OVERCOMING WEAK VISUAL-TEXTUAL ALIGNMENT FOR VIDEO MOMENT RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Video moment retrieval (VMR) identifies a specific moment in an untrimmed video for a given natural language query. This task is prone to suffer the weak visual-textual alignment problem innate in the video datasets. Due to the ambiguity, a query does not fully cover the relevant details of the corresponding moment, or the moment may contain misaligned and irrelevant frames, potentially limiting further performance gains and generalization capability. To tackle this problem, we propose a background-aware moment detection transformer (BM-DETR). Our model adopts a contrastive approach, carefully utilizing the negative queries matched to other moments in the video. Specifically, our model learns to predict the target moment from the joint probability of each frame given the positive query and the complement of negative queries. This leads to effective use of the surrounding background, improving moment sensitivity and enhancing overall alignments in videos. Our approach is efficient and outperforms previous methods, including contrastive learning-based, on multiple datasets with significantly reduced computational costs.

1 INTRODUCTION

Video moment retrieval (VMR) (Gao et al., 2017) retrieves the target moment in an untrimmed video corresponding to a natural language query. A successful VMR model requires a comprehensive understanding of videos, language queries, and correlations to predict relevant moments precisely. In contrast to traditional action localization tasks (Yeung et al., 2016; Shou et al., 2016) that predict a fixed set of actions like "throwing" or "jumping," VMR is a more difficult task requiring joint comprehension of semantic meanings in video and language.

A video is typically composed of short video clips, where query sentences describe each clip. However, query sentences are often ambiguous as to whether they fully express the events occurring within the matching moment, and boundary annotations might include frames unrelated to the query sentences (Zhou et al., 2021; Huang et al., 2022). As shown in Figure 1, the moment prediction can be imprecise and weakly aligned with annotations. For instance, the query "Person pours some water into a glass" does not describe an event for "drink water", but the boundary annotation includes it. Furthermore, queries like "Person sitting on the sofa eating out of a dish" may confuse the model, as the actions "sitting" and the object "sofa" overlap throughout the video.

Some prior works (Zhang et al., 2020b;a; Mun et al., 2020; Zeng et al., 2020; Gao & Xu, 2021; Liu et al., 2021) take only a single query as input to predict the moment. However, solely relying on a single query may learn only local-level alignment and make it easy to overlook the global context due to weak alignment problems. Whereas, contrastive learning-based methods (Ding et al., 2021; Nan et al., 2021; Wang et al., 2022; Li et al., 2023) learn the query and the ground-truth moment features close to each other while keeping others apart (Oord et al., 2018). Nevertheless, due to semantic overlap and sparse annotation dilemma (Zeng et al., 2020) in videos, Li et al. (2023) claimed that adopting vanilla contrastive learning into VMR is suboptimal. Negative queries from random videos used to have semantic overlap, making them false negatives, while negative moments also likely are false negatives due to the sparse annotation. Because of this, negative sampling from the same video limits the number of negative samples, inducing inaccurate estimation of marginal distribution used in contrastive methods of InfoNCE (Oord et al., 2018). If negative queries from the same video

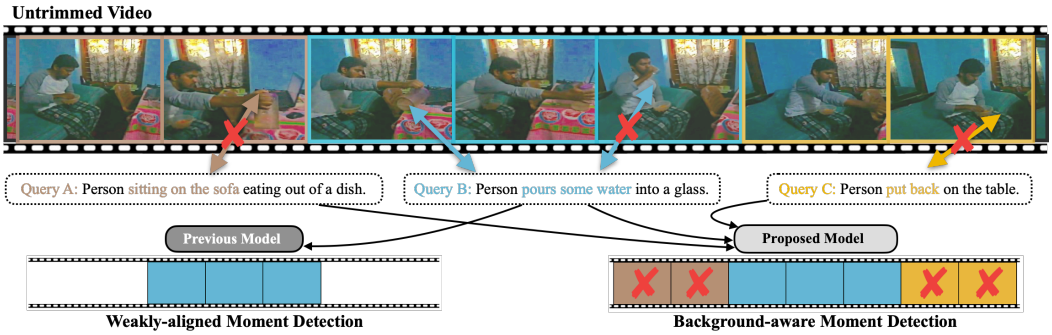

Figure 1: **Top:** We give an example of weak visual-textual alignment in the video. **Bottom:** We compare the current (left) and proposed (right) VMR approach.

semantically overlap with positive queries, vanilla contrastive learning wrongly forces them apart. To overcome this, Li et al. (2023) proposed a novel geodesic-guided contrastive learning scheme for VMR. Despite this, it still had to sample a large number of negative moments with varying lengths, leading to high computational costs to approximate the true partition and achieve sophisticated alignment faithfully.

In this paper, we propose a novel **B**ackground-aware **M**oment **DE**tection **TR**ansformer (BM-DETR). We utilize contexts outside of the target moments (*i.e.,*, negative queries) along with the positive query. Our model calculates the joint probability of each frame given a positive query and the complement of negative queries, resulting in frame attention scores for enhancing multimodal representations. By considering the relative relationships between queries within the video, the model learns how to best identify and focus on the relevant visual features of the target moment, improving *moment sensitivity*, or *true positive rate*. Then, we utilize cross-modal discrimination between other video-query pairs for fine-grained semantic alignment. Furthermore, we leverage a temporal shifting method as an auxiliary objective, improving the model's robustness to temporal content changes.

In contrast to previous approaches, which relied on a single query with complex multimodal reasoning or mining a multitude of negative moments with high cost, our model can attend to the target moment and be aware of the contextual meanings throughout the video. Moreover, our model is efficient and outperforms previous contrastive learning methods by eliminating dense moment features and reducing redundant computations.

To sum up, the contributions of our paper can be summarized as follows:

- We propose BM-DETR to mitigate the weak visual-textual alignment problem, which is crucial in video moment retrieval tasks.
- By considering temporal and contextual relationships within videos, our model can enhance overall alignment in videos and enable robust moment detection.
- Our model outperforms state-of-the-art methods in efficiency and transferability on four public datasets and two out-of-distribution VMR scenarios.

2 Related Work

Video moment retrieval. Video moment retrieval (VMR) aims to retrieve the target moment in a video based on a natural language sentence. Existing approaches are mainly classified into proposal-based methods and proposal-free methods. The proposal-based methods (Gao et al., 2017; Xu et al., 2019; Anne Hendricks et al., 2017; Chen et al., 2018; Zhang et al., 2019; 2020b; Gao & Xu, 2021; Wang et al., 2021; 2022) sample candidate moments from the video and select the most similar moment to the given query. In contrast, proposal-free methods (Yuan et al., 2019; He et al., 2019; Rodriguez et al., 2020; Chen et al., 2020; Zhang et al., 2020a; Mun et al., 2020; Zeng et al., 2020; Liu et al., 2021) regress target moments from video and language features without generating candidate moments. Breaking away from traditional paradigms, several studies (Lei et al., 2021; Cao et al., 2021; Woo et al., 2022; Liu et al., 2022b; Moon et al., 2023; Lin et al., 2023) utilized the DETR's

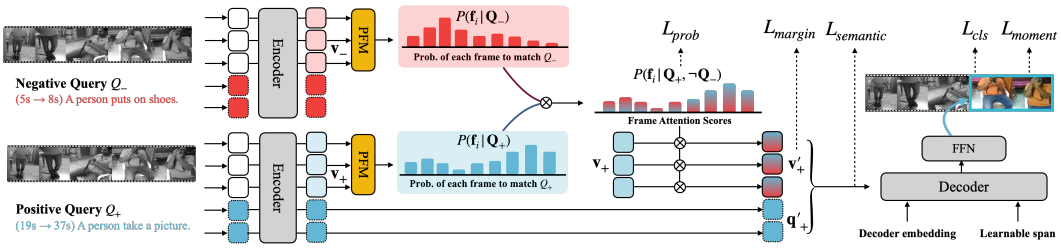

Figure 2: An overview of the proposed BM-DETR framework.

(Zhu et al., 2020) object detection ability for localization tasks. In this paper, our model also follows DETR's detection paradigm.

Visual-textual alignment problem in video domains. Labeling videos is expensive and cumbersome, making it difficult to build high-quality and scalable video datasets. This often leads to alignment issues, which have been observed (Miech et al., 2020; Ko et al., 2022; Han et al., 2022) as a crucial bottleneck of video understanding. VMR is sensitive to the above issues since it requires accurate temporal moment locations. To tackle these problems, a couple of studies (Zhou et al., 2021; Huang et al., 2022; Nan et al., 2021; Ding et al., 2021; Li et al., 2023) carefully designed their modeling to achieve sophisticated video alignments. For instance, Zhou et al. (2021) changed the phrases (e.g., verb) in language queries to improve semantic diversity, and Li et al. (2023) proposed a novel geodesic-guided contrastive learning scheme considering the semantic alignment and uniformity between video and query. In this paper, we propose background-aware moment detection to enhance overall alignments in videos.

3 METHOD

3.1 VIDEO MOMENT RETRIEVAL TASK

Given an untrimmed video V and language query Q, we represent the video as $V = \{f_i\}_{i=1}^{L_v}$ where f_i denotes the i-th frame. Likewise, the language query is denoted as $Q = \{w_i\}_{i=1}^{L_w}$ where w_i denotes the i-th word. L_v and L_w indicate the overall count of frames and words, respectively. We aim to localize the target moment $m = (t_s, t_e)$ in V from Q, where t_s and t_e represent the start and end times of the target moment, respectively.

3.2 BACKGROUND-AWARE MOMENT DETECTION

As mentioned earlier, a single query may not be sufficient to disambiguate the corresponding moment due to the weak alignment in videos. That said, predicting the target moment in V based solely on information from Q is less informative and ineffective, where the term "information" refers to the knowledge or cues used for accurate predictions of the target moment in V. Hence, we propose an alternative to resolve this problem inspired by *importance sampling* (Tokdar & Kass, 2010). Similar to the contrastive learning (Oord et al., 2018), a specific query Q_+ is designated as the target (positive), while we randomly sampled a negative query Q_- for each training step. Our main idea is based on two guiding principles, which are as follows:

Principle 1. *Queries from the same video V allow for disambiguation of the target query Q_+, as they have implicit contextual and temporal relationships with the corresponding video moments.*

Principle 2. *To avoid spurious correlations, we differentiate between negative query Q_- and target query Q_+ based on their temporal locations and semantic similarity.*

We use Q_- that has less intersection over union (IoU) than a certain threshold (*e.g.,* 0.5). Additionally, we remove Q_- that have high semantic similarity with Q_+ using SentenceBERT (Reimers & Gurevych, 2019) to reduce semantic overlap further.

Let $P(f_i \mid Q_+)$ and $P(f_i \mid Q_-)$ to be the likelihood of i-th frame to match the positive and negative query from the same video clip, respectively. We assume these likelihoods are independent since their

corresponding moments are at different temporal locations. Our model predicts the target moment by the joint probability of each frame, and the probability can be represented as:

$$P(f_i \mid Q_+, \neg Q_-) := P(f_i \mid Q_+) \cdot (1 - P(f_i \mid Q_-)).$$

Considering $P(f_i|Q_-)$, our model can focus on relatively more important meanings included in the target query. As a result, being aware of contexts preceding and following the target moment is more informative for the model's prediction, further improving *moment sensitivity*.

3.3 ARCHITECTURE

We give an overall architecture of the BM-DETR in Figure 2. First, our encoder takes the V, Q_+, and Q_- as the inputs. Then we obtain frame attention scores from encoder outputs and update them for background-aware moment detection. After our decoder predicts moments from given inputs, we calculate the losses from predictions and ground-truth moments, as in DETR (Zhu et al., 2020). In addition, we leverage a temporal shifting method to encourage the model's time-equivariant predictions. The details of the model components are described in the following sections.

3.3.1 ENCODER

Our encoder aims to catch the multimodal interaction between video V and query Q. Initially, the pre-trained model (*e.g.,*, CLIP (Radford et al., 2021)) is employed to convert each input into multi-dimensional features and normalize them. We utilize two projection layers to convert input features into the same hidden dimension d. Each projection layer consists of several MLPs. Then, we obtain video representations as $\mathbf{V} \in \mathbb{R}^{L_v \times d}$ and query representation as $\mathbf{Q} \in \mathbb{R}^{L_w \times d}$. Note that there are two query representations $\mathbf{Q}_+$ and $\mathbf{Q}_-$ for positive and negative queries, respectively. We direct them to the multimodal encoder $E(\cdot)$, a stack of transformer encoder layers denoted as:

$$E(\mathbf{V}, \mathbf{Q}) = E(PE(\mathbf{V}) \parallel \mathbf{Q}),$$

where *PE* means the positional encoding function (Vaswani et al., 2017), $\parallel$ indicates the concatenation on the feature dimension. We denote the length of concatenated features as $L = L_v + L_w$. Finally, we obtain multimodal features X_+ and X_- represented as:

$$X_+ = E(\mathbf{V}, \mathbf{Q}_+), \;\; X_- = E(\mathbf{V}, \mathbf{Q}_-),$$

where $X_+ \in \mathbb{R}^{L \times d}$ and $X_- \in \mathbb{R}^{L \times d}$.

3.3.2 IMPLEMENTING THE BACKGROUND-AWARE MOMENT DETECTION

Let us redefine the frame parts of the multimodal features X_+ and X_- are $\mathbf{v}_+ = \{\mathbf{f}_i^+\}_i^{L_v}$ and $\mathbf{v}_- = \{\mathbf{f}_i^-\}_i^{L_v}$, respectively. We compute the likelihood of each frame to match the positive and negative queries, denoted as $P(\mathbf{f}_i \mid \mathbf{Q}_+)$ and $P(\mathbf{f}_i \mid \mathbf{Q}_-)$, respectively. These probabilities can be obtained through a **P**robabilistic **F**rame-Query **M**atcher (PFM) defined as:

$$P(\mathbf{f}_i \mid \mathbf{Q}_+) = \text{PFM}(\mathbf{f}_i^+), \;\; P(\mathbf{f}_i \mid \mathbf{Q}_-) = \text{PFM}(\mathbf{f}_i^-).$$

PFM consists of two linear layers followed by tanh and sigmoid (σ) functions defined as:

$$\text{PFM}(\mathbf{f}_i) = \sigma(\textit{tanh}(\mathbf{f}_i\mathbf{W}_1)\mathbf{W}_2),$$

where $\mathbf{W}_1 \in \mathbb{R}^{d \times \frac{d}{2}}$ and $\mathbf{W}_2 \in \mathbb{R}^{\frac{d}{2} \times 1}$ are learnable matrices. As mentioned in Section 3.2, the joint probability of i-th frame $\mathbf{p}_i$ can be calculated as:

$$P(\mathbf{f}_i \mid \mathbf{Q}_+, \neg \mathbf{Q}_-) = P(\mathbf{f}_i \mid \mathbf{Q}_+) \cdot (1 - P(\mathbf{f}_i \mid \mathbf{Q}_-)).$$

After that, the softmax function is applied to obtain the frame attention scores $\mathbf{o}$:

$$\mathbf{o} = \text{Softmax}(\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_{L_v}).$$

Finally, we leverage $\mathbf{o}$ to enhance the positive frame features $\mathbf{v}_+$ in X_+ to $\mathbf{v'}_+$. This can be formulated as follows:

$$\mathbf{v'}_+ = \mathbf{o} \otimes \mathbf{v}_+,$$

where $\otimes$ is an element-wise product. If there is no negative query in the video, we only use positive visual features $\mathbf{v}_+$ to obtain frame attention scores $\mathbf{o}$. We denote the updated multimodal features as X'_+ and send it to the decoder to predict the target moment.

3.3.3 DECODER

Thanks to the advances in object detection (Liu et al., 2022a), we can directly use moment locations as queries rather than naively initialize them as learnable embeddings. We utilize learnable spans as $S = \{S_m\}_{m=1}^{M}$, and each learnable span is represented as $S_m = (c_m, w_m)$, where c_m and w_m refer to the center and width of the corresponding span. Given a span S_m, we utilize positional encoding and MLP layers to generate positional query P_m as:

$$P_m = \text{MLP}(\textit{PE}(S_m)) = \text{MLP}(\textit{PE}(c_m) \parallel \textit{PE}(w_m)),$$

where *PE* means fixed positional encoding to generate sinusoidal embeddings from the learnable span. Two key modules in our decoder are self-attention and cross-attention. In the self-attention module, the queries and keys additionally take P_m as:

$$Q_m = D_m + P_m, \;\; K_m = D_m + P_m, \;\; V_m = D_m,$$

where D_m is the decoder embedding, an input of the decoder layer. Each component in the cross-attention module can be represented as:

$$Q_m = (D_m \parallel \textit{PE}(S_m) \otimes \sigma(\text{MLP}(D_m))), \;\; K_m = (X'_+ \parallel \textit{PE}(X'_+)), \;\; V_m = X'_+.$$

The learnable spans are updated layer-by-layer, and we provide the reference spans $S_{m,\text{ref}}$ to utilize modulated attention, which helps to extract multimodal features with various lengths.

$$S_{m,\text{ref}} = \sigma(\text{MLP}(D_m))$$

Please refer to Liu et al. (2022a) for the implementation details.

3.4 FINE-GRAINED SEMANTIC ALIGNMENT

Since queries from different videos describe different semantic meanings specific to their respective video topics, they are more likely to have less semantic overlap with Q_+. Therefore, it is intuitively clear that aligning semantic meanings across different videos would be more effective than comparing with Q_-. Let the visual and textual representations from multimodal features X'_+ are $\mathbf{v}' \in \mathbb{R}^{L_v \times d}$ and $\mathbf{q}' \in \mathbb{R}^{L_v \times d}$, respectively. We first adopt an attentive pooling to extract the global context of each representation as:

$$\hat{\mathbf{v}} = \sum_{n=1}^{L_v} \mathbf{a}_i^v \mathbf{v}'_i, \;\; \mathbf{a}^v = \text{Softmax}(\mathbf{v}'\mathbf{W}_v),$$

$$\hat{\mathbf{q}} = \sum_{n=1}^{L_v} \mathbf{a}_i^q \mathbf{q}'_i, \;\; \mathbf{a}^q = \text{Softmax}(\mathbf{q}'\mathbf{W}_q),$$

where $\mathbf{W}_v \in \mathbb{R}^{d \times 1}$ and $\mathbf{W}_q \in \mathbb{R}^{d \times 1}$ are learnable matrices. Then we can compute semantic similarity score s as:

$$s = S(\hat{\mathbf{v}}, \hat{\mathbf{q}}) = \frac{\hat{\mathbf{v}}^T \cdot \hat{\mathbf{q}}}{\|\hat{\mathbf{v}}\|_2 \, \|\hat{\mathbf{q}}\|_2},$$

where $\|\cdot\|_2$ represents the L2-norm of a vector. As our background moment detection makes $\hat{\mathbf{v}}$ more sensitive to the semantic meaning within the target moment, we can learn fine-grained semantic matching with $\hat{\mathbf{q}}$. This is in contrast to vanilla contrastive learning, where moment features are pushed away by different queries, regardless of the semantic relationships. Finally, we compare the semantic scores obtained from different videos.

3.5 TEMPORAL SHIFTING

Recent studies (Xu et al., 2021; Zhang et al., 2021; Hao et al., 2022; Zhang et al., 2022) demonstrated that temporal augmentation techniques are effective for localization tasks. Inspired by this, we use a temporal shifting method that randomly moves the ground-truth moment to a new temporal location for VMR. This allows our model to make time-equivariant predictions, but we acknowledge that this technique may disrupt long-term temporal semantic information in videos. To address this issue, we empirically apply the temporal shifting method to videos with short durations (*i.e.,* $|V| < 60$s), during each training step.

3.6 LEARNING OBJECTIVES

Based on decoder outputs, we apply MLP layers to generate a set of M predictions denoted as $\hat{y} = \{\hat{y}_i\}_{i=1}^{M}$. Each prediction $\hat{y}_i$ contains two components: 1) the class label $\hat{c}_i$ to indicate whether the predicted moment is the ground-truth moment or not, and 2) temporal moment location $\hat{m}_i = (\hat{t}_s^i, \hat{t}_e^i)$. Following previous work (Lei et al., 2021), we find the optimal assignment i between the ground-truth y and the predictions $\hat{y}_i$, using the matching cost $\mathcal{C}_{\text{match}}$ as:

$$\mathcal{C}_{\text{match}}(y, \hat{y}_i) = -p(\hat{c}_i) + \mathcal{L}_{\text{moment}}(m, \hat{m}_i).$$

We use Hungarian algorithm to find the optimal pair and calculate the loss between the ground-truth moment and predictions.

$$i = \arg\min_{i \in N} \mathcal{C}_{\text{match}}(y, \hat{y}_i).$$

Details of loss formulation are described below.

Moment localization loss. The moment localization loss contains two losses: 1) L1 loss and 2) a generalized IoU loss (Rezatofighi et al., 2019). This loss is designed to calculate the accuracy of a prediction by comparing it to the ground-truth moment.

$$\mathcal{L}_{\text{moment}}(m, \hat{m}_i) = \lambda_{\text{L1}} \mid\mid m - \hat{m}_i \mid\mid + \lambda_{\text{iou}} \mathcal{L}_{\text{iou}}(m, \hat{m}_i),$$

where λ_{L1} and λ_{iou} are the coefficients to adjust weights.

Frame margin loss. The margin loss encourages frames within the ground-truth moment to have high scores via hinge loss. Let $\mathbf{f}_{\text{fore}}$ and $\mathbf{f}_{\text{back}}$ are frame features in $\mathbf{v}'_+$. We use a linear layer to predict the scores of these frame features. Note that $\mathbf{f}_{\text{fore}}$ is located within the ground-truth moment, and $\mathbf{f}_{\text{back}}$ is not. The loss can be formulated as follows:

$$\mathcal{L}_{\text{margin}} = \max(0, \Delta + \mathbf{f}_{\text{back}}\mathbf{W} - \mathbf{f}_{\text{fore}}\mathbf{W}),$$

where $\mathbf{W} \in \mathbb{R}^{d \times 1}$ and Δ is the margin.

Frame probability loss. We encourage frames within the target moment to have a high probability. Let $\mathcal{P}$ and $\mathcal{N}$ be the sets of positive and negative frame indices. Then we calculate the loss from the joint probability of frames $\mathbf{p} = \{\mathbf{p}_i\}_i^{L_v}$ as follows:

$$\mathcal{L}_{\text{prob}} = 1 - \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} \mathbf{p}_i + \frac{1}{|\mathcal{N}|} \sum_{j \in \mathcal{N}} \mathbf{p}_j.$$

Semantic alignment loss. We obtain multimodal features $\{\hat{\mathbf{v}}_i, \hat{\mathbf{q}}_i\}_{i=1}^{B}$ within a batch B that correspond to different videos and calculate semantic scores from both positive and irrelevant pairs. The loss can be formulated as follows:

$$\mathcal{L}_{\text{semantic}} = -\frac{1}{|B|} \sum_{i=1}^{B} \log \frac{\exp(S(\hat{\mathbf{v}}_i, \hat{\mathbf{q}}_i)/\tau)}{\sum_{j \in \mathcal{B}} \exp(S(\hat{\mathbf{v}}_i, \hat{\mathbf{q}}_j)/\tau)},$$

where τ is a temperature parameter and set as 0.07.

Overall loss. The overall loss $\mathcal{L}$ is determined by linearly combining the individual losses mentioned above, and we optimize our model as:

$$\mathcal{L} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{moment}}\mathcal{L}_{\text{moment}} + \lambda_{\text{margin}}\mathcal{L}_{\text{margin}} + \lambda_{\text{prob}}\mathcal{L}_{\text{prob}} + \lambda_{\text{semantic}}\mathcal{L}_{\text{semantic}},$$

where $\lambda_{\text{cls}}, \lambda_{\text{margin}}, \lambda_{\text{prob}}$, and $\lambda_{\text{semantic}}$ are hyper-parameters and λ_{moment} involves two hyper-parameters λ_{L1} and λ_{iou}, and $\mathcal{L}_{\text{cls}}$ is the cross-entropy function computed by $\hat{c}_i$ that classifies whether the predicted moment is the ground-truth moment.

Table 1: Performance results on Charades-STA test split. We use an asterisk (*) to indicate that we re-implemented the method with the same training scheme.

Method	C3D		VGG		SF+C	
	R1@0.5↑	R1@0.7↑	R1@0.5↑	R1@0.7↑	R1@0.5↑	R1@0.7↑
2D-TAN (Zhang et al., 2020b)	39.70	27.10	41.34	23.91	46.02	27.5
DRN (Zeng et al., 2020)	45.40	26.40	42.90	23.68	-	-
VSLNet (Zhang et al., 2020a)	47.31	30.19	-	-	42.69	24.14
CBLN (Liu et al., 2021)	47.94	28.22	43.67	24.44	-	-
FVMR (Gao & Xu, 2021)	38.16	18.22	42.36	24.14	-	-
MDETR* (Lei et al., 2021)	49.25	27.02	54.09	31.24	53.07	30.59
LVTR (Woo et al., 2022)	47.15	25.72	-	-	-	-
UMT (Liu et al., 2022b)	-	-	48.31	29.25	-	-
QD-DETR (Moon et al., 2023)	-	-	<u>52.77</u>	<u>31.13</u>	57.31	32.55
UniVTG (Lin et al., 2023)	-	-	-	-	<u>58.01</u>	<u>35.65</u>
CL-based:						
IVG-DCL (Nan et al., 2021)	<u>50.24</u>	<u>32.88</u>	-	-	-	-
SSCS (Ding et al., 2021)	-	-	43.15	25.54	-	-
MMN (Wang et al., 2022)	-	-	47.31	27.28	-	-
G2L (Li et al., 2023)	-	-	47.91	28.42	-	-
BM-DETR (ours)	**54.08**	**34.47**	**56.91**	**36.24**	**59.48**	**38.33**

4 EXPERIMENTS

4.1 EXPERIMENT SETUP

We experiment on four representative VMR datasets with various characteristics: Charades-STA (Gao et al., 2017), ActivityNet-Captions (Krishna et al., 2017), QVHighlights (Lei et al., 2021), and TACoS (Regneri et al., 2013). Charades-STA and ActivityNet-Captions cover human activities, whereas TACoS contains cooking scenarios. QVHighlights contains videos with diverse themes, including lifestyle vlogs and news.

Importantly, Charades-STA and ActivityNet-Captions has been widely used for VMR datasets, but there are significant bias problems (Otani et al., 2020; Yuan et al., 2021) that current models heavily rely on identifying frequent patterns in the temporal moment distribution of the datasets, rather than real comprehension. To conduct further reliable evaluations and validate generality, we also experiment on out-of-distribution test splits (*i.e.,* test-ood) in Charades-CD and ActivityNet-CD.

For a fair comparison, we use the same encoder in each dataset as used in previous models: VGG (Simonyan & Zisserman, 2014), C3D (Tran et al., 2015), I3D (Carreira & Zisserman, 2017), and SF+C, which is a concatenation of SlowFast (Feichtenhofer et al., 2019) and CLIP (Radford et al., 2021).

We use two metrics to compare our model with previous works; 1) R@n, IoU=m, which measures the percentage of the top-n predicted moments with an IoU greater than m (*i.e.,* 0.5) and 2) Mean Average Precision (mAP) over IoU thresholds.

Please refer to the Appendix for detailed experiment settings and model configurations.

4.2 COMPARISON WITH THE STATE-OF-THE-ART METHODS

In this section, we report the performance of BM-DETR on various datasets and compare it with previous models. We use bolds to mark the best score in each table, while the second-best score is underlined, and gray out methods that use additional input sources (*i.e.,* audio). We report the average performance of 5 runs with random seeds.

In Table 1-4, we compare the BM-DETR to previous models on Charades-STA, ActivityNet-Captions, TACoS, and QVHighlights, respectively. BM-DETR clearly outperforms SOTA methods, including contrastive learning-based, on Charades-STA regardless of the video encoder. For the other datasets,

Table 2: Performance results on QVHighlights test split.

Method	SF+C				
	R1@0.5↑	R1@0.7↑	mAP@0.5↑	mAP@0.75↑	mAP_{avg} ↑
MCN (Anne 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	44.63	31.73	32.14
XML+ (Lei et al., 2021)	46.69	33.46	47.89	34.67	34.90
MDETR (Lei et al., 2021)	52.89	33.02	54.82	29.40	30.73
UMT (Liu et al., 2022b)	56.23	41.18	53.83	37.01	36.12
QD-DETR (Moon et al., 2023)	**62.40**	**44.98**	62.52	39.88	39.86
UniVTG (Lin et al., 2023)	58.86	40.86	57.60	35.59	35.47
BM-DETR (ours)	60.12	43.05	**63.08**	**40.18**	**40.08**

Table 3: Performance results on ActivityNet-Captions val2 split.

Method	C3D	
	IoU@0.5↑	IoU@0.7↑
2D-TAN (Zhang et al., 2020b)	44.51	26.54
VSLNet (Zhang et al., 2020a)	43.22	26.16
DRN (Zeng et al., 2020)	45.45	24.39
CBLN (Liu et al., 2021)	48.12	27.60
SMIN (Wang et al., 2021)	48.46	30.34
GTR (Cao et al., 2021)	50.57	29.11
CL-based:		
IVG-DCL (Nan et al., 2021)	43.84	27.10
SSCS (Ding et al., 2021)	46.67	27.56
MMN (Wang et al., 2022)	48.59	29.26
G2L (Li et al., 2023)	**51.68**	**33.35**
BM-DETR (ours)	50.23	30.88

Table 4: Performance results on TACoS test split.

Method	C3D	
	IoU@0.5↑	IoU@0.7↑
2D-TAN (Zhang et al., 2020b)	37.29	25.32
VSLNet (Zhang et al., 2020a)	29.61	24.27
DRN (Zeng et al., 2020)	-	23.17
CBLN (Liu et al., 2021)	38.98	27.65
FVMR (Gao & Xu, 2021)	41.48	29.12
GTR (Cao et al., 2021)	40.39	30.22
CL-based:		
IVG-DCL (Nan et al., 2021)	38.84	29.07
SSCS (Ding et al., 2021)	41.33	29.56
MMN (Wang et al., 2022)	39.24	26.17
G2L (Li et al., 2023)	42.74	30.95
BM-DETR (ours)	**43.91**	**31.08**

Table 5: Performance results on Charades-CD test-ood split.

Method	I3D	
	IoU@0.5↑	IoU@0.7↑
2D-TAN (Zhang et al., 2020b)	35.88	13.91
LG (Mun et al., 2020)	42.90	19.29
DRN (Zeng et al., 2020)	31.11	15.17
VSLNet (Zhang et al., 2020a)	34.10	17.87
DCM (Yang et al., 2021)	45.47	22.70
Shuffling (Hao et al., 2022)	46.67	27.08
BM-DETR (ours)	**55.02**	**29.52**

Table 6: Performance results on ActivityNet-CD test-ood split.

Method	I3D	
	IoU@0.5↑	IoU@0.7↑
2D-TAN (Zhang et al., 2020b)	22.01	10.34
LG (Mun et al., 2020)	23.85	10.96
VSLNet (Zhang et al., 2020a)	20.03	10.29
DCM (Yang et al., 2021)	22.32	11.22
Shuffling (Hao et al., 2022)	24.57	13.21
BM-DETR (ours)	**26.67**	**15.33**

BM-DETR achieves competitive performance compared to baseline models. Notably, BM-DETR also shows superior performance on OOD datasets in Table 5 and 6, showcasing its robustness.

Compared with MMN, which utilizes negative samples from both the same and other videos based on IoU, we only use negative queries within the same video that have less semantic overlap with positive queries. Our model outperforms MMN on all datasets, despite using fewer negative samples, which confirms our design choices and sampling strategies. However, we observe that performance gain on ActivityNet-Captions is smaller than other datasets. The reason is that ActivityNet-Captions has a significant sparse annotation dilemma and semantic overlapping than other datasets. If there is too much semantic overlap in the video, only a small number of negative queries will be used, making our modeling less effective. This is also a significant problem that leads vanilla contrastive learning to wrong results. Nevertheless, BM-DETR performs competitively compared to SOTA methods while maintaining efficiency. We will discuss this in the following sections.

4.3 Ablation Study

Ablations on model components. To validate the effectiveness of each model component, we build up several baseline models with different model components. In Table 7, using background moment detection, fine-grained semantic matching, temporal shifting, and learnable spans are represented by BMD, FS, TS, and LS, respectively. We observe that there is significant performance degradation when BMD is unavailable, demonstrating BMD plays a crucial role in our modeling. Also, we confirm that all components contribute to performance improvement.

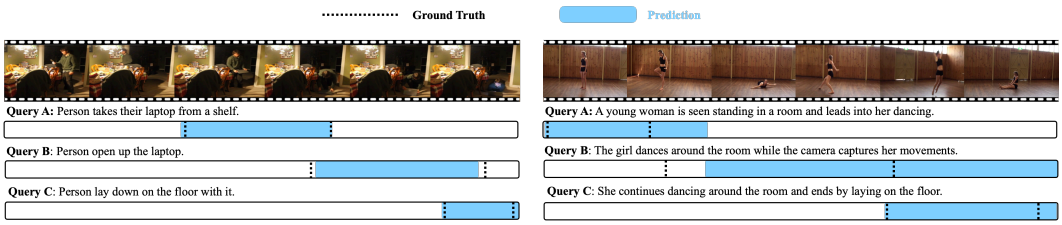

Figure 3: Visualization of predictions on Charades-STA (Left) and ActivityNet-Captions (Right).

Table 7: Ablation on model components.

Method	SF+C	
	IoU@0.5↑	IoU@0.7↑
Full Model	**59.48**	**38.33**
w/o TS	57.21	35.88
w/o LS	56.99	35.97
w/o FS	56.24	35.16
w/o BMD	55.32	35.02
w/o BMD+FS	54.89	34.57
w/o ALL	53.31	30.17

Table 8: Ablation on losses.

Method	SF+C	
	IoU@0.5↑	IoU@0.7↑
Full Losses	**59.48**	**38.33**
w/o $\mathcal{L}_{\mathrm{s}}$	58.24	37.14
w/o $\mathcal{L}_{\mathrm{m}}$	56.21	36.29
w/o $\mathcal{L}_{\mathrm{p}}$	55.79	36.03
w/o $\mathcal{L}_{\mathrm{m}}$, $\mathcal{L}_{\mathrm{p}}$, $\mathcal{L}_{\mathrm{s}}$	54.02	32.02

Table 9: Ablation on efficiency.

Method	Iteration↓	Inference↓
MMN (Wang et al., 2022)	0.32s	37s
G2L (Li et al., 2023)	0.84s	43s
BM-DETR (ours)	**0.19s**	**35s**

Albations on model losses. To investigate the impact of each loss, we turn off one loss at a time. In Table 8, $\mathcal{L}_{\mathrm{s}}$, $\mathcal{L}_{\mathrm{m}}$, and $\mathcal{L}_{\mathrm{p}}$ means margin, probability, and semantic losses, respectively. We observe that loss related to alignment within the video (*i.e.,* $\mathcal{L}_{\mathrm{m}}$ and $\mathcal{L}_{\mathrm{p}}$) helps in more accurate predictions.

Ablations on efficiency. In Table 9, we compute the average iteration time in the training stage and total inference time on ActivityNet-Captions. Our model is quite efficient compared to contrastive learning-based approaches. Moreover, G2L (Li et al., 2023) requires 8 A100 GPUs for training, but our model can be optimized with a single GPU.

4.4 Visualization Results

In Figure 3, we visualize moment predictions on Charades-STA and ActivityNet-Captions to show how our model behaves given weakly-aligned videos. For Charades-STA, the object "laptop" overlaps in ground-truth moments. Nevertheless, our model successfully provides accurate temporal moments. On the other hand, some of the predictions on ActivityNet-Captions are somewhat less accurate. Since the queries in the video have different temporal positions but overlap semantically, our model does not use negating queries on each other. This reduces the impact of our background moment detection, resulting in inaccurate moment detection. We will address our limitations in future research.

5 Conclusion

In this paper, we argue the weak visual-textual alignment problem and present BM-DETR to alleviate it. We carefully desgin our model by adopting a contrastive approach to overcome weak alignment in videos for VMR. With the proposed background-aware moment detection, BM-DETR can effectively learn to identify and focus on the most relevant visual features of the target moment. BM-DETR demonstrates its strength and robustness on four public datasets and two out-of-distribution datasets. Moreover, our model shows remarkable efficiency compared to previous contrastive learning-based methods without mining costly negative moments. Future work will focus on improving modeling for cases when low diversity negative queries cases and extending our method to a variety of video understanding tasks.

References

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

Meng Cao, Long Chen, Mike Zheng Shou, Can Zhang, and Yuexian Zou. On pursuit of designing multi-modal transformer for video grounding. *arXiv preprint arXiv:2109.06085*, 2021.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.

Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. Temporally grounding natural sentence in video. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 162–171, 2018.

Long Chen, Chujie Lu, Siliang Tang, Jun Xiao, Dong Zhang, Chilie Tan, and Xiaolin Li. Rethinking the bottom-up framework for query-based video localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 10551–10558, 2020.

Xinpeng Ding, Nannan Wang, Shiwei Zhang, De Cheng, Xiaomeng Li, Ziyuan Huang, Mingqian Tang, and Xinbo Gao. Support-set based cross-supervision for video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11573–11582, 2021.

Victor Escorcia, Mattia Soldan, Josef Sivic, Bernard Ghanem, and Bryan Russell. Temporal localization of moments in video collections with natural language. *arXiv preprint arXiv:1907.12763*, 2019.

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

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

Junyu Gao and Changsheng Xu. Fast video moment retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1523–1532, 2021.

Tengda Han, Weidi Xie, and Andrew Zisserman. Temporal alignment networks for long-term video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2906–2916, 2022.

Jiachang Hao, Haifeng Sun, Pengfei Ren, Jingyu Wang, Qi Qi, and Jianxin Liao. Can shuffling video benefit temporal bias problem: A novel training framework for temporal grounding. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pp. 130–147. Springer, 2022.

Dongliang He, Xiang Zhao, Jizhou Huang, Fu Li, Xiao Liu, and Shilei Wen. Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 8393–8400, 2019.

Jiabo Huang, Hailin Jin, Shaogang Gong, and Yang Liu. Video activity localisation with uncertainties in temporal boundary. In *European Conference on Computer Vision*, pp. 724–740. Springer, 2022.

Dohwan Ko, Joonmyung Choi, Juyeon Ko, Shinyeong Noh, Kyoung-Woon On, Eun-Sol Kim, and Hyunwoo J Kim. Video-text representation learning via differentiable weak temporal alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5016–5025, 2022.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pp. 706–715, 2017.

Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *ECCV*, 2020.

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

Hongxiang Li, Meng Cao, Xuxin Cheng, Yaowei Li, Zhihong Zhu, and Yuexian Zou. G2l: Semantically aligned and uniform video grounding via geodesic and game theory. *arXiv preprint arXiv:2307.14277*, 2023.

Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. Univtg: Towards unified video-language temporal grounding. *arXiv preprint arXiv:2307.16715*, 2023.

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

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

Ye Liu, Siyuan Li, Yang Wu, Chang-Wen Chen, Ying Shan, and Xiaohu Qie. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3042–3051, 2022b.

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9879–9889, 2020.

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

Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10810–10819, 2020.

Guoshun Nan, Rui Qiao, Yao Xiao, Jun Liu, Sicong Leng, Hao Zhang, and Wei Lu. Interventional video grounding with dual contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2765–2775, 2021.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkilä. Uncovering hidden challenges in query-based video moment retrieval. *arXiv preprint arXiv:2009.00325*, 2020.

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

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

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

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

Cristian Rodriguez, Edison Marrese-Taylor, Fatemeh Sadat Saleh, Hongdong Li, and Stephen Gould. Proposal-free temporal moment localization of a natural-language query in video using guided attention. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2464–2473, 2020.

Zheng Shou, Dongang Wang, and Shih-Fu Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1049–1058, 2016.

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

Surya T Tokdar and Robert E Kass. Importance sampling: a review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):54–60, 2010.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 4489–4497, 2015.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Hao Wang, Zheng-Jun Zha, Liang Li, Dong Liu, and Jiebo Luo. Structured multi-level interaction network for video moment localization via language query. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7026–7035, 2021.

Zhenzhi Wang, Limin Wang, Tao Wu, Tianhao Li, and Gangshan Wu. Negative sample matters: A renaissance of metric learning for temporal grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2613–2623, 2022.

Sangmin Woo, Jinyoung Park, Inyong Koo, Sumin Lee, Minki Jeong, and Changick Kim. Explore and match: End-to-end video grounding with transformer. *arXiv preprint arXiv:2201.10168*, 2022.

Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. Multilevel language and vision integration for text-to-clip retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 9062–9069, 2019.

Mengmeng Xu, Juan-Manuel Pérez-Rúa, Victor Escorcia, Brais Martinez, Xiatian Zhu, Li Zhang, Bernard Ghanem, and Tao Xiang. Boundary-sensitive pre-training for temporal localization in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7220–7230, 2021.

Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. Deconfounded video moment retrieval with causal intervention. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1–10, 2021.

Serena Yeung, Olga Russakovsky, Greg Mori, and Li Fei-Fei. End-to-end learning of action detection from frame glimpses in videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2678–2687, 2016.

Yitian Yuan, Tao Mei, and Wenwu Zhu. To find where you talk: Temporal sentence localization in video with attention based location regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 9159–9166, 2019.

Yitian Yuan, Xiaohan Lan, Xin Wang, Long Chen, Zhi Wang, and Wenwu Zhu. A closer look at temporal sentence grounding in videos: Dataset and metric. In *Proceedings of the 2nd International Workshop on Human-centric Multimedia Analysis*, pp. 13–21, 2021.

Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10287–10296, 2020.

Can Zhang, Tianyu Yang, Junwu Weng, Meng Cao, Jue Wang, and Yuexian Zou. Unsupervised pre-training for temporal action localization tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14031–14041, 2022.

Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. *arXiv preprint arXiv:2004.13931*, 2020a.

Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Towards debiasing temporal sentence grounding in video. *arXiv preprint arXiv:2111.04321*, 2021.

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

Zhu Zhang, Zhijie Lin, Zhou Zhao, and Zhenxin Xiao. Cross-modal interaction networks for query-based moment retrieval in videos. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 655–664, 2019.

Hao Zhou, Chongyang Zhang, Yan Luo, Yanjun Chen, and Chuanping Hu. Embracing uncertainty: Decoupling and de-bias for robust temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8445–8454, 2021.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

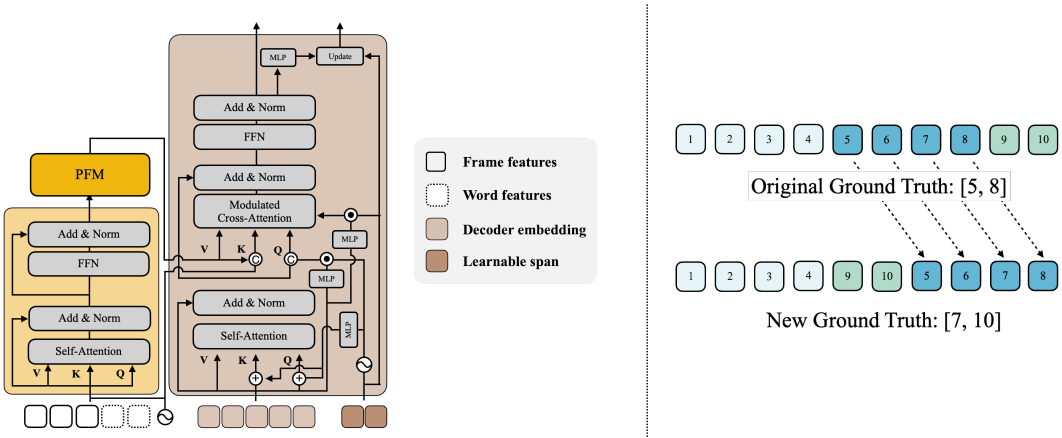

Figure 4: Visualization of our model architecture (Left) and temporal shifting method (Right).

A Appendix

A.1 Details of Model

Our model is built upon Moment-DETR (Lei et al., 2021) implemented in Pytorch and trained using a single Titan Xp. The encoder and decoder in our model are stacked with 3 layers of transformer block. We utilize AdamW to optimize our model and set the weight decay as 1e-4. We set the hidden dimension of transformers as 256, and the model weights are initialized with Xavier init. We set the batch size 32. We use a fixed number of 10 learnable spans, the same number of predicted moments.

A.2 Experiment Setup

We provide more details for training each dataset: Charades-STA (Gao et al., 2017), ActivityNet-Captions (Krishna et al., 2017), TACoS (Regneri et al., 2013), and QVHighlights (Lei et al., 2021). We extract visual features every 1s for Charades-STA and 2s for the other datasets. For Charades-STA and TACoS, we set the learning rate as 2e-4. We set the learning rate as 1e-4 for ActivityNet-Captions and QVHighlights. We train the model for 50 epochs on Charades-STA and 100 epochs on ActivityNet-Captions and TACoS. For QVHighlights (Lei et al., 2021), we train the model for 200 epochs. In addition, some queries in QVHighlights are matched to multiple ground-truth moments. In this case, our model finds the optimal bipartite matching between ground-truth moments and predictions.

A.3 Additional Visualization Results

In Figure 5, we provide four additional visualization results of our model's prediction. Also, we present frame attention scores $\mathbf{o}_+$ below. We can see the frame attention scores within the ground-truth moment are higher than others.

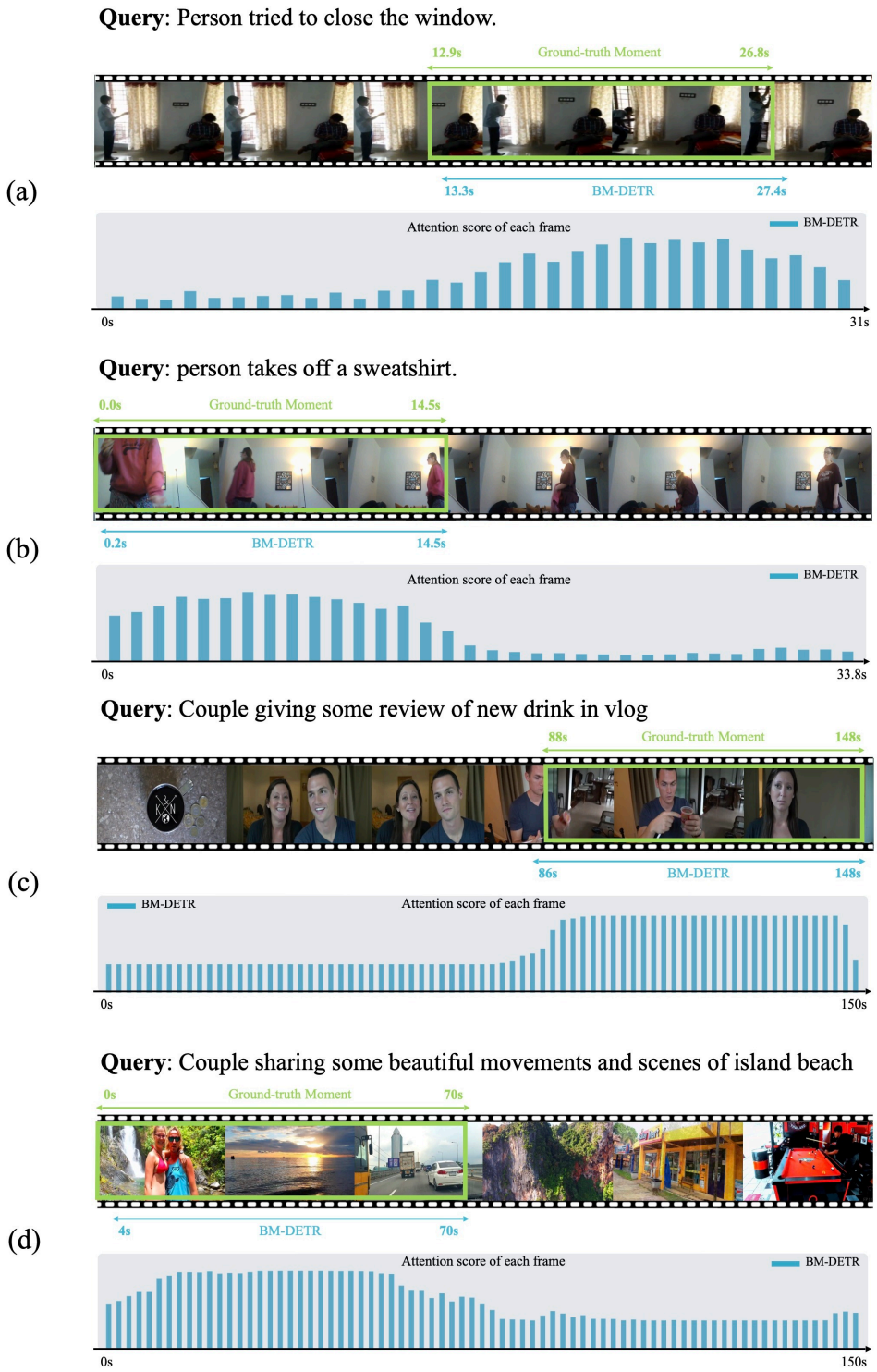

Figure 5: Four visualization examples of our model's moment prediction.