Overview

This file provides additional results not included in the main manuscript and describes the following sections.

1. Details of our experimental setup (Section A)
2. Details of our model components (Section B)
3. Additional visualization results (Section C)

A. Experimental Setup

A.1. Model configurations

We illustrate the transformer-based architecture of BM-DETR in Figure 5. The encoder and decoder in our model are stacked with 3 layers of transformer block. We set the hidden dimension of transformers as 256, and the model weights are initialized with Xavier init. We use a fixed number of 10 learnable spans, the same number of predicted moments. We utilize AdamW to optimize our model. For our sampling strategy, we set the IoU threshold as 0.7 for ActivityNet-Captions and 0.5 for the other datasets to eliminate overlapping video moments. We then extract the representations of positive and negative queries from Sentence-BERT [36] and compute the similarity between them. We set the threshold of similarity as 0.5.

A.2. Implementation Detail

We provide more details for training each dataset: Charades-STA [10], ActivityNet-Captions [18], TACoS [35], and QVHighlights [19]. We set the batch size to 32 and select loss hyper-parameters as $\lambda_{\text{L1}} = 1, \lambda_{\text{iou}} = 8$, and $\lambda_{\text{cls}} = 8$ for all datasets. We extract visual features every 1s for Charades-STA and 2s for ActivityNet, TACoS, and QVHighlights. 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 100 epochs on Charades-STA and 200 epochs on the other datasets.

B. Details of Model Components

B.1. Details of Temporal Shifting

Let us suppose the target video V has L frames as $V = \{f_i\}_{i=1}^{L}$, and the length of the ground-truth moment is l ($l < L$). First, we randomly select a new start/end index s_{start}/s_{end} as follows:

$$s_{start} \sim U(0, L-l), \quad l \in \mathcal{Z} \tag{23}$$

$$s_{end} = s_{start} + l. \tag{24}$$

Then we shift frames in the ground-truth moment to the new ground-truth moment while maintaining the sequence of frames. For ease of explanation, in Figure 6, let us define

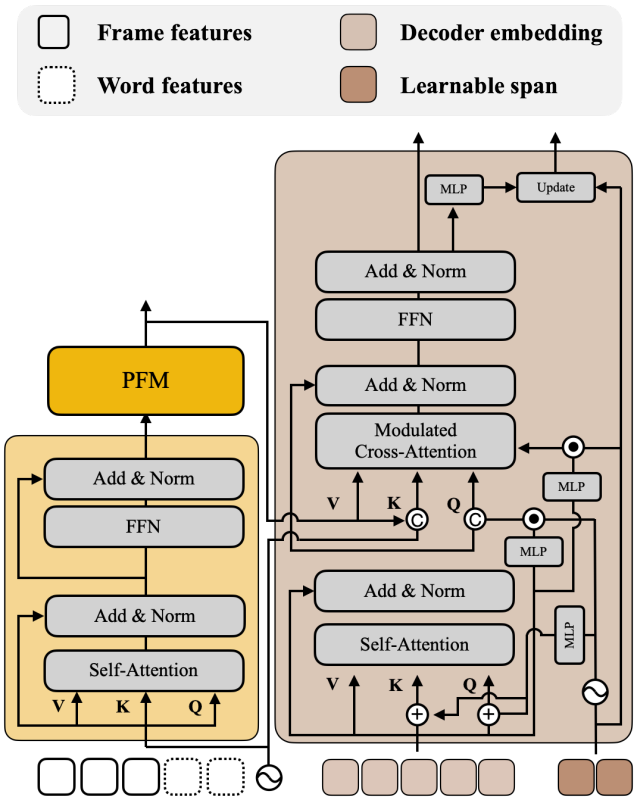

Figure 5. The detailed architecture of BM-DETR.

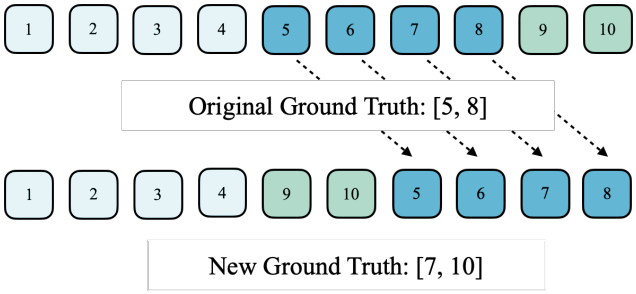

Figure 6. Visualization of temporal shifting method. When temporal shifting is applied to the video, we randomly move the frames in the ground-truth moment while keeping the sequence of frames

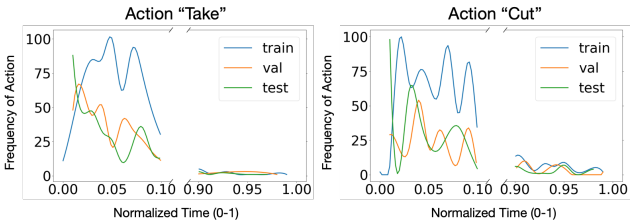

Figure 7. Normalized temporal distribution of action in TACoS.

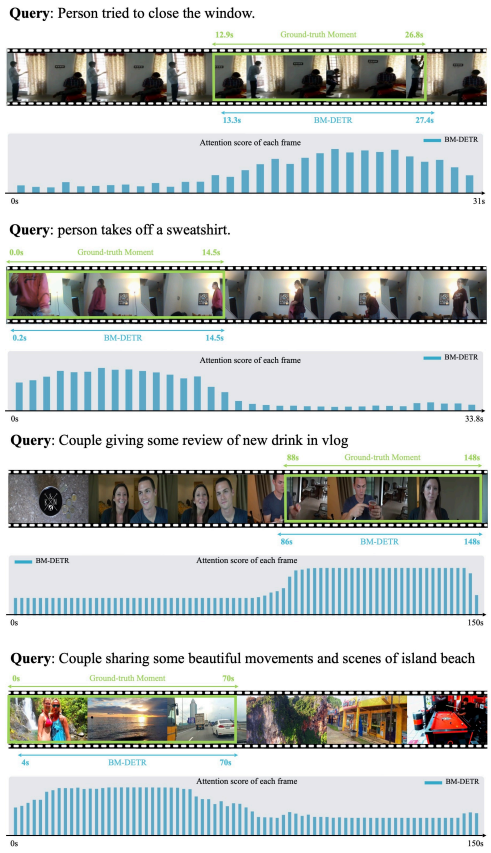

Figure 8. Four visualization examples of our model's moment prediction. We show predicted and ground-truth moments and present the attention scores $\mathbf{o}$ below.

the given video V contains 10 frames with 1 FPS, and the start and end times of the ground-truth moment in V are 5s and 8s. Then we randomly select s_{start} and s_{end} as 7s and 10s. Finally, the ground-truth moment (5s, 8s) is changed to the new ground-truth moment (7s, 10s).

B.2. Discussion of Temporal Shifting

As discussed before, we further investigate the temporal shifting's inconsistent impact. We find that TACoS videos have 60 times more queries on average than Charades-STA (135.2 vs 2.3), and due to the nature of cooking videos, learning procedural information from these queries appears to be crucial. As depicted in Figure 7, specific actions consistently appear in the early video segments but diminish in the later parts. Consequently, temporal shifting may not yield benefits for learning these temporal relationships. We will develop effective temporal augmentation methods for VMR in our future work.

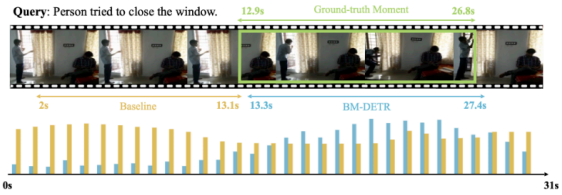

Figure 9. Visualization of model's predictions. We present the frame probability $\mathbf{p}$ below the ground-truth and predicted moments.

C. Additional Visualization Results

We provide further insight into how our model addresses the weak alignment problem in the video, providing additional support for the results in Table 6. Figure 9 presents the predictions of each model for a given video. The query describes only one person despite two individuals in the video, and the terms "person" and "window" in the query are present not only in the target moment but throughout the entire video. Our baseline model fails to predict the target moment due to the lack of clear distinction in frame probabilities between the ground-truth moment. In contrast, BM-DETR provides accurate predictions by assigning higher probabilities to frames in the ground-truth moments than to other frames, mitigating the weak alignment problem effectively. Additionally, we provide four additional visualization results of our model's prediction on QVHighlights in Figure 8.