BAM-DETR: Boundary-Aligned Moment Detection Transformer for Temporal Sentence Grounding in Videos

Pilhyeon Lee[1]* and Hyeran Byun[2]

[1] Department of Artificial Intelligence, Inha University
[2] Department of Computer Science, Yonsei University

Abstract. Temporal sentence grounding aims to localize moments relevant to a language description. Recently, DETR-like approaches achieved notable progress by predicting the center and length of a target moment. However, they suffer from the issue of center misalignment raised by the inherent ambiguity of moment centers, leading to inaccurate predictions. To remedy this problem, we propose a novel boundary-oriented moment formulation. In our paradigm, the model no longer needs to find the precise center but instead suffices to predict any anchor point within the interval, from which the boundaries are directly estimated. Based on this idea, we design a boundary-aligned moment detection transformer, equipped with a dual-pathway decoding process. Specifically, it refines the anchor and boundaries within parallel pathways using global and boundary-focused attention, respectively. This separate design allows the model to focus on desirable regions, enabling precise refinement of moment predictions. Further, we propose a quality-based ranking method, ensuring that proposals with high localization qualities are prioritized over incomplete ones. Experiments on three benchmarks validate the effectiveness of the proposed methods. The code is available here.

Keywords: Temporal sentence grounding · Detection transformer

1 Introduction

Recent years have witnessed a notable surge in the popularity of short-form video content on social media platforms like TikTok, YouTube Shorts, and Instagram Reels. As such, users prefer to selectively engage with short *moments* of interest rather than passively watch an entire long video. This trend highlights the importance of localizing desired moments. As a result, moment localization tasks have emerged as pivotal research topics in video understanding, including temporal action detection [20, 29, 75], video summarization [50, 71], and highlight detection [48, 58]. Within this context, we tackle temporal sentence grounding [1, 72], aiming to retrieve moments corresponding to free-form language descriptions.

To address temporal sentence grounding, numerous efforts have been undertaken in the last decade [9, 11, 31, 49, 64, 72]. Especially, taking inspiration

* Correspondence to: Pilhyeon Lee <pilhyeon.lee@inha.ac.kr>

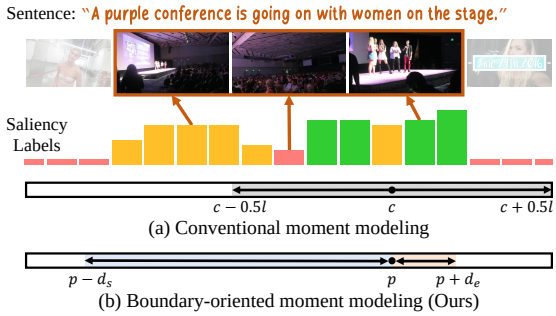


Fig. 1: Comparison of moment modeling approaches under the scenario of an ambiguous center from QVHighlights. (a) The conventional method formulates a moment with a tuple of (c, l). (b) In contrast, we propose to model it with a triplet of (p, d_s, d_e).

from DETR [3], query-based approaches have become a promising research direction owing to the architectural simplicity [17,21,26,37,41]. By decoding temporal spans (*i.e.*, moments) from a handful set of learnable queries, they achieve promising grounding performance while maintaining high inference speed.

Existing query-based models typically predict a moment using its center and width, *i.e.*, (c, l), under the assumption that the boundaries are equidistant from the center. However, such a formulation can be problematic as the center of a moment might be ambiguous. An illustrative example is presented in Fig. 1 (top), where the frame at the center of the ground-truth moment is less relevant to the given sentence, as demonstrated by the low saliency label. Indeed, a moment center does not necessarily serve as the best representative of the sentence. This ambiguity can challenge the model's ability to precisely locate the centers, and the misaligned centers lead to low-quality predictions (Fig. 1a). To probe the impact of center misalignment, we conducted a diagnostic experiment in Table 1. The results reveal that (i) existing methods struggle with detecting accurate center points and (ii) they suffer from significant performance drops when the predicted centers deviate from the ground-truth centers (*i.e.*, large center errors).

To address this challenge, we present a novel boundary-oriented formulation for moments, as illustrated in Fig. 1b, where each moment is represented by a triplet consisting of an anchor point and its distances to the boundaries, *i.e.*, (p, d_s, d_e). This asymmetric formulation liberates the model from the stringent requirement of predicting the center. Instead, it is sufficient for the model to predict *any* salient anchor within the target moment, and the distances from the anchor to the onset and offset are predicted subsequently. By directly locating the boundaries based on the anchor point, the model can achieve improved boundary alignment, even when the anchor point does not coincide with the actual center.

Building upon the proposed moment modeling, we introduce a new framework equipped with a dedicated decoder design, dubbed Boundary-Aligned Moment Detection Transformer (BAM-DETR). The design of its decoding layers originates from the intuition that the refining processes for an anchor and boundaries should be distinct from each other. That is to say, a model needs to scan over the whole video to find a potential anchor point that enables estimating the

Table 1: Impact of misaligned centers on QVHighlights. The top-1 predictions are grouped based on their center errors normalized by ground-truth lengths. For each group, we present the mean IoU (%) and the proportion (in parentheses). Only the predictions whose centers fall within the ground-truth moments are considered here.

Method	[0, 0.1)	[0.1, 0.2)	[0.2, 0.3)	[0.3, 0.4)	[0.4, 0.5)	All
Moment-DETR [21]	83.20 (46 %)	65.00 (24 %)	54.82 (14 %)	43.77 (9 %)	34.98 (7 %)	67.84 (100 %)
QD-DETR [41]	87.79 (56 %)	67.95 (18 %)	55.40 (11 %)	44.93 (9 %)	36.45 (5 %)	74.09 (100 %)
BAM-DETR† (Ours)	77.62 (24 %)	78.20 (21 %)	77.52 (21 %)	78.96 (22 %)	71.08 (12 %)	77.21 (100 %)

†anchor points are utilized for grouping.

rough location of the target moment. On the other hand, it is required to focus on fine-grained details in the vicinity to refine the boundaries to be aligned with those of the target moments. From this motivation, in contrast to existing methods, our model adopts a dual-pathway decoding pipeline to predict an anchor point and boundaries in a parallel way. To be specific, we leverage two different types of queries respectively for anchor and boundary refinement. The former aggregates global information with standard attention, while the latter concentrates on the sparse local neighborhood of the boundaries using the proposed boundary-focused attention. These distinct designs of two pathways allow for effective moment localization with minimal computational overhead increase.

In addition, we identify the problem of the conventional scoring method, where binary classification (or matching) scores are used for proposal ranking. This leads to suboptimal results for the grounding task since a fractional moment may have high matching scores with the given sentence. To handle this issue, we propose to rank proposals based on their localization qualities. Accordingly, we modify the typical matching function and training objectives of the query-based model to be localization-oriented by discarding the role of classification scores. In this way, our model can prioritize the moment proposals exhibiting high overlap with ground truths at inference, leading to improved grounding performance.

The advantage of our BAM-DETR is showcased in Table 1. It can be observed that the anchor points predicted by our model are evenly distributed within the ground-truth intervals. Importantly, our model consistently produces precise moments with high IoUs (> 0.7) across different groups, indicating that it does not depend on accurate center prediction. On average, our model shows superior grounding performance over the existing methods with the help of our moment formulation and quality-based scoring. In a later section, we validate the efficacy of the proposed methods through extensive experiments. Notably, our model outperforms previous methods by large margins on three public benchmarks.

2 Related Works

2.1 Temporal Sentence Grounding in Videos

Temporal sentence grounding requires seeking temporal spans semantically relevant to the given sentence in a video. Proposal-based approaches adopt the

two-stage pipeline, *i.e.*, proposal generation and ranking. They generate moment proposals by relying on sliding windows [11,13,34,73] or utilizing pre-defined anchors [4,59,64,67,74]. Several works process all possible candidates at once with 2D maps [25,32,56,57,72]. Meanwhile, proposal-free methods are developed for efficient grounding by directly predicting the moments [42,65] or estimating the probabilities of each frame being starting and ending positions [15,70]. Some approaches perform dense regression by predicting the boundaries from individual frames [6,39,66]. Recently, query-based models streamline the complicated sentence grounding pipeline by removing handcrafted techniques [2,17,26,37,60]. There are also attempts to unify temporal sentence grounding with other video understanding tasks into a single framework [28,61].

Our method belongs to the query-based group [21,41], inheriting the benefit of architectural simplicity. In contrast to others, we employ a boundary-oriented formulation of moments to relieve the heavy reliance on center predictions, leading to better boundary alignment. Our method also relates to dense regression methods [6,28,66] that predict boundaries from each frame as an anchor. Comparatively, our model leverages dynamic anchors that are gradually adjusted through decoding, enabling precise grounding using a small set of predictions.

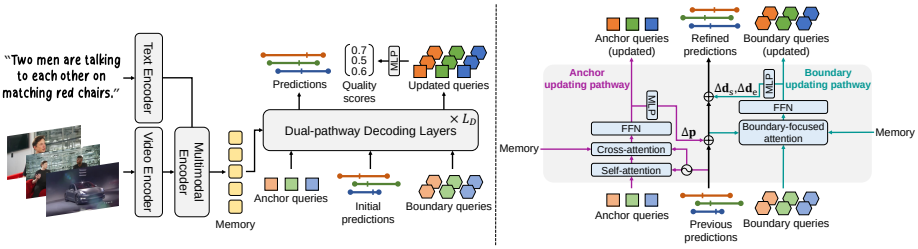
2.2 Detection Transformers

Query-based temporal sentence grounding models by design are closely related to the family of detection transformers (DETR) [3]. Since the advent of DETR, a number of variants have been introduced to improve it from various perspectives [30,33,36,53]. Some works focus on reducing the excessive computational costs of vanilla transformers in order to leverage multi-scale features [7,23,47,76,78]. On the other hand, several methods attempt to speed up the model convergence by manipulating the attention operations [12,35,62,68] or incorporating the denoising process during the model training [24,69].

The most relevant works to ours are those which propose explicit anchor modeling for object queries using center points [40,55] or boxes (center, width, and height) [35,78]. In comparison to an object, a moment in temporal sentence grounding has its own challenges such as center ambiguity and indistinct boundaries. To accommodate the discrepancies, we propose a novel boundary-oriented modeling of moments to replace the conventional center-based 1D box modeling. The advantages of our approach are clearly verified in the experiments.

3 Method

Given an untrimmed video and a sentence, the goal of temporal sentence grounding is to localize relevant moments $\{\varphi_n = (t_{s_n}, t_{e_n})\}_{n=1}^{N}$, where N denotes the number of ground truths in the video and φ_n indicates the temporal interval of the n-th moment. Note that a video may have multiple moments that match the sentence, *i.e.*, $N \geq 1$. During the test time, the model is expected to produce a total of M predictions, $\{(\hat{\varphi}_m, q_m)\}_{m=1}^{M}$, where $\hat{\varphi}_m = (\hat{t}_{s_m}, \hat{t}_{e_m})$ is the m-th prediction, while q_m is its score for ranking.



(a) Overall framework of BAM-DETR (b) Dual-pathway decoding layer

Fig. 2: (a) Overview of the proposed BAM-DETR. (b) Details of the proposed dual-pathway decoding layer. It consists of two parallel pathways respectively for anchor and boundary updates, which refine previous moment predictions in a sequential manner.

Motivation. We tackle two main problems in the current prediction process. On one hand, existing query-based approaches adopt the symmetric design of (c, l) to predict a moment, *i.e.*, $\hat{\varphi} = (c-0.5l, c+0.5l)$. As we discuss in Sec. 1, this strategy suffers from the issue of over-reliance on the center prediction, leading to unstable performance. To address this, we propose a boundary-oriented modeling with a triplet of (p, d_s, d_e), where a moment is represented by $\hat{\varphi} = (p - d_s, p + d_e)$. This asymmetric design enables direct alignment of boundaries without relying on precise center prediction. On the other hand, previous models use classification scores as q_m. This scoring is prone to sub-optimal solutions since only a fraction of the moment may well match the sentence, leading incomplete predictions to be highly ranked. To handle this, we propose quality-based scoring to sort the proposals based on their localization qualities rather than the degree of matching.

3.1 Overview

As shown in Fig. 2a, our BAM-DETR follows the encoder-decoder pipeline. Briefly, it first extracts multimodal video features using the encoders. Taking them as memory, the dual-pathway decoder predicts temporal spans and their quality scores by progressively refining learnable initial spans.

3.2 Feature Extraction

Following the convention [42, 70, 72], we employ pre-trained encoders for token-level unimodal feature extraction. It is worth noting that all unimodal encoders are kept frozen during training to avoid memory exploding. In this stage, we obtain D_v-dim video features $E_v \in \mathbb{R}^{N_v \times D_v}$ and D_t-dim text features $E_t \in \mathbb{R}^{N_t \times D_t}$, where N_v and N_t are the numbers of clips and words, respectively.

3.3 Multimodal Encoder

We feed the unimodal features of the video and text to the multimodal encoder, so as to fuse them into text-aware video representations for temporal sentence grounding. While various multimodal encoders are explored [21, 28, 37], we

adopt the text-to-video encoder design [41] consisting of cross- and self-attention blocks [54]. Before getting into the encoder, unimodal features are projected into a shared space to facilitate cross-modal interaction, *i.e.*, $\mathcal{V} = f_v(E_v) \in \mathbb{R}^{N_v \times D}$, $\mathcal{T} = f_t(E_t) \in \mathbb{R}^{N_t \times D}$, where D is the embedding dimension. Afterward, multi-head cross-attention blocks are used to inject textual information into the clip-level video representations. In specific, we project $\mathcal{V}$ to $\mathbf{Q}_\mathcal{V}$ (query) while $\mathcal{T}$ to $\mathbf{K}_\mathcal{T}$ (key) and $\mathbf{V}_\mathcal{T}$ (value). Then a cross-attention block can be formulated as:

$$\begin{aligned} \mathcal{V}' &= \text{softmax}\Big(\frac{\mathbf{Q}_\mathcal{V}\mathbf{K}_\mathcal{T}^\top}{\sqrt{D}}\Big)\mathbf{V}_\mathcal{T} + \mathcal{V}, \\ \mathcal{V}'' &= \text{FFN}(\mathcal{V}') + \mathcal{V}', \end{aligned} \tag{1}$$

where $\text{FFN}(\cdot)$ is a feed-forward network. Although we here present the single-head attention block, it can readily generalize to a multi-headed version [54]. We denote the resulting multimodal representations obtained after L_E multi-head cross-attention blocks by $\tilde{\mathcal{V}} \in \mathbb{R}^{N_v \times D}$.

Subsequently, self-attention blocks are leveraged to enhance the representations by allowing the inter-clip interaction. Here we project $\tilde{\mathcal{V}}$ to $\mathbf{Q}_{\tilde{\mathcal{V}}}$, $\mathbf{K}_{\tilde{\mathcal{V}}}$, and $\mathbf{V}_{\tilde{\mathcal{V}}}$. We note that the query and the key are supplemented with fixed sinusoidal positional encoding [3,54] for temporal awareness. Then the self-attention block is defined in a similar way to Eq. (1) but with different inputs. The enhanced clip-level representations after L_E multi-head self-attention blocks are denoted by $\hat{\mathcal{V}} \in \mathbb{R}^{N_v \times D}$, which will serve as the *memory* for the decoder.

It is widely known that providing saliency guidance to the memory features helps the model to better understand the semantic relationship between the video and text [21, 28]. As in previous works [17, 26], we impose saliency score constraints on the memory features. Specifically, we leverage a saliency predictor $S(\cdot)$ and train the model with the following margin-based training objective.

$$\mathcal{L}_{\text{margin}} = \max(0, \alpha + S(\hat{v}^{\text{low}}) - S(\hat{v}^{\text{high}})), \tag{2}$$

where α is a margin and $(\hat{v}^{\text{low}}, \hat{v}^{\text{high}})$ is the sampled feature pair satisfying that the saliency label of $\hat{v}^{\text{low}}$ is lower than that of $\hat{v}^{\text{high}}$. In case of the absence of saliency labels, we collect clips within and outside the ground-truth moment intervals to build a pair. In addition, we employ the rank-aware contrastive loss $\mathcal{L}_{\text{cont}}$ and the negative relation loss $\mathcal{L}_{\text{neg}}$, following Moon *et al.* [41]. Due to space limits, we refer the readers to Appendix for the loss formulations. In summary, the overall saliency loss is defined as $\mathcal{L}_{\text{sal}} = \mathcal{L}_{\text{margin}} + \mathcal{L}_{\text{cont}} + \mathcal{L}_{\text{neg}}$.

3.4 Dual-pathway Decoder

With the multimodal representations $\hat{\mathcal{V}}$ as memory features, we aim to localize temporal spans corresponding to the sentence. We adopt the new boundary-oriented formulation for moment prediction, where each prediction is represented by a triplet of (p, d_s, d_e), where p is the anchor point, while d_s and d_e are the distances from the anchor to the starting and ending points, respectively. To make full use of the proposed formulation, we design a dual-pathway decoding

layer with two parallel pathways (Fig. 2b). Formally, the inputs of the l-th layer are anchor queries $\mathbf{C}_p^l \in \mathbb{R}^{M\times D}$, boundary queries $\mathbf{C}_s^l, \mathbf{C}_e^l \in \mathbb{R}^{M\times D}$, and previous moment predictions $\mathbf{A}^l = [\mathbf{p}^l; \mathbf{d}_s^l; \mathbf{d}_e^l] \in \mathbb{R}^{M\times 3}$, where M is the number of queries (predictions). Note that the initial queries and spans, *i.e.*, $\{\mathbf{C}_p^0, \mathbf{C}_s^0, \mathbf{C}_e^0, \mathbf{A}^0\}$, are learnable parameters. We elaborate on the two pathways in the following.

Anchor updating pathway. Given the predictions and the anchor queries from the preceding layer, the goal of this pathway is to adjust the anchor position so that the boundaries can be predicted based on it. Intuitively, in order to obtain a valuable anchor point without redundancy, an anchor query should communicate with other queries as well as the memory features. To this goal, the anchor updating pathway consists of a self-attention layer, a cross-attention layer, and a feed-forward network. In the self-attention layer, the anchor queries $\mathbf{C}_p^l$ are first projected into $\mathbf{Q}_{\mathbf{C}_p^l}$, $\mathbf{K}_{\mathbf{C}_p^l}$, and $\mathbf{V}_{\mathbf{C}_p^l}$. Since an anchor query itself lacks positional information, we build positional encoding. Following the previous works [17,41], we extend the current spans $\mathbf{A}^l$ to the positional information of the queries, *i.e.*, $\mathbf{P}_{\mathbf{A}^l} = \mathrm{MLP}(\mathrm{PE}(\mathbf{A}^l)) \in \mathbb{R}^{M\times D}$, where $\mathrm{PE}(\cdot)$ denotes the point-wise mapping from a position to the corresponding sinusoidal encoding while $\mathrm{MLP}(\cdot)$ is multi-layer perceptron. The self-attention for anchor queries is defined as:

$$\tilde{\mathbf{C}}_p^l = \mathrm{softmax}\Big(\frac{\big(\mathbf{Q}_{\mathbf{C}_p^l} + \mathbf{P}_{\mathbf{A}^l}\big)\big(\mathbf{K}_{\mathbf{C}_p^l} + \mathbf{P}_{\mathbf{A}^l}\big)^\top}{\sqrt{D}}\Big)\mathbf{V}_{\mathbf{C}_p^l} + \mathbf{C}_p^l. \tag{3}$$

After inter-query interaction, we employ a global cross-attention layer to aggregate multi-modal features from the memory. The anchor queries $\tilde{\mathbf{C}}_p^l$ is projected into $\mathbf{Q}_{\tilde{\mathbf{C}}_p^l}$ while the memory $\hat{\mathcal{V}}$ is projected to $\mathbf{K}_{\hat{\mathcal{V}}}$ and $\mathbf{V}_{\hat{\mathcal{V}}}$. To make the query location-aware, we leverage the sinusoidal encoding of current anchor positions, *i.e.*, $\mathbf{P}_{\mathbf{p}^l} = \mathrm{PE}(\mathbf{p}^l) \in \mathbb{R}^{M\times D}$. Similarly, the memory leverages the positional encoding, *i.e.*, $\mathbf{P}_{\hat{\mathcal{V}}} = \mathrm{PE}(\hat{\mathcal{V}}) \in \mathbb{R}^{N_v\times D}$. We use concatenation instead of summation to separate the roles of features and positional encoding [35, 40]. The cross-attention between anchor queries and the memory can be expressed as:

$$\hat{\mathbf{C}}_p^l = \mathrm{softmax}\Big(\frac{\big(\mathbf{Q}_{\tilde{\mathbf{C}}_p^l} \parallel \mathbf{P}_{\mathbf{p}^l}\big)\big(\mathbf{K}_{\hat{\mathcal{V}}} \parallel \mathbf{P}_{\hat{\mathcal{V}}}\big)\big)^\top}{\sqrt{2D}}\Big)\mathbf{V}_{\hat{\mathcal{V}}} + \tilde{\mathbf{C}}_p^l. \tag{4}$$

After all, the anchor queries are updated with a feed-forward network, *i.e.*, $\mathbf{C}_p^{(l+1)} = \mathrm{FFN}(\hat{\mathbf{C}}_p^l) + \hat{\mathbf{C}}_p^l$. Lastly, we adjust the anchor positions using sigmoid-based refinement [78]: $\hat{\mathbf{A}}^l = [\mathbf{p}^{(l+1)}; \mathbf{d}_s^l; \mathbf{d}_e^l]$ where $\mathbf{p}^{(l+1)} = \sigma(\sigma^{-1}(\mathbf{p}^l) + \Delta\mathbf{p}^l)$ with $\Delta\mathbf{p}^l = \mathrm{MLP}(\mathbf{C}_p^{(l+1)}) \in \mathbb{R}^M$ and the sigmoid function $\sigma(\cdot)$.

Boundary updating pathway. After the anchor update, we refine the boundaries of the predictions. It is widely perceived that a model needs to focus on fine-grained features in the neighborhood rather than far ones to adjust temporal boundaries [27, 29, 51]. Inspired by this, we devise a boundary-focused attention layer (Fig. 3). For brevity, we explain the process of starting boundary update.

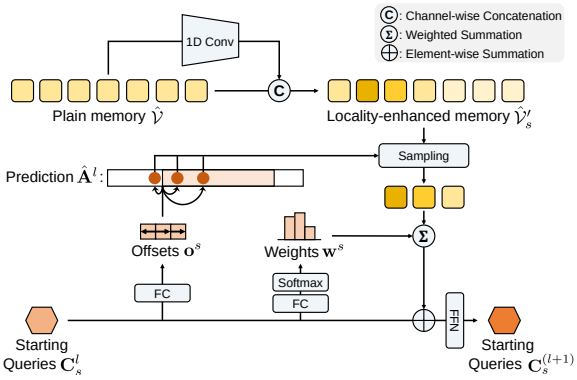


Fig. 3: Boundary-focused attention layer for starting queries.

The plain memory $\hat{\mathcal{V}}$ lacks the inductive bias due to the property of attentional layers [8]. Thus we first build locality-enhanced memory features for effective boundary refinement. To this goal, we obtain boundary-sensitive features with several 1D convolutional layers, *i.e.*, $\hat{\mathcal{V}}_s = f_s(\hat{\mathcal{V}})$. Then we encourage them to highly activate around the starting position of target moments. In detail, we impose regularization on the clip-wise activation scores obtained by channel-mean, *i.e.*, $\hat{g}^s = \text{mean}(\sigma(\hat{\mathcal{V}}_s)) \in \mathbb{R}^{N_v}$. The regularization loss is defined as:

$$\mathcal{L}^s_{\text{regul}} = -\frac{1}{N_v}\sum_{i=1}^{N_v}\big(g_i^s \log(\hat{g}_i^s) + (1-g_i^s)\log(1-\hat{g}_i^s)\big), \tag{5}$$

where g_i^s is the binary label obtained as $g_i^s = \mathbb{1}[i \in \mathcal{B}^s]$, where $\mathcal{B}^s$ is the neighbor clip set around starting points with a radius of r^s (set to 1/10 of the moment length). We merge boundary-sensitive features with the plain ones to form locality-enhanced memory, *i.e.*, $\hat{\mathcal{V}}'_s = [\hat{\mathcal{V}} \parallel \hat{\mathcal{V}}_s] \in \mathbb{R}^{N_v \times 2D}$. We can obtain $\mathcal{L}^e_{\text{regul}}$ and $\hat{\mathcal{V}}'_e$ in the same way. The total regularization term is $\mathcal{L}_{\text{regul}} = \mathcal{L}^s_{\text{regul}} + \mathcal{L}^e_{\text{regul}}$.

Given the locality-enhanced memory, the boundary queries need to capture fine-grained details around the boundaries for refinement. To efficiently aggregate useful features near the boundaries, we employ deformable attention [77, 78]. Regarding the starting boundary, *i.e.*, $\mathbf{p}^{(l+1)} - \mathbf{d}_s^l$, as the origin, we predict offsets and weights to select K neighbors, *i.e.*, $\mathbf{o}^s = \phi_o(\mathbf{C}_s^l) \in \mathbb{R}^{M\times K}$, $\mathbf{w}^s = \text{softmax}(\phi_w(\mathbf{C}_s^l)) \in \mathbb{R}^{M\times K}$, where ϕ_* are fully-connected layers. Features from the sampled neighbors are then aggregated into the starting queries as:

$$\hat{\mathbf{C}}_s^l = \sum_{k=1}^{K}\big[\mathbf{w}_k^s \cdot \hat{\mathcal{V}}'_s[\mathbf{p}^{(l+1)} - \mathbf{d}_s^l + \mathbf{o}_k^s]\big] + \mathbf{C}_s^l, \tag{6}$$

where we denote the sampling process from memory by $\hat{\mathcal{V}}'_s[\cdot]$. Lastly, we adopt a feed-forward network to obtain the updated starting queries, *i.e.*, $\mathbf{C}_s^{(l+1)} = \text{FFN}(\hat{\mathbf{C}}_s^l)$. We also obtain $\mathbf{C}_e^{(l+1)}$ in the same way. With the updated queries, we refine the boundaries to be better aligned with those of ground truths using sigmoid-based refinement as similar in the anchor update, leading to the refined

predictions $\mathbf{A}^{(l+1)} = [\mathbf{p}^{(l+1)}; \mathbf{d}_s^{(l+1)}; \mathbf{d}_e^{(l+1)}]$. Note that we utilize the deformable attention for the purpose of local feature aggregation, which sharply differs from the original purpose of efficient multi-scale global operation [78]. We provide comparison experiments regarding their roles in Appendix.

Moment prediction. We repeatedly update the predictions through a total of L_D dual-pathway decoding layers. We denote the resulting predictions by $\mathbf{A} = \big[\mathbf{p}; \mathbf{d}_s; \mathbf{d}_e\big]$, the anchor queries by $\mathbf{C}_p$, and the boundary queries by $\mathbf{C}_s$ and $\mathbf{C}_e$. Then we cast the predictions in the form of starting and ending timestamps, *i.e.*, $\hat{\varphi} = \big[\mathbf{p} - \mathbf{d}_s; \mathbf{p} + \mathbf{d}_e\big] \in \mathbb{R}^{M\times 2}$, which serves as the final results.

3.5 Quality-based Scoring

After producing the moment predictions, we opt to rank them for evaluation. In the convention, query-based models utilize classification scores as the measure, which exhibits how well the proposals semantically match the sentence. However, it does not necessarily represent the localization qualities of proposals. Hence we propose to estimate the localization quality of each moment prediction. Formally, the quality score can be derived as $\mathbf{q} = \sigma(\text{MLP}([\mathbf{C}_p \parallel \mathbf{C}_s \parallel \mathbf{C}_e])) \in \mathbb{R}^M$, where σ is sigmoid activation. Then the quality loss is defined as follows.

$$\mathcal{L}_{\text{qual}} = \sum_{m=1}^{M} \Big| q_m - \max_{\forall n} \big(\frac{|\hat{\varphi}_m \cap \varphi_n|}{|\hat{\varphi}_m \cup \varphi_n|}\big) \Big|, \tag{7}$$

where the objective of the quality head is to predict the maximum IoUs of the proposals with ground-truth moments.

3.6 Matching

As in the standard of query-based models [3, 21], we perform Hungarian matching [19] between predictions and ground truths. The optimal matching results ψ^* can be derived as follows.

$$\begin{aligned} \psi^* &= \arg\min\nolimits_{\psi\in\mathfrak{G}_N} \sum_{n=1}^{N} \mathcal{C}(\varphi_n, \hat{\varphi}_{\psi(n)}), \\ \mathcal{C}(\varphi_n, \hat{\varphi}_{\psi(n)}) &= \lambda_{l_1}\mathcal{L}_{l_1}(\varphi_n, \hat{\varphi}_{\psi(n)}) + \lambda_{\text{iou}}\mathcal{L}_{\text{iou}}(\varphi_n, \hat{\varphi}_{\psi(n)}), \end{aligned} \tag{8}$$

where $\mathfrak{G}_N$ denotes the combination pool and $\psi(n)$ is the index of the prediction matched by the n-th ground truth. $\mathcal{L}_{l_1}$ and $\mathcal{L}_{\text{iou}}$ respectively represent the L_1 distance and the generalized IoU [46] between the moments, while λ_{l_1} and λ_{iou} are their weights. Note that in contrast to other works, the classification term is not involved in the matching process, leading to localization-oriented matching.

Once the matching is completed, we minimize the matching cost between each pair of the matching results ψ^*. The localization loss is defined as:

$$\mathcal{L}_{\text{loc}} = \sum_{n=1}^{N} \mathcal{C}(\varphi_n, \hat{\varphi}_{\psi^*(n)}). \tag{9}$$

Overall training objectives. Our model is trained in an end-to-end fashion and the overall training objective is defined as follows.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{loc}} + \lambda_{\text{qual}}\mathcal{L}_{\text{qual}} + \lambda_{\text{sal}}\mathcal{L}_{\text{sal}} + \lambda_{\text{regul}}\mathcal{L}_{\text{regul}}, \tag{10}$$

where λ_* are the balancing parameters.

4 Experiments

4.1 Experimental Settings

Datasets. QVHighlights [21] is a recently built dataset, containing a total of 10,148 videos and 10,310 sentences from vlog and news domains. In addition to moments, it provides segment-level saliency score annotations within each moment. Importantly, it allows a single sentence corresponding to multiple disjoint moments (1.8 on average). Due to this practical setup, we utilize QVHighlights as the main benchmark. Charades-STA [11] includes 16,128 sentence-moment pairs with 9,848 indoor videos. The average duration of videos and moments is 30.6 and 8.1 seconds, respectively. TACoS [45] contains 127 cooking videos encompassing a total of 18,818 sentence-moment pairs. This dataset is known to be challenging since the moments occupy only a small portion (6.1 sec on average) within considerably long videos (4.8 min on average).

Evaluation metrics. Following the standard protocol, we measure the Recall@1 (R1) under the IoU thresholds of 0.3, 0.5, and 0.7 by default. Since QVHighlights contains multiple ground-truth moments per sentence, we report the mean average precision (mAP) with IoU thresholds of 0.5 and 0.75 as well as the average mAP over a set of IoU thresholds [0.5:0.05:0.95]. Meanwhile, we compute the mean IoU of top-1 predictions on Charades-STA and TACoS. Note that the performances at high IoU thresholds (*e.g.*, 0.7) exhibit how well the predictions align with the ground truths.

4.2 Implementation Details

For a fair comparison, we adopt the same feature extraction strategy with the competitors [21,26,41]. Specifically, we adopt the CLIP [44] text features for text and the concatenation of Slowfast [10] (ResNet-50) and CLIP [44] (ViT-B/32) features for videos unless otherwise specified. The video features are extracted every 1 second for Charades-STA and 2 seconds for QVHighlihgts and TACoS. To compare with audio-augmented models [37], we optionally employ audio features extracted by PANNs [18] pre-trained on AudioSet [14]. Due to the long video duration, we uniformly sample 200 feature vectors from each video for TACoS.

We set the embedding dimension D to 256, the number of attention heads to 8, the number of queries M to 10, the number of boundary points K to 3, and the margin α to 0.2. We determine the numbers of encoding and decoding layers as same with the prior work [41], *i.e.*, $L_E = L_D = 2$. The balancing parameters are set as: $\lambda_{l1} = 10$, $\lambda_{\text{iou}} = \lambda_{\text{sal}} = \lambda_{\text{regul}} = 1$, $\lambda_{\text{qual}} = 2$. As in previous works [26,41], we increase λ_{sal} to 4 when saliency labels are unavailable, *i.e.*, for Charades-STA

Table 2: Results on the QVHighlights test split.

Method	R1		mAP		
	@0.5	@0.7	@0.5	@0.75	Avg.
w/o pre-training					
MCN [1]	11.41	2.72	24.94	8.22	10.67
CAL [9]	25.49	11.54	23.40	7.65	9.89
CLIP [44]	16.88	5.19	18.11	7.00	7.67
XML [22]	46.69	33.46	47.89	34.67	34.90
Moment-DETR [21]	52.89	33.02	54.82	29.40	30.73
UMT† [37]	56.23	41.18	53.83	37.01	36.12
MH-DETR [60]	60.05	42.48	60.75	38.13	38.38
QD-DETR [41]	62.40	44.98	62.52	39.88	39.86
QD-DETR† [41]	63.06	45.10	63.04	40.10	40.19
UniVTG [28]	58.86	40.86	57.60	35.59	35.47
EaTR‡ [17]	57.98	42.41	59.95	39.29	39.00
MomentDiff [26]	57.42	39.66	54.02	35.73	35.95
BAM-DETR	**62.71**	**48.64**	**64.57**	**46.33**	**45.36**
BAM-DETR†	**64.07**	**48.12**	**65.61**	**47.51**	**46.91**
w/ pre-training on 4.2M data labeled by CLIP					
UniVTG [28]	65.43	50.06	64.06	45.02	43.63
w/ pre-training on 236K ASR captions					
Moment-DETR [21]	59.78	40.33	60.51	35.36	36.14
UMT† [37]	60.83	43.26	57.33	39.12	38.08
QD-DETR [41]	63.18	45.19	63.37	40.35	39.96
BAM-DETR	**63.88**	**47.92**	**66.33**	**48.22**	**46.67**

†additional use of audio modality ‡reproduced by official checkpoint

and TACoS. Our model is trained from scratch for 200 epochs on QVHighlights and 100 epochs on the other datasets using the AdamW optimizer [38] with a learning rate of 1e-4 and a batch size of 32.

4.3 Comparison with State-of-the-arts

Results on QVHighlights. We compare our model with existing state-of-the-arts including recent query-based approaches [17, 21, 26, 37, 41, 60] on the test split. As shown in Table 2, our BAM-DETR consistently outperforms the comparative models under various settings. In detail, without pre-training, our model surpasses the previous state-of-the-art model [41] by large margins, *e.g.*, 3.66% in R1@0.7 and 6.45% in mAP@0.75. These improvements under strict IoU thresholds verify the superior localization ability of our method. When leveraging auxiliary audio features, the performance further boosts especially in R1@0.5 and mAPs, enlarging the gap between the competitors including those with audio features [37,41]. To compare with the methods with pretraining, we pretrain our model on middle-scale ASR caption data [21]. Again, our BAM-DETR achieves state-of-the-art results, while showing the least gap with the method [28] that leverages large-scale data for pre-training. Notably, our model even outperforms it in terms of mAPs with much fewer (about 18×) pre-training data, manifesting the effectiveness of the proposed methods.

Results on Charades-STA. Experimental results on the test split are shown in Table 3. Not only does our BAM-DETR outperform the query-based com-

Table 3: Results on the Charades-STA and TACoS test splits.

Method	Charades-STA				TACoS			
	R1@0.3	R1@0.5	R1@0.7	mIoU	R1@0.3	R1@0.5	R1@0.7	mIoU
2D-TAN [72]	58.76	46.02	27.40	41.25	40.01	27.99	12.92	27.22
VSLNet [70]	60.30	42.69	24.14	41.58	35.54	23.54	13.15	24.99
Moment-DETR [21]	65.83	52.07	30.59	45.54	37.97	24.67	11.97	25.49
QD-DETR [41]	-	57.31	32.55	-	52.39	36.77	21.07	35.76
UniVTG [28]	70.81	58.01	35.65	50.10	51.44	34.97	17.35	33.60
MomentDiff [26]	-	55.57	32.42	-	46.64	28.92	12.37	30.36
BAM-DETR	**72.93**	**59.95**	**39.38**	**52.33**	**56.69**	**41.54**	**26.77**	**39.31**

Table 4: Results on the anti-biased Charades-STA test split against the moment location and length. VGG [52] and Glove [43] features are employed for all models.

Method	w.r.t. moment location				w.r.t. moment length			
	R1@0.3	R1@0.5	R1@0.7	mAP_{avg}	R1@0.3	R1@0.5	R1@0.7	mAP_{avg}
2D-TAN [72]	27.81	20.44	10.84	17.23	39.68	28.68	17.72	22.79
MMN [56]	33.58	27.20	14.12	19.18	43.58	34.31	19.94	26.85
Moment-DETR [21]	29.94	22.16	11.56	18.66	42.73	34.39	16.12	24.02
QD-DETR‡ [41]	56.17	46.82	28.13	30.70	67.39	54.44	32.87	36.99
MomentDiff [26]	48.39	33.59	15.71	21.37	51.25	38.32	23.38	28.19
BAM-DETR	**59.83**	**50.00**	**32.08**	**31.68**	**68.40**	**55.46**	**40.74**	**43.21**

‡reproduced by official codebase

petitors [21, 26, 41], but it achieves a new state-of-the-art by surpassing the best performing anchor-free model [28] for all metrics. Notably, a large gap of 3.73% is observed in R1@0.7, which confirms the strong localization ability of our model.

Results on TACoS. We present the comparison results on the test set in Table 3. It can be observed that our BAM-DETR achieves a new state-of-the-art with pronounced performances under the strict IoU thresholds, which is consistent with the above results on other datasets. On this challenging benchmark, our model surpasses the previous best model [41] by 5.7% (relatively 27%) in R1@0.7. These results clearly exhibit the superiority of the proposed model.

4.4 Robustness Evaluation

Query-based models potentially have a temporal bias [16, 63] against the locations and lengths of moments. To measure robustness, we evaluate our model on the anti-biased Charades-STA [26] with distribution shifts of the moment location and length between training and test sets. Table 4 summarizes the results, where our model outperforms all competitors under both anti-biased settings. Especially, it shows significant performance gaps under the moment length bias. This can be expected since our model directly localizes boundaries instead of predicting lengths, which lessens the effect of bias. This robustness test corroborates the advantage of our boundary-oriented moment modeling.

4.5 Analysis

We conduct analytical experiments on the QVHighlights validation split.

Table 5: Ablation study of components on QVHighlights.

Method	R1		mAP		
	@0.5	@0.7	@0.5	@0.75	Avg.
Baseline	62.39	47.87	62.64	41.54	41.75
+ boundary-oriented modeling	63.42	49.23	62.86	43.24	42.42
+ dual-pathway decoder	63.61	50.26	63.01	44.98	44.16
+ quality-based scoring	65.10	51.61	65.41	48.56	47.61

Table 6: Ablative experiments on QVHighlights. Default settings are marked gray.

(a) Number of predictions

M	R1		mAP
	@0.5	@0.7	Avg.
5	62.19	49.03	45.18
10	65.10	51.61	47.61
15	65.48	50.32	48.23
20	65.23	51.61	48.01

(b) Choice of memory features

Memory	R1		mAP
	@0.5	@0.7	Avg.
plain	63.74	49.68	46.21
sensitive	63.81	49.87	46.34
merged	65.10	51.61	47.61

(c) Number of boundary points

K	R1		mAP
	@0.5	@0.7	Avg.
1 (fixed)	63.23	48.32	45.17
1	64.32	49.87	47.14
3	65.10	51.61	47.61
5	64.90	50.45	46.70

(d) Query choices for quality prediction

$\mathbf{C}_p$	$\mathbf{C}_s$	$\mathbf{C}_e$	R1		mAP
			@0.5	@0.7	Avg.
✓			62.97	49.03	46.26
	✓		64.39	50.71	46.66
		✓	64.58	49.23	46.71
✓	✓	✓	65.10	51.61	47.61

(e) Combinations of attention layers for updating pathways

Anchor query	Boundary query	R1		mAP	FLOPs	Params
		@0.5	@0.7	Avg.		
Global (shared)		62.26	49.03	44.18	0.71G	8.2M
Global	Global	63.61	49.81	44.54	0.72G	11.5M
Global	Focused	63.74	49.68	46.21	0.65G	9.5M
Focused	Focused	62.48	49.16	45.87	0.62G	7.6M

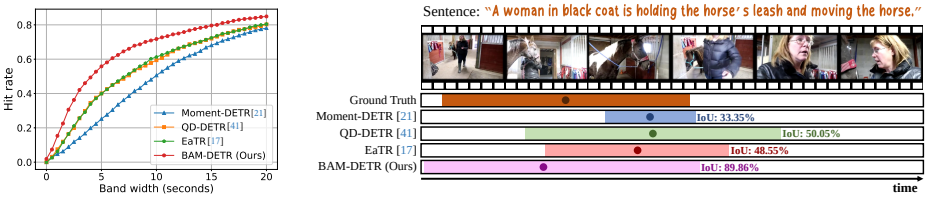
Effect of each component. We analyze the effect of each component in Table 5. The direct adoption of our moment modeling solely improves the performance, particularly at strict thresholds, showcasing its advantage in boundary alignment. Employing the dual-pathway decoder leads to further gains, which suggests the essential role of separate pathways. Lastly, the quality-based scoring considerably elevates the scores, especially in terms of mAPs.

Number of predictions. We experiment with varying numbers of predictions in Table 6a, where the model achieves robust results when M is sufficiently large. By default, we set M to 10 for a fair comparison with the previous works [21,41].

Locality-enhanced features. We analyze the effect of the choice of input features for boundary-focused attention in Table 6b. The results indicate that boundary-sensitive features are slightly more helpful than plain ones in precise localization. In addition, the merged features achieve the best performance.

Number of sampled boundary points. We analyze the effect of the number of sampled boundary points K in Table 6c. For comparisons, we report the case where the fixed boundary features are sampled without using offsets (1^{st} row). As shown in the table, the dynamic selection of neighborhoods rather than fixed boundaries is important for accurate moment localization. The performance improves when sampling multiple points, while it saturates at $K = 3$.

Query choices for quality prediction. Our model leverages both anchor and boundary queries for quality prediction. We investigate the effect of query choices in Table 6d. As a result, utilizing all the queries shows better localization



(a) Boundary hit rate **(b)** Visualization results

Fig. 4: Analytical experiments on QVHighlights.

performance than using one of them. This might be trivial as different queries contain complementary information for quality prediction.

Attention layers. Table 6e compares different combinations of attention layers for updating pathways in terms of performances and costs. For an apple-to-apple comparison, we leverage the plain memory features for both attention layers. First of all, employing separate global cross-attention for different queries leads to a huge parameter increase yet limited score gains. Replacing the global one for boundary queries with our focused attention layer substantially reduces the cost, while achieving the best performance. Meanwhile, the use of focused attention for both types of queries leads to inferior results.

Boundary alignment. Standard IoU-based metrics are indirect measures for boundary alignment. To precisely diagnose the ability, inspired by Deeplab [5], we compute the boundary hit rate of predictions with varying band widths. We expand ground-truth boundary points with a band width to form starting/ending zones and regard a prediction as correct if both of its boundaries fall within the corresponding zones. More details can be found in Appendix. To disentangle the effect of ranking, we mark a video as correct if at least one prediction is correct and measure the video-level hit rate. As shown in Fig. 4a, our model greatly outperforms the recent competitors. The sharp increase in low band widths clearly validates the superiority of our model in boundary alignment.

Qualitative results. As shown in Fig. 4b, previous models fail to localize accurate moments with misleading center predictions. Especially, QD-DETR [41] accurately predicts the moment length but suffers from the misaligned center, leading to the limited IoU. In contrast, thanks to the novel moment modeling, our model predicts well-aligned boundaries without relying on center prediction.

5 Conclusion

In this paper, we identified the center misalignment issue of existing query-based models for sentence grounding. To address it, we presented boundary-oriented moment modeling where boundaries are directly predicted without relying on centers. Based on the modeling, we designed the boundary-aligned moment detection transformer characterized by dual-pathway decoding. Further, we proposed localization quality-based scoring of predictions. The efficacy of the proposed methods is validated by thorough examinations. We hope this work sheds light on the issue of center-based moment modeling in detection transformers.

Acknowledgements

This project was supported by the National Research Foundation of Korea grant funded by the Korea government (MSIT) (No. 2022R1A2B5B02001467; RS-2024-00346364).

References

1. Anne Hendricks, L., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.: Localizing moments in video with natural language. In: ICCV. pp. 5803–5812 (2017) 1, 11
2. Cao, M., Chen, L., Shou, M.Z., Zhang, C., Zou, Y.: On pursuit of designing multimodal transformer for video grounding. In: EMNLP. pp. 9810–9823 (2021) 4
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV. pp. 213–229. Springer (2020) 2, 4, 6, 9
4. Chen, J., Chen, X., Ma, L., Jie, Z., Chua, T.S.: Temporally grounding natural sentence in video. In: EMNLP. pp. 162–171 (2018) 4
5. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. PAMI **40**(4), 834–848 (2017) 14
6. Chen, L., Lu, C., Tang, S., Xiao, J., Zhang, D., Tan, C., Li, X.: Rethinking the bottom-up framework for query-based video localization. In: AAAI. vol. 34, pp. 10551–10558 (2020) 4
7. Dai, X., Chen, Y., Yang, J., Zhang, P., Yuan, L., Zhang, L.: Dynamic detr: End-to-end object detection with dynamic attention. In: ICCV. pp. 2988–2997 (2021) 4
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021) 8
9. Escorcia, V., Soldan, M., Sivic, J., Ghanem, B., Russell, B.: Temporal localization of moments in video collections with natural language. arXiv preprint arXiv:1907.12763 (2019) 1, 11
10. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: ICCV. pp. 6202–6211 (2019) 10
11. Gao, J., Sun, C., Yang, Z., Nevatia, R.: Tall: Temporal activity localization via language query. In: CVPR. pp. 5267–5275 (2017) 1, 4, 10
12. Gao, P., Zheng, M., Wang, X., Dai, J., Li, H.: Fast convergence of detr with spatially modulated co-attention. In: ICCV. pp. 3621–3630 (2021) 4
13. Ge, R., Gao, J., Chen, K., Nevatia, R.: Mac: Mining activity concepts for language-based temporal localization. In: WACV. pp. 245–253. IEEE (2019) 4
14. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: ICASSP. pp. 776–780. IEEE (2017) 10
15. Ghosh, S., Agarwal, A., Parekh, Z., Hauptmann, A.G.: Excl: Extractive clip localization using natural language descriptions. In: NAACL. pp. 1984–1990 (2019) 4
16. Hao, J., Sun, H., Ren, P., Wang, J., Qi, Q., Liao, J.: Can shuffling video benefit temporal bias problem: A novel training framework for temporal grounding. In: ECCV. pp. 130–147. Springer (2022) 12

17. Jang, J., Park, J., Kim, J., Kwon, H., Sohn, K.: Knowing where to focus: Event-aware transformer for video grounding. In: ICCV. pp. 13846–13856 (2023) 2, 4, 6, 7, 11
18. Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., Plumbley, M.D.: Panns: Large-scale pretrained audio neural networks for audio pattern recognition. IEEE/ACM Trans. on Audio, Speech, and Language Processing **28**, 2880–2894 (2020) 10
19. Kuhn, H.W.: The hungarian method for the assignment problem. Naval research logistics quarterly **2**(1-2), 83–97 (1955) 9
20. Lee, P., Kim, T., Shim, M., Wee, D., Byun, H.: Decomposed cross-modal distillation for rgb-based temporal action detection. In: CVPR. pp. 2373–2383 (2023) 1
21. Lei, J., Berg, T.L., Bansal, M.: Detecting moments and highlights in videos via natural language queries. In: Neurips. vol. 34, pp. 11846–11858 (2021) 2, 3, 4, 5, 6, 9, 10, 11, 12, 13
22. Lei, J., Yu, L., Berg, T.L., Bansal, M.: Tvr: A large-scale dataset for video-subtitle moment retrieval. In: ECCV. pp. 447–463. Springer (2020) 11
23. Li, F., Zeng, A., Liu, S., Zhang, H., Li, H., Zhang, L., Ni, L.M.: Lite detr: An interleaved multi-scale encoder for efficient detr. In: CVPR. pp. 18558–18567 (2023) 4
24. Li, F., Zhang, H., Liu, S., Guo, J., Ni, L.M., Zhang, L.: Dn-detr: Accelerate detr training by introducing query denoising. In: CVPR. pp. 13619–13627 (2022) 4
25. Li, H., Cao, M., Cheng, X., Li, Y., Zhu, Z., Zou, Y.: G2l: Semantically aligned and uniform video grounding via geodesic and game theory. In: ICCV. pp. 12032–12042 (2023) 4
26. Li, P., Xie, C.W., Xie, H., Zhao, L., Zhang, L., Zheng, Y., Zhao, D., Zhang, Y.: Momentdiff: Generative video moment retrieval from random to real. In: Neurips (2023) 2, 4, 6, 10, 11, 12
27. Lin, C., Xu, C., Luo, D., Wang, Y., Tai, Y., Wang, C., Li, J., Huang, F., Fu, Y.: Learning salient boundary feature for anchor-free temporal action localization. In: CVPR. pp. 3320–3329 (2021) 7
28. Lin, K.Q., Zhang, P., Chen, J., Pramanick, S., Gao, D., Wang, A.J., Yan, R., Shou, M.Z.: Univtg: Towards unified video-language temporal grounding. In: ICCV. pp. 2794–2804 (2023) 4, 5, 6, 11, 12
29. Lin, T., Zhao, X., Su, H., Wang, C., Yang, M.: Bsn: Boundary sensitive network for temporal action proposal generation. In: ECCV. pp. 3–19 (2018) 1, 7
30. Lin, Y., Yuan, Y., Zhang, Z., Li, C., Zheng, N., Hu, H.: Detr does not need multi-scale or locality design. In: ICCV. pp. 6545–6554 (2023) 4
31. Liu, D., Qu, X., Dong, J., Zhou, P., Cheng, Y., Wei, W., Xu, Z., Xie, Y.: Context-aware biaffine localizing network for temporal sentence grounding. In: CVPR. pp. 11235–11244 (2021) 1
32. Liu, D., Qu, X., Dong, J., Zhou, P., Cheng, Y., Wei, W., Xu, Z., Xie, Y.: Context-aware biaffine localizing network for temporal sentence grounding. In: CVPR. pp. 11235–11244 (2021) 4
33. Liu, F., Wei, H., Zhao, W., Li, G., Peng, J., Li, Z.: Wb-detr: transformer-based detector without backbone. In: ICCV. pp. 2979–2987 (2021) 4
34. Liu, M., Wang, X., Nie, L., Tian, Q., Chen, B., Chua, T.S.: Cross-modal moment localization in videos. In: ACM MM. pp. 843–851 (2018) 4
35. Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J., Zhang, L.: Dab-detr: Dynamic anchor boxes are better queries for detr. In: ICLR (2022) 4, 7
36. Liu, S., Ren, T., Chen, J., Zeng, Z., Zhang, H., Li, F., Li, H., Huang, J., Su, H., Zhu, J., Zhang, L.: Detection transformer with stable matching. In: ICCV. pp. 6491–6500 (2023) 4

37. Liu, Y., Li, S., Wu, Y., Chen, C.W., Shan, Y., Qie, X.: Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In: CVPR. pp. 3042–3051 (2022) 2, 4, 5, 10, 11
38. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019) 11
39. Lu, C., Chen, L., Tan, C., Li, X., Xiao, J.: Debug: A dense bottom-up grounding approach for natural language video localization. In: EMNLP-IJCNLP. pp. 5144–5153 (2019) 4
40. Meng, D., Chen, X., Fan, Z., Zeng, G., Li, H., Yuan, Y., Sun, L., Wang, J.: Conditional detr for fast training convergence. In: ICCV. pp. 3651–3660 (2021) 4, 7
41. Moon, W., Hyun, S., Park, S., Park, D., Heo, J.P.: Query-dependent video representation for moment retrieval and highlight detection. In: CVPR. pp. 23023–23033 (2023) 2, 3, 4, 6, 7, 10, 11, 12, 13, 14
42. Mun, J., Cho, M., Han, B.: Local-global video-text interactions for temporal grounding. In: CVPR. pp. 10810–10819 (2020) 4, 5
43. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: EMNLP. pp. 1532–1543 (2014) 12
44. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PMLR (2021) 10, 11
45. Regneri, M., Rohrbach, M., Wetzel, D., Thater, S., Schiele, B., Pinkal, M.: Grounding action descriptions in videos. Trans. Assoc. Comput. Linguistics **1**, 25–36 (2013) 10
46. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: CVPR. pp. 658–666 (2019) 9
47. Roh, B., Shin, J., Shin, W., Kim, S.: Sparse detr: Efficient end-to-end object detection with learnable sparsity. In: ICLR (2022) 4
48. Rui, Y., Gupta, A., Acero, A.: Automatically extracting highlights for tv baseball programs. In: ACM MM. pp. 105–115 (2000) 1
49. Shao, D., Xiong, Y., Zhao, Y., Huang, Q., Qiao, Y., Lin, D.: Find and focus: Retrieve and localize video events with natural language queries. In: ECCV. pp. 200–216 (2018) 1
50. Sharghi, A., Laurel, J.S., Gong, B.: Query-focused video summarization: Dataset, evaluation, and a memory network based approach. In: CVPR. pp. 4788–4797 (2017) 1
51. Shi, D., Zhong, Y., Cao, Q., Ma, L., Li, J., Tao, D.: Tridet: Temporal action detection with relative boundary modeling. In: CVPR. pp. 18857–18866 (2023) 7
52. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015) 12
53. Sun, Z., Cao, S., Yang, Y., Kitani, K.M.: Rethinking transformer-based set prediction for object detection. In: ICCV. pp. 3611–3620 (2021) 4
54. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Neurips. vol. 30 (2017) 6
55. Wang, Y., Zhang, X., Yang, T., Sun, J.: Anchor detr: Query design for transformer-based detector. In: AAAI. vol. 36, pp. 2567–2575 (2022) 4
56. Wang, Z., Wang, L., Wu, T., Li, T., Wu, G.: Negative sample matters: A renaissance of metric learning for temporal grounding. In: AAAI. vol. 36, pp. 2613–2623 (2022) 4, 12

57. Xiao, S., Chen, L., Zhang, S., Ji, W., Shao, J., Ye, L., Xiao, J.: Boundary proposal network for two-stage natural language video localization. In: AAAI. vol. 35, pp. 2986–2994 (2021) 4
58. Xiong, B., Kalantidis, Y., Ghadiyaram, D., Grauman, K.: Less is more: Learning highlight detection from video duration. In: CVPR. pp. 1258–1267 (2019) 1
59. Xu, H., He, K., Plummer, B.A., Sigal, L., Sclaroff, S., Saenko, K.: Multilevel language and vision integration for text-to-clip retrieval. In: AAAI. vol. 33, pp. 9062–9069 (2019) 4
60. Xu, Y., Sun, Y., Li, Y., Shi, Y., Zhu, X., Du, S.: Mh-detr: Video moment and highlight detection with cross-modal transformer. In: ACM MM (2023) 4, 11
61. Yan, S., Xiong, X., Nagrani, A., Arnab, A., Wang, Z., Ge, W., Ross, D., Schmid, C.: Unloc: A unified framework for video localization tasks. In: ICCV. pp. 13623–13633 (2023) 4
62. Ye, M., Ke, L., Li, S., Tai, Y.W., Tang, C.K., Danelljan, M., Yu, F.: Cascade-detr: Delving into high-quality universal object detection. In: ICCV. pp. 6704–6714 (2023) 4
63. Yuan, Y., Lan, X., Wang, X., Chen, L., Wang, Z., Zhu, W.: A closer look at temporal sentence grounding in videos: Dataset and metric. In: Proc. 2nd Int. Workshop on Human-Centric Multimedia Analysis. pp. 13–21 (2021) 12
64. Yuan, Y., Ma, L., Wang, J., Liu, W., Zhu, W.: Semantic conditioned dynamic modulation for temporal sentence grounding in videos. In: Neurips. vol. 32 (2019) 1, 4
65. Yuan, Y., Mei, T., Zhu, W.: To find where you talk: Temporal sentence localization in video with attention based location regression. In: AAAI. vol. 33, pp. 9159–9166 (2019) 4
66. Zeng, R., Xu, H., Huang, W., Chen, P., Tan, M., Gan, C.: Dense regression network for video grounding. In: CVPR. pp. 10287–10296 (2020) 4
67. Zhang, D., Dai, X., Wang, X., Wang, Y.F., Davis, L.S.: Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In: CVPR. pp. 1247–1257 (2019) 4
68. Zhang, G., Luo, Z., Yu, Y., Cui, K., Lu, S.: Accelerating detr convergence via semantic-aligned matching. In: CVPR. pp. 949–958 (2022) 4
69. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In: ICLR (2023) 4
70. Zhang, H., Sun, A., Jing, W., Zhou, J.T.: Span-based localizing network for natural language video localization. In: ACL (2020) 4, 5, 12
71. Zhang, K., Chao, W.L., Sha, F., Grauman, K.: Video summarization with long short-term memory. In: ECCV. pp. 766–782. Springer (2016) 1
72. Zhang, S., Peng, H., Fu, J., Luo, J.: Learning 2d temporal adjacent networks for moment localization with natural language. In: AAAI. vol. 34, pp. 12870–12877 (2020) 1, 4, 5, 12
73. Zhang, S., Su, J., Luo, J.: Exploiting temporal relationships in video moment localization with natural language. In: ACM MM. pp. 1230–1238 (2019) 4
74. Zhang, Z., Lin, Z., Zhao, Z., Xiao, Z.: Cross-modal interaction networks for query-based moment retrieval in videos. In: ACM SIGIR. pp. 655–664 (2019) 4
75. Zhao, Y., Xiong, Y., Wang, L., Wu, Z., Tang, X., Lin, D.: Temporal action detection with structured segment networks. In: ICCV. pp. 2914–2923 (2017) 1
76. Zheng, D., Dong, W., Hu, H., Chen, X., Wang, Y.: Less is more: Focus attention for efficient detr. In: ICCV. pp. 6674–6683 (2023) 4

77. Zhu, X., Hu, H., Lin, S., Dai, J.: Deformable convnets v2: More deformable, better results. In: CVPR. pp. 9308–9316 (2019) 8
78. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. In: ICLR (2021) 4, 7, 8, 9