

Zien Xie, School of Electronic Science and Engineering, Nanjing University, China

Benxiang Zhai, School of Electronic Science and Engineering, Nanjing University, China

Youyao Jia, Gosuncn Chuanglian Technology Co., Ltd., Guangzhou, China

Sidan Du, School of Electronic Science and Engineering, Nanjing University, China

ABSTRACT

With the surge in online video content, finding highlights and key video segments have garnered widespread attention. Given a textual query, video highlight detection (HD) and temporal grounding (TG) aim to predict frame-wise saliency scores from a video while concurrently locating all relevant spans. Despite recent progress in DETR-based works, these methods crudely fuse different inputs in the encoder, which limits effective cross-modal interaction. To solve this challenge, the authors design QD-Net (query-guided refinement and dynamic spans network) tailored for HD&TG. Specifically, they propose a query-guided refinement module to decouple the feature encoding from the interaction process. Furthermore, they present a dynamic span decoder that leverages learnable 2D spans as decoder queries, which accelerates training convergence for TG. On QVHighlights dataset, the proposed QD-Net achieves 61.87 HD-HIT@1 and 61.88 TG-mAP@0.5, yielding a significant improvement of +1.88 and +8.05, respectively, compared to the state-of-the-art method.

KEYWORDS

Multi-Modal Learning, Transformer, Video Highlight Detection, Video Temporal Grounding

INTRODUCTION

The rapid advancement of artificial intelligence has significantly elevated video content creation technologies, resulting in tens of millions of new videos being uploaded to online platforms daily (Taleb & Abbas, 2022; Abbas et al., 2021). Given this vast volume of content, users urgently desire

DOI: 10.4018/IJSWIS.332768

*Corresponding Author

to see highlights or retrieve precise frames in a video that are most pertinent to a given textual query, allowing them to quickly skip to relevant video segments (Hamza et al., 2022; Sahoo & Gupta, 2021). In this paper, we focus on two video understanding tasks: highlight detection (HD) and temporal grounding (TG), as depicted in Fig. 1. Given a video paired with its corresponding natural language query, the objective of HD is to predict highlights for each video clip (Y. Liu et al., 2022). TG aims to retrieve all spans in a video that are most relevant to the query, where each span consists of a start and end clip (Gao et al., 2017). Since the goal of both tasks is to find the most appropriate clip, recent work (Lei et al., 2021) proposes the QVHighlights dataset to conduct HD and TG concurrently.

The primary challenge of the HD&TG task lies in effectively generating cross-modal features that contain query-related information, since such features are utilized to predict highlights and locate the query-matched spans. Inspired by DETR (Carion et al., 2020), Moment-DETR (Lei et al., 2021) designed a transformer encoder-decoder pipeline to tackle this challenge, as shown in Fig. 2 (a). However, Moment-DETR opts to directly concatenate video and text for coarse fusion in the encoder. This approach mixes intra-modal contextual modeling with cross-modal feature interaction. When the similarity between video frames far surpasses the video-query similarity, the resulting cross-modal features are irrelevant to the query, leading to diminished performance. Moreover, tasks like object detection (OD) and TG both necessitate decoder-based localization. Recent DETR-based research (S. Liu et al., 2022) indicates that utilizing dynamic bounding box anchors as queries within the decoder helps alleviate the problem of slow convergence in OD training. Yet, Moment-DETR solely

Figure 1. A depiction of HD&TG. Given a video paired with its corresponding textual query, the goal of HD&TG is to predict frame-wise saliency scores and locate all the most relevant spans simultaneously

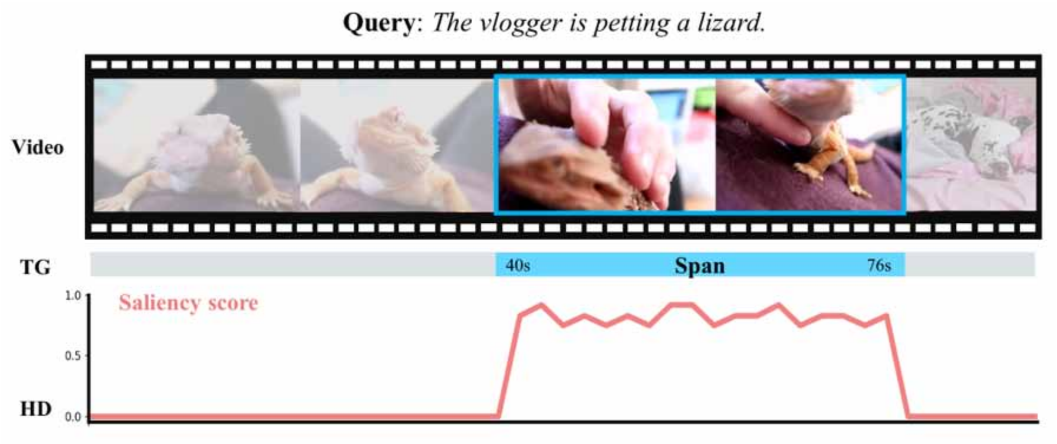
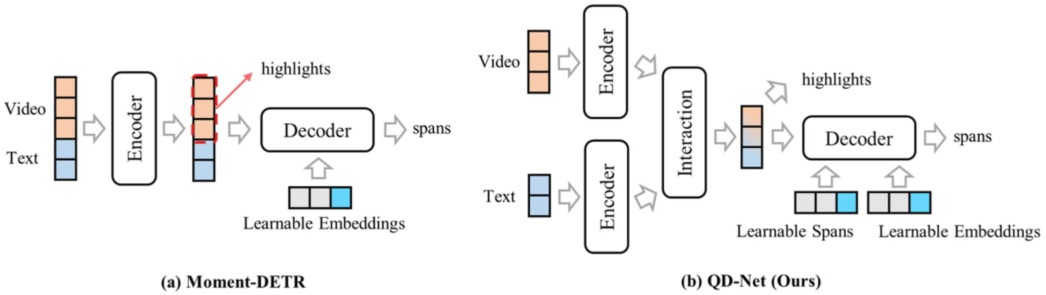


Figure 2. Comparison between Moment-DETR (a) and QD-Net (b)



employs learnable embeddings in the decoder and lacks adequate temporal span modeling, which hinders convergence speed and accuracy for a given TG task.

In this paper, we newly propose a HD&TG model named QD-Net (Query-guided refinement and Dynamic spans Network) to tackle the above issues. As shown in Figure 2(b), QD-Net decouples the feature encoding and interaction processes using a query-guided refinement module. This module fuses video and text tokens, which produce query-relevant cross-modal features. To capture intra-modal context from the global perspective, we introduce the straightforward yet efficient PoolFormer (Yu et al., 2022), which is applied to both visual and text encoders. In addition, we design a span decoder, which can more explicitly associate learnable embeddings with predicted span positions and speed up training convergence for the TG task. Specifically, the decoder contains learnable 2D spans that are dynamically updated at each layer, and their size can modulate the cross-attention weights within the decoder. To demonstrate the superiority of QD-Net, we execute comprehensive experiments and ablations on three publicly accessible datasets (QVHighlights, TVSum, and Charades-STA). The results reveal that QD-Net outperforms current state-of-the-art (SOTA) approaches. Notably, on the QVHighlights dataset, our model scores 61.87 HD-HIT@1 and 61.88 TG-mAP@0.5, showing gains of +1.88 and +8.05 over the SOTA method. In summary, our principal contributions include:

(1) We propose a QD-Net tailored for HD&TG tasks. We design a query-guided refinement module to generate query-relevant cross-modal features and decouple the feature encoding from the interaction process. We introduce a simplified pooling mechanism in the encoder to model the global information within a single modality.
(2) We propose a span decoder to dynamically associate learnable embeddings with span information and ensure faster training convergence for TG.
(3) Extensive experimental results indicate that our approach achieves state-of-the-art performance on three publicly available datasets.

The remainder of this paper is structured as follows: The following section reviews related works in HD&TG. Next, we provide a detailed description of our proposed method, including the uni-modal encoder, query-guided refinement module, and span decoder. Experimental results and ablation analysis are then presented. Finally, we conclude the paper and outline future directions.

RELATED WORKS

Along with the development of artificial intelligence, an increasing number of researchers are leveraging deep learning to address challenges in areas such as big data (Stergiou et al., 2021; Galiautdinov, 2021), health diagnosis (Shankar et al., 2021; Anil et al., 2022), fake information detection (Li et al., 2022; Tembhurne et al., 2022), and video understanding (Zhang et al., 2022; Xu et al., 2023). Video highlight detection (HD) is designed to output saliency scores for each video clip. Trailer (Wang et al., 2020) and sLSTM (Zhang et al., 2016) formulate HD as a ranking problem, training a network to rank highlight moments higher than non-highlight moments. SL-Module (Tang et al., 2022) employs an attention-based model to discern multiple moments contributing to the desired moment. Since early HD datasets are without textual queries, existing multi-modal learning methods in HD primarily process visual and audio cues. Joint-VA (Badamdorj et al., 2021) utilizes the noise sentinel method and combines audio and visual information in a bi-attention module.

Video temporal grounding (TG) involves locating all pertinent video spans in response to a textual prompt. DRFT (Chen et al., 2021) integrates multi-modal features such as depth (Xu et al., 2021a; Xu et al., 2021b) and optical flow to learn complementary visual sources, but without a decoder it results in low accuracy of TG. 3D features (Chu et al., 2022; Srivastava et al., 2022) in the video also improve multi-modal representation, and GTR (Cao et al., 2021) utilizes a cubic embedding extractor to capture 3D features in videos. Yet 3D feature extracting is time-consuming. Imbalanced annotations may introduce data bias (Hasib et al., 2021; Hammad et al., 2021), and therefore CMA (B. Zhang et al., 2020) employs a cross-modality network to rectify this bias.

Moment-DETR (Lei et al., 2021) proposes QVHighlights, a unique dataset specifically designed to support query-based HD and TG. Moment-DETR also provides an encoder-decoder model. To combine different input sources (text, audio, and video), UMT (Y. Liu et al., 2022) proposes a more unified transformer-based model for HD&TG. However, by removing the crucial decoder and Hungarian matching present in Moment-DETR, UMT compromises the performance on the TG task.

Uni-modal encoding is intended to capture contextual information from single-modal features (Zhang et al., 2022). The attention mechanism in the transformer (Vaswani et al., 2017) is adept at modeling long-range dependencies, which leads to some researchers (Lin et al., 2023) emphasizing its importance as the encoder. However, recent work (Yu et al., 2022) indicates that the effectiveness of transformers largely hinges on their token mixer blocks and MLP. Moreover, utilizing a straightforward pooling mechanism in token mixers has proven fruitful (Yu et al., 2022). In addition, Sparse-MLP (Tang et al., 2022) also achieves satisfying results using only MLPs. Inspired by these works, we introduce PoolFormer (Yu et al., 2022) during the encoding stage.

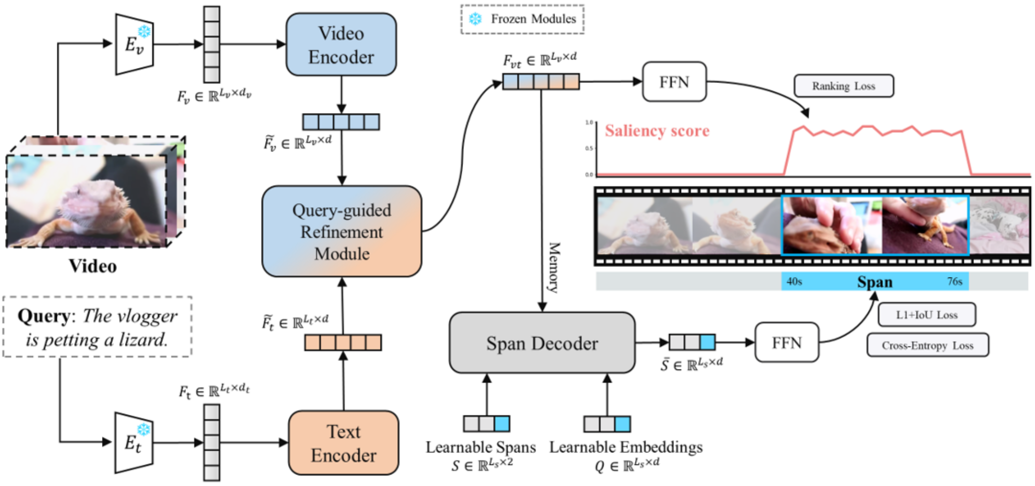
For a pair of video and textual queries, visual features F_v and textual features F_t are derived using the pre-trained visual extractor E_v and textual extractor E_t. Next, the uni-modal encoder processes these features to model global information. By leveraging the query-guided refinement module, we fuse features from various modalities to derive cross-modal features F_{vt}. Subsequently, we use the span decoder containing learnable spans S with learnable embeddings Q to get span features $\bar{S}$. Ultimately, the prediction module yields HD&TG outcomes, with optimization informed by the depicted loss function.

METHOD

Overview

HD&TG (video highlight detection and temporal grounding) can be defined in this manner: for an untrimmed video $V \in R^{L_v \times H \times W \times 3}$ containing L_v frames and a corresponding text query $T \in \mathbb{R}^{L_t}$ with L_t words, HD&TG aims to compute highlights (saliency scores) $H \in \mathbb{R}^{L_v}$ for each frame and

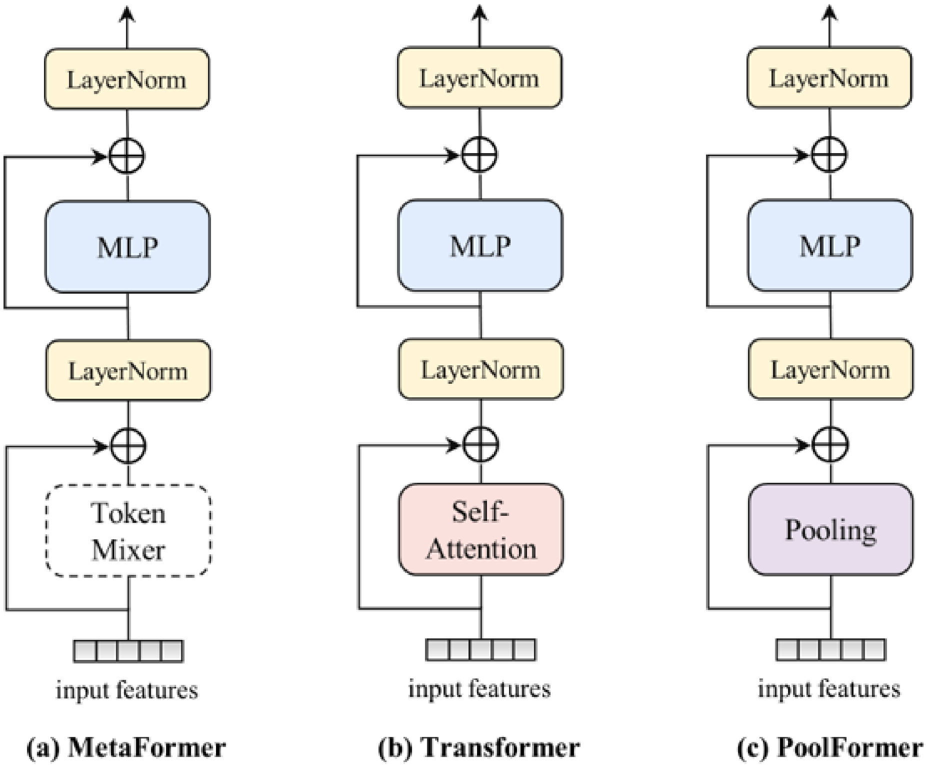
Figure 3. An overview of our proposed QD-Net framework


E_v
$F_v \in \mathbb{R}^{L_v \times d_v}$
Video Encoder
$\tilde{F}_v \in \mathbb{R}^{L_v \times d}$
Query-guided Refinement Module
Frozen Modules
$F_{vt} \in \mathbb{R}^{L_v \times d}$
FFN
Ranking Loss
Saliency score
1.0
0.5
0.0
Memory
Video
Query: The vlogger is petting a lizard.
$\tilde{F}_t \in \mathbb{R}^{L_t \times d}$
$F_t \in \mathbb{R}^{L_t \times d_t}$
E_t
Text Encoder
Span Decoder
$\bar{S} \in \mathbb{R}^{L_s \times d}$
FFN
40s
Span
76s
L1+IoU Loss
Cross-Entropy Loss
Learnable Spans
$S \in \mathbb{R}^{L_s \times 2}$
Learnable Embeddings
$Q \in \mathbb{R}^{L_s \times d}$

concurrently locate video spans $\tilde{S} \in \mathbb{R}^{L_s \times 2}$, which are highly relevant to T, where each span comprises a start frame and an end frame.

Figure 3 illustrates our proposed model QD-Net, comprising five main parts: feature extractor, uni-modal encoder, query-guide refinement module, span decoder, and prediction heads. We first utilize the semantic web and frozen pre-trained model (detailed below, under "Experimental Settings") to obtain video features $F_v \in \mathbb{R}^{L_v \times d_v}$ and query features $F_t \in \mathbb{R}^{L_t \times d_t}$. Here, the visual and textual extractors are abbreviated as E_v and E_t, respectively. Next, we utilize separate 2-layer MLP (multi-layer perceptron) complemented by layer normalization (Ba et al., 2016) to map both the visual and textual tokens into a common embedding space with dimensionality d. To capture intra-model correlations under the global perspective, the contextual video features $\tilde{F}_v \in \mathbb{R}^{L_v \times d}$ and text features $\tilde{F}_t \in \mathbb{R}^{L_t \times d}$ are derived using the uni-modal encoder. The query-guided refinement component combines these features to obtain query-relevant cross-modal features $F_{vt} \in \mathbb{R}^{L_v \times d}$, which are strictly aligned with the visual features in the temporal length. Subsequently, we use the span decoder and learnable spans $S \in \mathbb{R}^{L_s \times 2}$ with learnable embeddings $Q \in \mathbb{R}^{L_s \times d}$ to get span features $\bar{S} \in \mathbb{R}^{L_s \times d}$. Finally, the naïve prediction heads are employed to estimate highlight scores $H \in \mathbb{R}^{L_v}$ and spans $\tilde{S} \in \mathbb{R}^{L_s \times 2}$.

Figure 4. (a) MetaFormer is a general architecture that does not specify a token mixer. (b) By integrating attention into token mixing, MetaFormer manifests as a Transformer. (c) In PoolFormer, a straightforward pooling mechanism is employed for fundamental token mixing



(a) MetaFormer (b) Transformer (c) PoolFormer

Uni-Modal Encoder

For the TG task, the 1D sliding-window strategy is employed in prior work (Hendricks et al., 2017) to pre-select video proposals. This strategy results in increased computational cost and diminished efficiency due to the necessity of densely overlapping sampling required for optimal accuracy. Additionally, a notable drawback is its tendency to seize on local temporal details while overlooking the temporally global information. For tasks related to video understanding, a thorough comprehension of the entirety of a video's content is crucial for improved performance.

Considering the superiority of the transformer (Vaswani et al., 2017) in capturing long-range dependencies, studies (Lin et al., 2023) underscore the significance of attention techniques, directing their efforts towards crafting diverse attention-driven token mixers during encoding processes. However, a recent method (Yu et al., 2022) suggests that the primary drivers of a transformer's success stem from the token mixing module and MLP, as depicted in Figure 4 (b). And Yu et al. (2022) propose a general architecture, MetaFormer-like transformer that does not specify a token mixer, as illustrated in Figure 4 (a). PoolFormer (Yu et al., 2022) harnesses a simple pooling mechanism as its token mixer, excelling in computer vision tasks compared to the traditional approach based on a transformer. As represented in Figure 4 (c), this non-parametric mechanism allows tokens to uniformly assimilate the information from nearby tokens, thus modelling the global intra-modal context. Through our experiments, we have been surprised to find that this simplified module achieves reasonable gain, as demonstrated in Tab. 1. Consequently, our uni-modal encoder is derived from PoolFormer, which incorporates a pooling mechanism paired with MLP. In addition, each encoder layer is equipped with layernorm (Ba et al., 2016) and residual block (He et al., 2016). To obtain the contextualized $\tilde{F}_v \in \mathbb{R}^{L_v \times d}$ and textual features $\tilde{F}_t \in \mathbb{R}^{L_t \times d}$, the encoding process is:

$$\bar{F}_x = \mathrm{Norm}\left(F_x + \mathrm{Pool}\left(F_x\right)\right) \tag{1}$$

$$\tilde{F}_x = \mathrm{Norm}\left(\bar{F}_x + \mathrm{MLP}\left(\bar{F}_x\right)\right) \tag{2}$$

where $F_x \in \left\{F_v, F_t\right\}$ and $\mathrm{Norm}\left(\cdot\right)$ represents layernorm.

Query-Guided Refinement Module

In the encoding stage of Moment-DETR (Lei et al., 2021), video and text are directly concatenated and then coarsely fused. However, if the similarity between video clips substantially exceeds the video-

Table 1. Performance comparison of QD-Net using Transformer or PoolFormer as uni-modal encoder on the QVHighlights validation set

Uni-Modal Encoder	Layers	Temporal Grounding					Highlight Detection	
		R1		mAP			$\geq$ Very Good	
		@0.5	@0.7	@0.5	@0.75	Avg.	mAP	HIT@1
Transformer	1	**62.98**	42.10	**61.79**	**40.47**	**39.97**	**38.10**	**61.91**
	2	61.76	**43.92**	62.98	40.26	38.63	38.16	61.48
	3	60.72	41.24	60.56	39.74	38.10	37.56	60.62
PoolFormer	0	60.13	41.65	60.93	39.77	38.92	38.26	60.45
	1	61.89	43.55	62.12	41.79	40.87	38.05	61.78
	2	**62.32**	**45.61**	**63.15**	**42.05**	**41.46**	**38.56**	**62.06**
	3	60.61	42.59	62.14	40.74	40.20	37.93	61.31

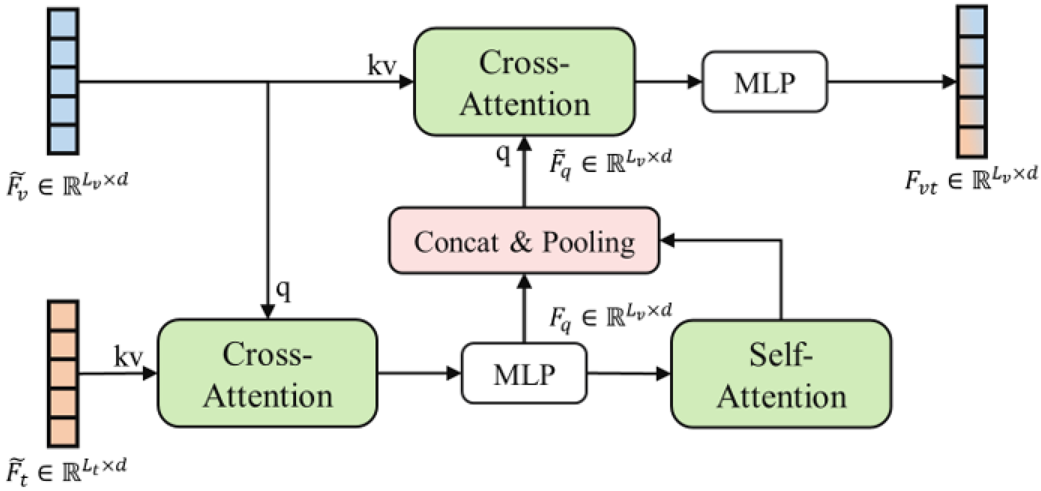
query similarity, the resulting cross-modal features become unrelated to the query, compromising overall performance. Therefore, we design a plug-and-play query-guided refinement module to decouple the feature encoding and interaction processes. This module fuses contextual features from different modalities and generates query-relevant cross-modal features by emphasizing the segments of visual clips most related to the textual words.

The framework of the query-guided refinement module is illustrated in Fig. 5. In this structure, the first cross-attention layer and MLP dynamically produce query features $F_q \in \mathbb{R}^{L_v \times d}$, which are derived from contextual text features $\tilde{F}_t \in \mathbb{R}^{L_t \times d}$. Here, $\tilde{F}_v$ serves as *query* of cross-attention, while $\tilde{F}_t$ is *key* and *value*. The cross-attention weights assess the relational significance between video clips and textual tokens, which allows every video clip to discern which textual concepts correspond to it. Subsequently, a self-attention mechanism refines query features F_q, and we concatenate these features and feed them into pooling to obtain refined query features $\tilde{F}_q \in \mathbb{R}^{L_v \times d}$. The above process can be summarized as:

$$\tilde{F}_q = \mathrm{Pool}\left(\mathrm{cat}\left[F_q, \quad \mathrm{SA}\left(F_q\right)\right]\right) \tag{3}$$

where SA means self-attention and cat is an abbreviation of concatenation. Finally, we use another cross-attention and MLP to obtain query-relevant cross-modal features $F_{vt} \in \mathbb{R}^{L_v \times d}$, where $\tilde{F}_q$ serves as *query*. Kindly observe that in Figure 5, we omit residual blocks and layernorm, yet they are incorporated across all layers. We also add learnable position embeddings (Vaswani et al., 2017) at the onset of every attention stratum. The cross-modal features F_{vt} also represent a fusion of span and highlight information. They are subsequently sent to prediction heads and the span decoder, which respectively output saliency scores $H \in \mathbb{R}^{L_v}$ and span features $\bar{S} \in \mathbb{R}^{L_s \times d}$.

Figure 5. The overview of the query-guided refinement module


kv
Cross-Attention
MLP
q
$\tilde{F}_q \in \mathbb{R}^{L_v \times d}$
$\tilde{F}_v \in \mathbb{R}^{L_v \times d}$
$F_{vt} \in \mathbb{R}^{L_v \times d}$
Concat & Pooling
q
$F_q \in \mathbb{R}^{L_v \times d}$
kv
Cross-Attention
MLP
Self-Attention
$\tilde{F}_t \in \mathbb{R}^{L_t \times d}$

Span Decoder

DETR (Carion et al., 2020) represents a fundamental transformer encoder-decoder architecture in object detection (OD) tasks. Inspired by DETR, Moment-DETR (Lei et al., 2021) introduced a DETR-based model explicitly tailored for TG. This is due to the striking similarity between TG and OD, where the outputs are predicted 2D and 4D boxes, respectively. UMT (Y. Liu et al., 2022) found that the decoder is the cause of slow convergence, thus it was subsequently removed to expedite training. This approach aligns with the conclusions drawn in TSP (Z. Sun et al., 2021). However, DAB-DETR (S. Liu et al., 2022) identified that the multi-modal property of queries could be the root cause of slow training. To address this, DAB-DETR proposed an explicit prior position, termed the dynamic anchor box, to compel each query to focus on a specific area. This method enabled faster training convergence and higher detection accuracy in OD.

Drawing inspiration from the above works, we design a dynamic span decoder with learnable spans tailored for TG, as shown in Figure 6. The span decoder contains N_s layers. Each layer of the span decoder comprises a simple self-attention layer and the width-modulated cross-attention layer, which are utilized for query updates and feature probing, respectively. 2D learnable spans (dynamic spans) represent the most relevant span location in cross-modal features $F_{vt} \in \mathbb{R}^{L_v \times d}$ and are updated in each decoder layer. This design explicitly associates learnable embeddings (high-dimensional span information) $Q \in \mathbb{R}^{L_s \times d}$ with cross-modal features F_{vt}. For ease of description and computation, we denote the beginning and end clips of a span as the central position and width of the span, respectively.

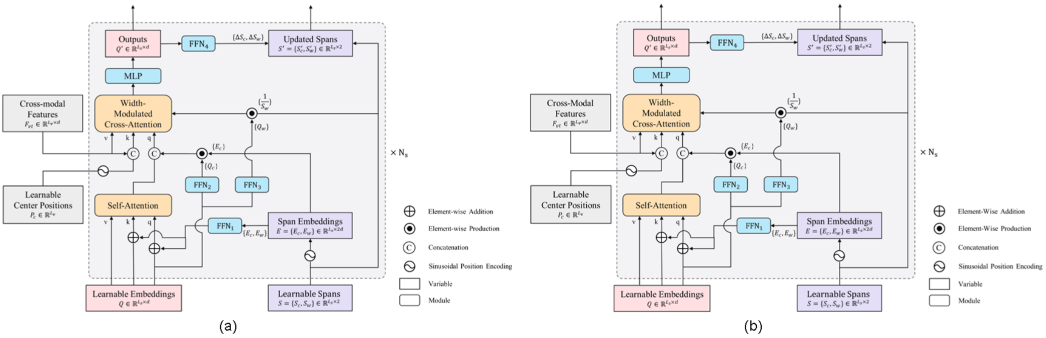
We define dynamic spans as $S = \left\{S_c, S_w\right\} \in \mathbb{R}^{L_s \times 2}$, where $S_c \in \mathbb{R}^{L_s}$ is the center of spans and $S_w \in \mathbb{R}^{L_s}$ is the width of spans. PE represents sinusoidal position encoding (Vaswani et al., 2017) to generate position embeddings, which projects input to the embedding space of size d. PE shares parameters in all span decoder layers. Since learnable spans S is a binary number, we overload the PE operator here:

$$\mathrm{PE}\left(S\right) = \mathrm{PE}\left(S_c, S_w\right) = \mathrm{cat}\left[\mathrm{PE}\left(S_c\right),\ \mathrm{PE}\left(S_w\right)\right] \tag{4}$$

In the self-attention layer, we formulate *query* q_s, *key* k_s and *value* v_s as follows:

$$q_s = Q + \mathrm{FFN}_1\left(\mathrm{E}\right) \tag{5}$$
$$k_s = Q + \mathrm{FFN}_1\left(\mathrm{E}\right),\quad v_s = Q \tag{6}$$

Figure 6. The framework of the span decoder


Outputs
$Q' \in \mathbb{R}^{L_s \times d}$
FFN$_4$
$\{\Delta S_c, \Delta S_w\}$
Updated Spans
$S' = \{S'_c, S'_w\} \in \mathbb{R}^{L_s \times 2}$
MLP
Cross-modal Features
$F_{vt} \in \mathbb{R}^{L_v \times d}$
Width-Modulated Cross-Attention
$\{\frac{1}{S_w}\}$
$\{Q_w\}$
v
k
q
$\{E_c\}$
$\{Q_c\}$
× N$_s$
Learnable Center Positions
$P_c \in \mathbb{R}^{L_v}$
FFN$_2$
FFN$_3$
Self-Attention
FFN$_1$
$\{E_c, E_w\}$
Span Embeddings
$E = \{E_c, E_w\} \in \mathbb{R}^{L_s \times 2d}$
Learnable Embeddings
$Q \in \mathbb{R}^{L_s \times d}$
Learnable Spans
$S = \{S_c, S_w\} \in \mathbb{R}^{L_s \times 2}$
Element-wise Addition
Element-wise Production
Concatenation
Sinusoidal Position Encoding
Variable
Module
(a)
Cross-Modal Features
Element-Wise Addition
Element-Wise Production
(b)

where FFN_1 maps input to d dimensions, $\mathrm{FFN}_1 : \mathbb{R}^{L_s \times 2d} \rightarrow \mathbb{R}^{L_s \times d}$. Span embeddings . ..

Following DAB-DETR, we employ FFN_2 to derive scaled features $Q_c \in \mathbb{R}^{L_s \times d}$ that are dependent on the center information. These scaled features subsequently undergo element-wise production with the span embeddings $E_c = \mathrm{PE}\left(S_c\right) \in \mathbb{R}^{L_s \times d}$, which leads to an effective rescaling of these embeddings. In the width-modulated cross-attention layer, we define *query* q_c, *key* k_c and *value* v_c as follows:

$$q_c = \mathrm{cat}\left[Q, E_c \cdot Q_c\right] \tag{7}$$

$$k_c = \mathrm{cat}\left[F_{vt}, \mathrm{PE}\left(P_c\right)\right], v_c = F_{vt} \tag{8}$$

where $Q_c = \mathrm{FFN}_2\left(Q\right)$, $\mathrm{FFN}_2 : \mathbb{R}^{L_s \times d} \rightarrow \mathbb{R}^{L_s \times d}$. $\cdot$ is the element-wise production. $\mathrm{PE}\left(P_c\right) \in \mathbb{R}^{L_v \times d}$ denotes center position embeddings for cross-modal features F_{vt}. The position embeddings for *query* and *key* are produced using 1D center coordinates.

To improve the span position prior, width information is incorporated into the cross-attention map, leading to the computation of width-modulated cross-attention (WMCA) as follows:

$$\mathrm{WMCA} = \mathrm{Softmax}\left(\frac{q_c \cdot k_c^T}{\sqrt{d}} \cdot \frac{Q_w}{S_w}\right) \cdot v_c \tag{9}$$

In this equation, $\sqrt{d}$ represents the scaling factor (Vaswani et al., 2017) in the softmax function. $S_w \in \mathbb{R}^{L_s}$ denotes the width of learnable spans S. $Q_w = \mathrm{FFN}_3\left(Q\right) \in \mathbb{R}^{L_s}$ indicates the reference width derived from learnable embeddings Q, FFN_3 : $\mathbb{R}^{L_s \times d} \rightarrow \mathbb{R}^{L_s}$. For additional details, refer to Fig. 6.

Finally, relative span positions $\left\{\Delta S_c, \Delta S_w\right\} \in \mathbb{R}^{L_s \times 2}$, derived from the outputs $Q' \in \mathbb{R}^{L_s \times d}$, are utilized to dynamically update the spans:

$$S' = \left\{S_c^{'}, S_w^{'}\right\} = \left\{S_c + \Delta S_c, S_w + \Delta S_w\right\} \tag{10}$$

In this formula, $S' \in \mathbb{R}^{L_s \times 2}$ represents the updated spans, and $\left\{\Delta S_c, \Delta S_w\right\} = \mathrm{FFN}_4\left(\ Q'\right)$, $\mathrm{FFN}_4 : \mathbb{R}^{L_s \times d} \rightarrow \mathbb{R}^{L_s \times 2}$. We denote the outputs from the final layer of the decoder as $\bar{S} \in \mathbb{R}^{L_s \times d}$, which represent the span features. These features are subsequently fed into the prediction heads to generate the results for TG.

Prediction Heads and Loss Function

For query-relevant cross-modal features $F_{vt} \in \mathbb{R}^{L_v \times d}$ contain joint span and highlight details, we apply 1-layer FFN to estimate the highlight scores $H \in \mathbb{R}^{L_v}$. With regards to the span features $\bar{S} \in \mathbb{R}^{L_s \times d}$, we utilize 2-layer FFN and a sigmoid function to obtain the normalized center and width of spans $\tilde{S} \in \mathbb{R}^{L_s \times 2}$. Additionally, another 1-layer FFN with a softmax is used to derive span class labels. In the TG task, predicted spans that align with the ground truth (GT) are assigned a *foreground* label; others are set as *background*. To facilitate a more accurate comparison with the baseline model, Moment-DETR, we adopt its training loss and loss hyperparameters λ_*.

In the calculation of the saliency loss $\mathcal{L}_h$, a ranking loss is applied to distinguish the relative significance between two specific pairs of moments, with an emphasis on challenging moments. One pair contrasts the moment with the top score moment s_{high} and the one with the lowest score s_{low} among the ground truth moments. Another pair contrasts moments inside and outside within the ground truth, and we denote their scores as s_{in} and s_{out}. Letting the margin be denoted by m, $\mathcal{L}_h$ can be formulated as:

$$\mathcal{L}_h = \max\left(0,\ m + s_{low} - s_{high}\right) + \max\left(0,\ m + s_{out} - s_{in}\right) \tag{11}$$

Building upon Moment-DETR, we adopt the Hungarian method to determine the best bipartite pairing between the predicted moments and their corresponding ground truths. Given that there might not always be a one-to-one mapping between the predicted and actual moments. We assume that L_n represents the pairs of matched predictions and ground truth within a video. We employ the span loss $\mathcal{L}_{\mathrm{s}}$ to evaluate the differences between the estimated span $\tilde{s}$ and the actual span s. This span loss integrates both L1 loss λ_{L1} and IoU (Rezatofighi et al., 2019) loss $\mathcal{L}_{IoU}$:

$$\mathcal{L}_{\mathrm{s}} = \sum_{i=1}^{L_n}\left[\lambda_{L1} \parallel \tilde{s} - s \parallel_1 + \lambda_{IoU}\mathcal{L}_{IoU}\left(\tilde{s}, s\right)\right] \tag{12}$$

Furthermore, we utilize the cross-entropy metric, denoted as $\mathcal{L}_{\mathrm{cls}}$, to categorize the estimated spans into *foreground* or *background* categories. This is formulated as:

$$\mathcal{L}_{\mathrm{cls}} = -\sum_{i=1}^{L_s}\left[w_p z_i \log\left(p_i\right) + \left(1 - z_i\right)\log\left(1 - p_i\right)\right] \tag{13}$$

In the expression, p_i denotes the forecasted likelihood of the *foreground*, while z_i represents its label. To counterbalance label disproportion, the *foreground* label receives an elevated weight w_p. The total loss is represented as $\mathcal{L}_{\mathrm{total}}$:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_h + \mathcal{L}_{\mathrm{s}} + \lambda_{cls}\mathcal{L}_{\mathrm{cls}} \tag{14}$$

EXPERIMENTS

Datasets

To demonstrate the robustness and superior performance of our proposed QD-Net, we carry out comprehensive experiments on three benchmark datasets: QVHighlights (Lei et al., 2021), TVSum (Song et al., 2015), and Charades-STA (Gao et al., 2017).

QVHighlights

QVHighlights is currently the only dataset that supports both temporal grounding (TG) and highlight detection (HD) tasks. This dataset contains a diverse set of 10,148 videos from online database

YouTube, each with a maximum length of 150 seconds. It comprises about 10,000 annotations, which include a free-text query, video spans (averaging about 1.8 spans for each query), and highlight scores on a per-clip basis. A distinctive feature of QVHighlights is its provision of a just benchmarking system, where evaluations can be obtained solely by submitting predictions for the testing split to the QVHighlights online server. In terms of data division, we adhere to the original QVHighlights splits, allocating 0.7 for training, 0.15 for validation, and 0.15 for testing.

TVSum

TVSum is a notable dataset tailored for HD tasks, which encompasses videos from 10 distinct categories, each containing 5 videos. Following the setting of UMT (Y. Liu et al., 2022), we divide the dataset, designating 80% for training purposes and the residual 20% for evaluation.

Charades-STA

Charades-STA is utilized as a standard dataset for the TG tasks. Originating from the initial Charades (Sigurdsson et al., 2016) dataset, Charades-STA features approximately 9,800 videos of everyday indoor actions and close to 16,000 annotations. Traditionally, the dataset is divided into 12,300 annotations for training and 3,700 for testing.

Evaluation Metrics

For the QVHighlights dataset, our TG evaluation metrics include Recall@1 at IoU of 0.5 and 0.7, mAP at IoU of 0.5 and 0.75, and the average mAP at IoU between 0.5 and 0.95 (incremented in steps of 0.05). When assessing HD, we utilize mAP and HIT@1, with the latter measuring the hit ratio for the highest-scored moment. For the TVSum dataset, we opt to use the top-5 mAP metric. Meanwhile, we adopt Recall@1 at 0.5 and 0.7 IoU settings for Charades-STA.

Experimental Settings

For the QVHighlights dataset, we follow Moment-DETR and apply semantic web technologies (Narayanasamy et al., 2022) to aid annotation generation. Then, we use the image-encoder of CLIP (Radford et al., 2021) and SlowFast (Feichtenhofer et al., 2019) to extract features, and then concatenate to generate visual features $F_v \in \mathbb{R}^{L_v \times 2816}$. We employ the textual encoder of CLIP for extracting query

Table 2. Comparison of results on the QVHighlights *testing* set. Each model only employs video and text features. M-DETR is the abbreviation of Moment-DETR. Top-performing results are bold

Methods	Temporal Grounding					Highlight Detection	
	R1		mAP			$\geq$ Very Good	
	@0.5	@0.7	@0.5	@0.75	Avg.	mAP	HIT@1
BeautyThumb (Song et al., 2016)	-	-	-	-	-	14.36	20.88
DVSE (Liu et al., 2015)	-	-	-	-	-	18.75	21.79
MCN (Hendricks et al., 2017)	11.41	2.72	24.94	8.22	10.67	-	-
CAL (Escorcia et al., 2019)	25.49	11.54	23.40	7.65	9.89	-	-
XML (Lei et al., 2020)	41.83	30.35	44.36	31.73	32.14	34.49	55.25
XML+ (Lei et al., 2020)	46.69	33.46	47.89	34.67	34.90	35.38	55.06
M-DETR (Lei et al., 2021)	52.89	33.02	54.82	29.40	30.73	35.69	55.60
UMT (Y. Liu et al., 2022)	56.23	41.18	53.83	37.01	36.12	38.18	59.99
QD-Net (Ours)	**61.71**	**44.76**	**61.88**	**39.84**	**40.34**	**38.78**	**61.87**

features $F_t \in \mathbb{R}^{L_t \times 512}$. As for TVSum, we adopt the approach of UMT (Y. Liu et al., 2022), utilizing I3D video features $F_v \in \mathbb{R}^{L_v \times 1024}$ that are pre-trained on Kinetics-400 (Kay et al., 2017). Additionally, textual embeddings $F_t \in \mathbb{R}^{L_t \times 512}$ are derived from video titles via CLIP. For Charades-STA, the standard VGG (Simonyan & Zisserman, 2014) features $F_v \in \mathbb{R}^{L_v \times 4096}$ are used, alongside GloVe text features $F_t \in \mathbb{R}^{L_t \times 512}$. For all the input video and text, we configure their maximum length, L_v / L_t, as 75/32 for QVHighlights, 100/10 for TVSum, and 120/10 for Charades-STA.

In our proposed model, we utilize a 2-layer uni-modal encoder adopting a pooling size of 3 and a stride of 2. The query-guided refinement module and span decoder N_s are respectively set as 1 and 4 layers. We configure the attention dimensionality as $d = 256$ and the number of multi-heads as 8. The number of dynamic spans L_s is set to 30. In addition, our model employs the post-style layernorm (Ba et al., 2016) and 0.1 dropout rate. The loss coefficients are denoted as: $\lambda_{L1} = 10$, $\lambda_{IoU} = 2$, $\lambda_{cls} = 4$, $m = 0.2$, $w_p = 10$. We train our model using 2 NVIDIA 3090 GPUs. For QVHighlights, TVSum, and Charades-STA datasets, the batch size and epochs are 64/4/16 and 200/1000/100, respectively. We employ AdamW (Loshchilov & Hutter, 2017) optimizer with a learning rate of 1e-4 and a weight decay of 1e-4.

Experimental Results

Firstly, we report a comparison between our proposed QD-Net and prior works on the QVHighlights *testing* set, as shown in Tab. 2. To ensure a fair comparison, all methods exclusively input video and textual queries. The results indicate that our method surpasses the state-of-the-art method UMT by 5.48%, 8.05%, and 1.88% in terms of TG-R1@0.5, TG-mAP@0.5, and HD-HIT@1, respectively. Tab. 3 contrasts the performance of our model against existing models on the QVHighlights *validation* set. Significantly, our model also achieves the top performance.

Table 3. Results of different models on QVHighlights *validation* set

Methods	Temporal Grounding					Highlight Detection	
	R1		mAP			$\geq$ Very Good	
	@0.5	@0.7	@0.5	@0.75	Avg.	mAP	HIT@1
M-DETR (Lei et al., 2021)	53.94	34.84	-	-	32.20	35.65	55.55
UMT (Y. Liu et al., 2022)	-	-	-	-	37.79	**38.97**	59.99
QD-Net (Ours)	**62.32**	**45.61**	**63.15**	**42.05**	**41.46**	38.56	**62.06**

Table 4. Quantitative results on TVSum dataset. Comparison against the state-of-the-art highlight detection methods

Methods	MS	VT	PK	BK	VU	FM	GA	PR	DS	BT	Avg.
sLSTM (Zhang et al., 2016)	47.7	41.1	44.8	40.6	46.2	45.2	46.3	46.1	45.5	47.1	45.1
Trailer (Wang et al., 2020)	60.8	61.3	59.1	64.7	54.6	58.2	65.7	70.1	68.1	65.6	62.8
SL-Module (Liu et al., 2015)	86.2	86.5	79.0	72.6	68.7	58.9	74.9	63.2	64.0	78.9	73.3
Joint-VA (Badamdorj et al., 2021)	86.1	83.7	80.1	73.0	57.3	70.0	78.5	69.2	67.5	**97.4**	76.3
UMT (Y. Liu et al., 2022)	78.8	**87.5**	81.4	**86.9**	81.5	76.0	88.2	**87.0**	**79.6**	84.4	83.1
QD-Net (Ours)	**87.5**	86.1	**82.6**	84.5	**82.4**	**79.2**	**89.7**	84.3	76.9	85.6	**84.0**

Furthermore, we present a visualization of the outputs on QVHighlights, as illustrated in Fig. 8. The illustration displays the input query and corresponding video in a top-down sequence, accompanied by the forecasted spans and moment-wise highlight scores. Fig. 8 (a) and (b) demonstrate that our model is capable of effectively predicting single or multiple spans, as well as proficiently predicting highlights.

Tab. 4 exhibits the quantitative results on the TVSum dataset for highlight detection. Here, QD-Net outperforms most other methods across most categories. To be concrete, although our model trails UMT marginally in a few categories, it surpasses UMT by 0.9% in average top-5 mAP across all categories. Moreover, QD-Net outperforms UMT by 8.7% and 3.2% in the MS (Making Sandwich) and FM (Flash Mob gathering) categories, respectively.

In Tab. 5, we show the performance evaluation of QD-Net against other methods on the Charades-STA *testing* set. Our method exhibits superior performance across all metrics. Specifically, regarding the R1@0.5 and R1@0.7 metrics, QD-Net outperforms the existing SOTA method PEARL (Zhang & Radke, 2022) by margins of 3.93% and 0.99%, respectively.

Ablation Studies

To assess the impact of individual components in our introduced framework, we perform detailed ablation studies, as shown in Tab. 1, 6, and 7. These ablations are executed on the QVHighlights *validation* subset, as the online submission limit of five trials on the test split server. In row 1 of Tab. 6, we first replicate Moment-DETR as our baseline model. Subsequently, to decouple the feature

Table 5. Quantitative results on Charades-STA *testing* set. Comparison against the state-of-the-art temporal grounding models

Methods	R1		R5	
	@0.5	**@0.7**	**@0.5**	**@0.7**
MCN (Hendricks et al., 2017)	4.05	-	-	-
2D-TAN (S. Zhang et al., 2020)	39.70	23.31	80.32	51.26
UMT (Y. Liu et al., 2022)	49.35	26.16	-	-
M-DETR (Lei et al., 2021)	53.63	31.37	-	-
PEARL (Zhang & Radke, 2022)	53.50	35.40	-	-
QD-Net (Ours)	**57.43**	**36.39**	**87.38**	**63.01**

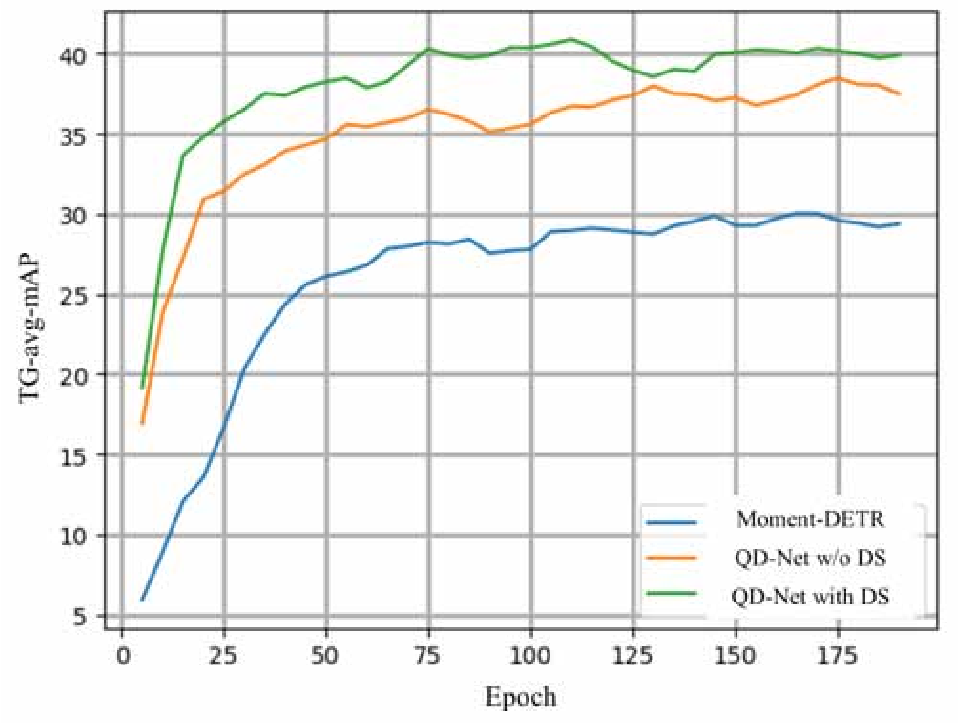
Table 6. Ablation studies of query-guided refinement module (QRM) on QVHighlights *validation* split. All results are our reproduction or experimental results. Moment-DETR (Lei et al, 2021) is abbreviated as M-DETR

Methods	Layers	TG			HD ($\geq$ VG)	
		R1@0.5	**R1@0.7**	**mAP Avg.**	**mAP**	**HIT@1**
M-DETR (Lei et al., 2021)	-	53.94	34.84	32.20	35.65	55.55
M-DETR + QRM	1	**59.87**	43.42	**36.86**	**37.06**	**58.26**
M-DETR + QRM	2	58.71	**43.35**	36.01	36.99	57.16
QD-Net w/o QRM	-	59.74	42.65	38.21	37.59	60.14
QD-Net + QRM	1	**62.32**	**45.61**	**41.46**	38.56	**62.06**
QD-Net + QRM	2	60.39	42.58	40.11	38.19	60.00
QD-Net + QRM	3	60.84	44.90	39.26	**38.77**	61.74

Table 7. Ablation experiments of span decoder with dynamic spans (DS) on QVHighlights *validation* split. All results are our reproduction or experimental results

Methods	Layers	TG			HD ($\geq$ VG)	
		R1@0.5	R1@0.7	mAP Avg.	mAP	HIT@1
M-DETR (Lei et al., 2021)	-	53.94	34.84	32.20	35.65	55.55
M-DETR +DS	4	**59.29**	**42.71**	**36.96**	**38.12**	**61.21**
UMT (Y. Liu et al., 2022)	-	60.06	43.42	38.13	**39.01**	**62.71**
UMT+ DS	4	**60.97**	**45.48**	**40.34**	38.57	61.29
QD-Net w/o DS	-	59.87	43.16	37.03	**39.13**	**63.24**
QD-Net + DS	2	62.23	46.32	41.37	38.46	61.16
QD-Net + DS	4	**62.32**	**45.61**	**41.46**	38.56	62.06
QD-Net + DS	6	61.16	46.24	40.83	38.48	61.68

Figure 7. Compared to Moment-DETR on the QVHighlights *validation* split, QD-Net with dynamic spans (DS) converges faster and performs better



encoding with interaction process, we incorporate a plug-and-play query-guided refinement module (QRM) into Moment-DETR, as shown in rows 2 and 3 of Tab. 6. Our findings suggest that QRM effectively fuses visual and textual features, improving performance considerably.

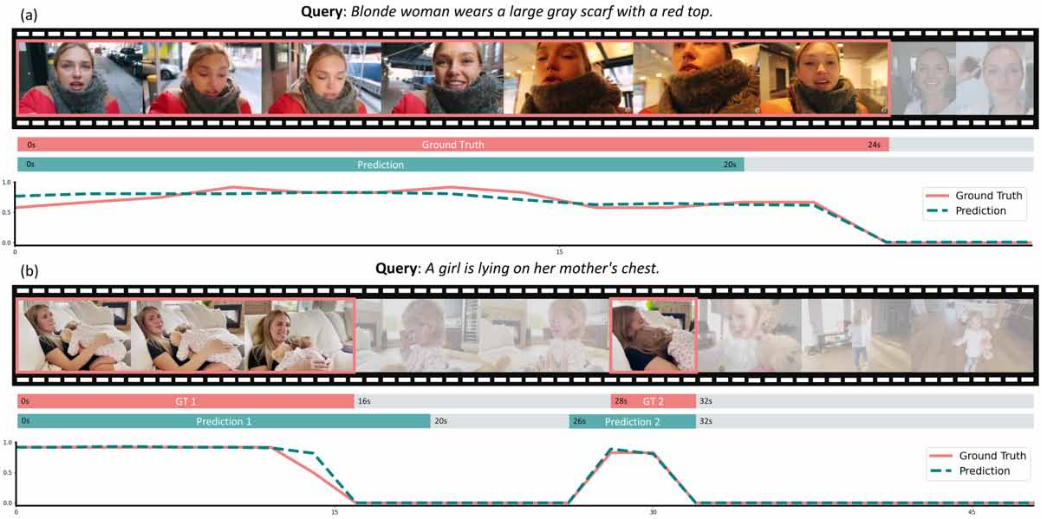
To validate the effectiveness of the span decoder with dynamic spans (DS), we incorporate DS into the original Moment-DETR and UMT. The former initially utilizes a straightforward transformer decoder, while the latter removes both the decoder and the Hungarian algorithm. As illustrated in lines 1 to 4 of Tab. 7, our results confirm that DS considerably enhances the model's performance on the TG task. Additionally, it is evident that UMT's performance on TG deteriorates when the decoder is removed. Then, we combine QRM and DS, as shown in line 4 of Tab. 1, and find a marked improvement in the model's performance. Building upon this, we attempt to integrate transformer or PoolFormer (Yu et al., 2022) as a uni-modal encoder, as illustrated in Tab. 1. Surprisingly, using a simple pooling operation in the encoder effectively captures contextual information within a single modality, yielding better results than self-attention. However, excessive encoder layers may lead to model overfitting. Consequently, we configure the encoder with two layers.

In Tab. 6, lines 4 to 7 display the results after removing QRM from our framework. The Average mAP (mAP Avg.) and HIT@1 drop significantly by 3.25% and 1.92%, respectively, illustrating that QRM is an indispensable part of cross-modal fusion. Moreover, an excessive number of QRM layers can lead to model overfitting, so we limit the number of QRM layers to 1. As shown in lines 4 to 7 of Tab. 7, adding DS into QD-Net results in a remarkable improvement in temporal grounding (TG) performance, where mAP Avg. is increased by 4.43%. However, the metrics of HD on QD-Net with DS see a slight decline. We speculate that over-optimizing the decoder may introduce noise into the cross-modal features. Furthermore, Figure 7 contrasts the efficiency of QD-Net with the inclusion and exclusion of the DS component. QD-Net with DS converges at the 75th epoch and performs better. In contrast, QD-Net without DS only converges at the 125th epoch and displays inferior performance. This comparison indicates that using learnable 2D spans to represent queries leads training to converge faster and results in improved performance.

CONCLUSION

In this paper, we propose a novel model called QD-Net (Query-guided refinement and Dynamic spans Network) for video highlight detection and temporal grounding (HD&TG) in online databases. Unlike previous transformer encoder decoder–based methods that combine multi-modal features in a coarse manner, our proposed QD-Net includes a query-guided refinement module. This module effectively separates feature encoding from the interaction process and produces query-relevant cross-modal

Figure 8. Visual results on the QVHighlights *validation* set. (a) Our method adaptively predicts single span and clip-wise saliency scores. (b) Our method accurately processes intricate video containing multiple spans



features. Subsequently, we introduce an encoder containing a simplified pooling mechanism to model the intra-modal information under a global view. In addition, we design a dynamic span decoder to accelerate training convergence for TG. This decoder utilizes learnable 2D spans to represent queries of the decoder, thereby strengthening the connection between learnable embedding and temporal span information. Finally, to demonstrate the superiority and robustness of our method, we undertake detailed experiments and ablation studies on three benchmark datasets: QVHighlights, TVSum, and Charades-STA. The results indicate that QD-Net outperforms the current state-of-the-art methods.

In future work, it will be essential to further investigate the span decoder. As we have observed in QD-Net, while the decoder significantly enhances the convergence and performance of TG, there is a slight decrease in accuracy for HD. Additionally, there is great value in leveraging large-scale visual-linguistic models to generate highlight scores. Finally, attempting additional multi-modal features, such as optical flow and depth maps, would be worthwhile.

AUTHOR NOTE

Yifang Xu: https://orcid.org/0009-0009-8332-4791

This research received no external funding. The authors declare there is no conflict of interest.

REFERENCES

Abbas, N., Taleb, S., & Hajj, H. (2021). Video features with impact on user quality of experience. In *2021 3rd IEEE Middle East and North Africa COMMunications Conference (MENACOMM)*. IEEE. doi:10.1109/MENACOMM50742.2021.9678269

Anil, B. C., Dayananda, P., Nethravathi, B., & Raisinghani, M. S. (2022). Efficient local cloud-based solution for liver cancer detection using deep learning. *International Journal of Cloud Applications and Computing*, *12*(1), 1–13. doi:10.4018/IJCAC.2022010109

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). *Layer normalization.* ArXiv Prepr. ArXiv160706450.

Badamdorj, T., Rochan, M., Wang, Y., & Cheng, L. (2021). Joint visual and audio learning for video highlight detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. doi:10.1109/ICCV48922.2021.00802

Cao, M., Chen, L., Shou, M. Z., Zhang, C., & Zou, Y. (2021). On pursuit of designing multi-modal transformer for video grounding. In *Empirical Methods in Natural Language Processing*. EMNLP. doi:10.18653/v1/2021.emnlp-main.773

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. *European Conference on Computer Vision (ECCV).*

Chen, Y. W., Tsai, Y. H., & Yang, M. H. (2021). End-to-end multi-modal video temporal grounding. *NeurIPS*, *34*, 28442–28453.

Chu, J., Zhao, X., Song, D., Li, W., Zhang, S., Li, X., & Liu, A. (2022). Improved semantic representation learning by multiple clustering for image-based 3D model retrieval. *International Journal on Semantic Web and Information Systems*, *18*(1), 1–20. doi:10.4018/IJSWIS.297033

Escorcia, V., Soldan, M., Sivic, J., Ghanem, B., & Russell, B. (2019). *Temporal localization of moments in video collections with natural language.* ArXiv Prepr. ArXiv190712763.

Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). SlowFast networks for video recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE.

Galiautdinov, R. (2021). Digitally-signed video/audio streams as prevention of AI-based attacks. *International Journal of Software Science and Computational Intelligence*, *13*(4), 54–63. doi:10.4018/IJSSCI.2021100104

Gao, J., Sun, C., Yang, Z., & Nevatia, R. (2017). Tall: Temporal activity localization via language query. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE. doi:10.1109/ICCV.2017.563

Hammad, M., Alkinani, M. H., Gupta, B. B., & El-Latif, A. A. A. (2021). Myocardial infarction detection based on deep neural network on imbalanced data. *Multimedia Systems*, *28*(4), 1373–1385. doi:10.1007/s00530-020-00728-8

Hamza, A., Javed, A. R., Iqbal, F., Kryvinska, N., Almadhor, A. S., Jalil, Z., & Borghol, R. (2022). Deepfake audio detection via MFCC features using machine learning. *IEEE Access : Practical Innovations, Open Solutions*, *10*, 134018–134028. doi:10.1109/ACCESS.2022.3231480

Hasib, K. M., Towhid, N. A., & Islam, M. R. (2021). HSDLM: A hybrid sampling with deep learning method for imbalanced data classification. *International Journal of Cloud Applications and Computing*, *11*(4), 1–13. doi:10.4018/IJCAC.2021100101

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE.

Hendricks, L. A., Wang, O., Shechtman, E., Sivic, J., Darrell, T., & Russell, B. (2017). Localizing moments in video with temporal language. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., & Zisserman, A. (2017). *The kinetics human action video dataset.* ArXiv Prepr. ArXiv170506950.

Lei, J., Berg, T. L., & Bansal, M. (2021). Detecting moments and highlights in videos via natural language queries. *NeurIPS*, *34*, 11846–11858.

Lei, J., Yu, L., Berg, T. L., & Bansal, M. (2020). TVR: A large-scale dataset for video-subtitle moment retrieval. In *European Conference on Computer Vision (ECCV).* Springer. doi:10.1007/978-3-030-58589-1_27

Li, S., Qin, D., Wu, X., Li, J., Li, B., & Han, W. (2022). False alert detection based on deep learning and machine learning. *International Journal on Semantic Web and Information Systems*, *18*(1), 1–21. doi:10.4018/IJSWIS.313190

Lin, T., Wang, Y., Liu, X., & Qiu, X. (2023). *A survey of transformers.* ArXiv preprint arXiv: 2106.04554.

Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., & Zhang, L. (2022). *DAB-DETR: Dynamic anchor boxes are better queries for DETR.* ArXiv preprint arXiv:2201.12329.

Liu, W., Mei, T., Zhang, Y., Che, C., & Luo, J. (2015). Multi-task deep visual-semantic embedding for video thumbnail selection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE. doi:10.1109/CVPR.2015.7298994

Liu, Y., Li, S., Wu, Y., Chen, C. W., Shan, Y., & Qie, X. (2022). UMT: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE. doi:10.1109/CVPR52688.2022.00305

Loshchilov, I., & Hutter, F. (2017). *Decoupled weight decay regularization.* ArXiv Prepr. ArXiv171105101.

Narayanasamy, S. K., Srinivasan, K., Hu, Y. C., Masilamani, S. K., & Huang, K. Y. (2022). A contemporary review on utilizing semantic web technologies in healthcare, virtual communities, and ontology-based information processing systems. *Electronics (Basel)*, *11*(3), 453. doi:10.3390/electronics11030453

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *International Conference on Machine Learning (ICML).*

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., & Savarese, S. (2019). Generalized intersection over union: A metric and a loss for bounding box regression. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE. doi:10.1109/CVPR.2019.00075

Sahoo, S. R., & Gupta, B. B. (2021). Multiple features based approach for automatic fake news detection on social networks using deep learning. *Applied Soft Computing*, *100*, 106983. doi:10.1016/j.asoc.2020.106983

Shankar, K., Perumal, E., Elhoseny, M., Taher, F., Gupta, B. B., & El-Latif, A. A. A. (2021). Synergic deep learning for smart health diagnosis of COVID-19 for connected living and smart cities. *ACM Transactions on Internet Technology*, *22*(3), 1–14. doi:10.1145/3453168

Sigurdsson, G. A., Varol, G., Wang, X., Farhadi, A., Laptev, I., & Gupta, A. (2016). Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision (ECCV).* Springer.

Simonyan, K., & Zisserman, A. (2014). *Very deep convolutional networks for large-scale image recognition.* ArXiv Prepr. ArXiv140915569.

Song, Y., Redi, M., Vallmitjana, J., & Jaimes, A. (2016). To click or not to click: Automatic selection of beautiful thumbnails from videos. In *CIKM '16: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management.* Springer. doi:10.1145/2983323.2983349

Song, Y., Vallmitjana, J., Stent, A., & Jaimes, A. (2015). TVSum: Summarizing web videos using titles. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE.

Srivastava, A. M., Rotte, P. A., Jain, A., & Prakash, S. (2022). Handling data scarcity through data augmentation in training of deep neural networks for 3D data processing. *International Journal on Semantic Web and Information Systems*, *18*(1), 1–16. doi:10.4018/IJSWIS.297038

Stergiou, C. L., Psannis, K. E., & Gupta, B. B. (2021). InFeMo: Flexible big data management through a federated cloud system. *ACM Transactions on Internet Technology*, *22*(2), 1–22. doi:10.1145/3426972

Sun, Z., Cao, S., Yang, Y., & Kitani, K. M. (2021). Rethinking transformer-based set prediction for object detection. ICCV. doi:10.1109/ICCV48922.2021.00359

Taleb, S., & Abbas, N. (2022). Hybrid machine learning classification and inference of stalling events in mobile videos. In *2022 4th IEEE Middle East and North Africa COMMunications Conference (MENACOMM)*. IEEE. doi:10.1109/MENACOMM57252.2022.9998209

Tang, C., Zhao, Y., Wang, G., Luo, C., Xie, W., & Zeng, W. (2022). Sparse MLP for image recognition: Is self-attention really necessary? *Proceedings of the AAAI Conference on Artificial Intelligence.* doi:10.5772/intechopen.95124

Tembhurne, J. V., Almin, M. M., & Diwan, T. (2022). Mc-DNN: Fake news detection using multi-channel deep neural networks. *International Journal on Semantic Web and Information Systems*, *18*(1), 1–20. doi:10.4018/IJSWIS.295553

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*.

Wang, L., Liu, D., Puri, R., & Metaxas, D. N. (2020). Learning trailer moments in full-length movies with co-contrastive attention. In *European Conference on Computer Vision (ECCV)*. Springer. doi:10.1007/978-3-030-58523-5_18

Xu, Y., Li, M., Peng, C., Li, Y., & Du, S. (2021a). Dual attention feature fusion network for monocular depth estimation. In *CICAI 2021*. Springer. doi:10.1007/978-3-030-93046-2_39

Xu, Y., Peng, C., Li, M., Li, Y., & Du, S. (2021b). Pyramid feature attention network for monocular depth prediction. In *IEEE International Conference on Multimedia and Expo (ICME)*. IEEE. doi:10.1109/ICME51207.2021.9428446

Xu, Y., Sun, Y., Li, Y., Shi, Y., Zhu, X., & Du, S. (2023). *MH-DETR: Video moment and highlight detection with cross-modal transformer.* ArXiv preprint arXiv:2305.00355.

Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., & Yan, S. (2022). Metaformer is actually what you need for vision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. doi:10.1109/CVPR52688.2022.01055

Zhang, B., Li, Y., Yuan, C., Xu, D., Jiang, P., & Shan, Y. (2020). *A simple yet effective method for video temporal grounding with cross-modality attention.* ArXiv Prepr. ArXiv200911232.

Zhang, H., Sun, A., Jing, W., & Zhou, J. T. (2022). *Temporal sentence grounding in videos: A survey and future directions.* ArXiv preprint arXiv: 2201.08071.

Zhang, K., Chao, W. L., Sha, F., & Grauman, K. (2016). Video summarization with long short-term memory. In *European Conference on Computer Vision (ECCV)*. Springer.

Zhang, L., & Radke, R. J. (2022). Natural language video moment localization through query-controlled temporal convolution. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE. doi:10.1109/WACV51458.2022.00258

Zhang, S., Peng, H., Fu, J., & Luo, J. (2020). Learning 2D temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI. doi:10.1609/aaai.v34i07.6984

Yifang Xu obtained his master's from Nanjing University in 2023. He has worked in the Nanjing University School of Electronic Science and Engineering. His research interests include computer vision, video understanding, and multimodal learning.

Yunzhuo Sun holds a master's degree from Nanjing University, Jinling college, where he graduated in 2020. Now a postgraduate student at Hubei Normal University's School of Physics and Electronics, his research interests include video understanding and generative learning.

Zien Xie graduated from Nanjing University with a master's degree in 2022. Now a postgraduate student at Nanjing University's School of Electronic Science and Engineering, his research interests include computer vision and video retrieval.

Benxiang Zhai graduated from Nanjing University in 2023 with a master's degree. Now a postgraduate student at Nanjing University's School of Electronic Science and Engineering, his research interests include computer vision and deep learning.

Youyao Jia has worked in Gosuncn Chuanglian Technology Co., Ltd. His research interests include 3D reconstruction and machine learning.

Sidan Du holds a Ph.D. A professor at Nanjing University's School of Electronic Science and Engineering, his research interests include probability and statistics, computer vision, and deep learning.