Towards Balanced Alignment: Modal-Enhanced Semantic Modeling for Video Moment Retrieval

Zhihang Liu[1], Jun Li[2], Hongtao Xie[1]*, Pandeng Li[1], Jiannan Ge[1], Sun-Ao Liu[1], Guoqing Jin[2]

[1] University of Science and Technology of China, Hefei, China
[2] People's Daily Online
{liuzhihang, lpd, gejn, lsa1997}@mail.ustc.edu.cn, htxie@ustc.edu.cn, {lijun, jinguoqing}@people.cn

Abstract

Video Moment Retrieval (VMR) aims to retrieve temporal segments in untrimmed videos corresponding to a given language query by constructing cross-modal alignment strategies. However, these existing strategies are often sub-optimal since they ignore the modality imbalance problem, *i.e.*, the semantic richness inherent in videos far exceeds that of a given limited-length sentence. Therefore, in pursuit of better alignment, a natural idea is enhancing the video modality to filter out query-irrelevant semantics, and enhancing the text modality to capture more segment-relevant knowledge. In this paper, we introduce Modal-Enhanced Semantic Modeling (MESM), a novel framework for more balanced alignment through enhancing features at two levels. First, we enhance the video modality at the frame-word level through word reconstruction. This strategy emphasizes the portions associated with query words in frame-level features while suppressing irrelevant parts. Therefore, the enhanced video contains less redundant semantics and is more balanced with the textual modality. Second, we enhance the textual modality at the segment-sentence level by learning complementary knowledge from context sentences and ground-truth segments. With the knowledge added to the query, the textual modality thus maintains more meaningful semantics and is more balanced with the video modality. By implementing two levels of MESM, the semantic information from both modalities is more balanced to align, thereby bridging the modality gap. Experiments on three widely used benchmarks, including the out-of-distribution settings, show that the proposed framework achieves a new start-of-the-art performance with notable generalization ability (*e.g.*, 4.42% and 7.69% average gains of R1@0.7 on Charades-STA and Charades-CG). The code will be available at https://github.com/lntzm/MESM.

Introduction

Video moment retrieval (VMR) poses a meaningful and challenging task in video understanding. Given a natural language query that describes a moment segment in an untrimmed video, VMR aims to determine the start and end timestamps of the segment in the video (Anne Hendricks et al. 2017; Gao et al. 2017). Therefore, it necessitates an accurate understanding of both the video content and the language query, as well as their alignment (Li et al. 2023b).

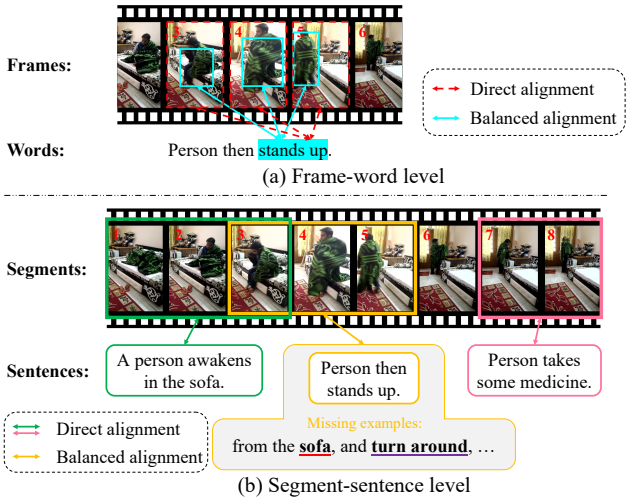

Figure 1: We creatively analyze the modality imbalance problem in VMR and the comparison between existing direct alignment and our balanced alignment, which manifests in two levels: (a) Frame-word level, the description of a word should typically align with specific parts within a frame (balanced) rather than the entire frame (direct). (b) Segment-sentence level, there is some semantic information in the segment but absent in the given sentence. The segment should typically align with the expanded sentence semantics (balanced) rather than the sentence only (direct).

Modality alignment in existing VMR methods is primarily implemented at two distinct levels. Some previous studies (Li, Guo, and Wang 2021; Liu et al. 2022a) align frame-level and word-level features, devising efficient alignment strategies to accurately regress the moments. Another line of methods (Chen and Jiang 2019; Wang et al. 2022) generates proposals to extract segment-level features, aligning them with sentence-level features to identify the most matching segment as the answer. There are also some methods considering both the frame-word and segment-sentence level (Wang et al. 2021; Moon et al. 2023). Normally, they first align frame-level and word-level features, then pool the segments for further alignment and moment retrieval.

Despite the achievements of existing alignment strategies,

*Corresponding author

most of them disregard a crucial modality imbalance problem at both the frame-word and segment-sentence levels, resulting in a modality gap. At the frame-word level, as shown in Figure 1(a), the words in a sentence are typically aligned with specific parts within a frame rather than the entire frame (*e.g.*, the action *stands up*), which poses difficulties to understand the fine-grained relationship between two modalities. Figure 1(b) shows the case of the segment-sentence level. First, the semantic information of the segment (*e.g.*, frame #3 to #5) surpasses the details provided in the given sentence and humans can easily infer missing information (*e.g.*, *from the sofa*). Second, the sentence itself may be ambiguous for VMR due to the annotation subjectivity. For example, frame #5 captures the action of *turn around*, which is entirely absent in the given sentence. Both scenarios result in a negative impact on video understanding. In summary, due to the inherent semantic richness of the video modality, the textual modality should only align to a subset of video modality at both levels, and direct alignment with the entire modalities thus results in sub-optimal solutions.

To tackle the problem, a natural idea is to enhance both modalities simultaneously. The video modality should be enhanced to filter out irrelevant semantics for the query, and the textual modality should be enhanced to capture more knowledge related to the segment. Therefore, we propose a novel framework named Modal-Enhanced Semantic Modeling (MESM) to enhance them at two levels. At the frame-word level, we enhance the video modality by reconstructing words via a weight-shared cross-attention mechanism. Since the words typically refer to certain portions of the frames, the reconstruction renders the model more sensitive to these semantically relevant portions and suppresses irrelevant ones. Consequently, there is less redundant semantic information in the output enhanced video feature, thus more balanced with words. At the segment-sentence level, we enhance the textual modality by learning complementary knowledge for the given query. As shown in Figure 1(b), the absence of semantic knowledge typically originates from both the given sentences within a video (*e.g.*, *sofa*, underlined in red) and the scenes of video segments (*e.g.*, *turn around*, underlined in purple). Therefore, we can acquire the absent semantics by learning from both sources. We mask the given sentence and regenerate the semantic knowledge supervised by the corresponding segment. As the generated knowledge complements the query, the semantic information of the query becomes stronger and is thus more balanced with the segment. Extensive experiments show our MESM achieves new state-of-the-art performance on three benchmarks and the out-of-distribution settings, demonstrating improved modality alignment and generalization.

The main contributions of our paper can be listed as follows. (1) As far as we know, we are the first to analyze the modality imbalance problem in VMR from both the frame-word and segment-sentence levels. (2) To alleviate the modality imbalance problem, we propose a novel framework MESM to model the enhanced semantic information from two levels, balancing the alignment to bridge the modality gap. (3) Extensive experimental results demonstrate the effectiveness of the proposed method.

Related Work

Video Moment Retrieval. Different from Video Retrieval (Li et al. 2022b), Video Moment Retrieval is a cross-modal task that emphasizes the ability to understand both video and textual modalities, including their alignment. The alignment can be typically split into the frame-word and segment-sentence levels. Some methods align the frame-level feature with the word-level feature (Yuan, Mei, and Zhu 2019; Zhang et al. 2020a; Liu et al. 2021a; Li, Guo, and Wang 2021; Liu et al. 2022a). Normally, they design various alignment strategies to directly predict the start and end moments. Other methods focus on aligning the segment-level feature with the sentence-level feature (Gao et al. 2017; Chen and Jiang 2019; Zhang et al. 2020b; Wang et al. 2022). They usually generate proposals to obtain the segment-level feature and align it with the sentence-level feature to select the best matching segment. There are recently some methods implementing the alignment from both levels (Wang et al. 2021; Sun et al. 2022; Moon et al. 2023; Wang et al. 2023a). SMIN (Wang et al. 2021) carefully designs multi-level alignment based on 2D-TAN. DETR-based methods (Lei, Berg, and Bansal 2021; Moon et al. 2023; Wang et al. 2023a; Li et al. 2023a) usually do the frame-word level alignment, and then pool the segments with learnable proposals for further interaction, which yields promising results. However, most of these methods overlook the modality imbalance problem, leading to the modality gap.

Modality Imbalance Problem. The modality imbalance problem seems widely existing in video-text representation tasks. (Ko et al. 2022) points out the non-sequential alignment problem between the video and the text due to the ambiguity of labeling and designed a differentiable weak temporal alignment. (Wu et al. 2023) used large language models to generate auxiliary captions for a video to complete the video-text retrieval task. In VMR, the modality imbalance problem is also crucial but few researchers focus on it. (Ding et al. 2021) builds a support set using generative captions, considering the co-existence of some visual entities. Still, many methods only use one video-query pair as their input, ignoring the causal relationship among different sentences of segments within the same video and simply considering these sentences as negative ones (Wang et al. 2022; Luo et al. 2023). Different from them, we utilize this information with the video modality together, enhancing both the video and textual modalities, leading to a more balanced alignment and bridging the modality gap.

Proposed Method

Overview

Problem Formulation. Given a pair of an untrimmed video $V = \{f_i\}_{i=1}^{N_v}$ and a language query $Q^\dagger = \{w_i^\dagger\}_{i=1}^{N_w}$, VMR aims to predict a video segment of moment $\hat{m} = (\hat{t_s}, \hat{t_e})$ that is most relevant to $Q^\dagger$, where N_v and N_w represent the number of frames and words, respectively, $\hat{t_s}$ and $\hat{t_e}$ indicate the predicted start and end time of the video segment.

Pipeline. Figure 2 shows the pipeline of the proposed MESM, which consists of three steps. First, an offline video and text feature extractor is utilized to obtain frame-level and

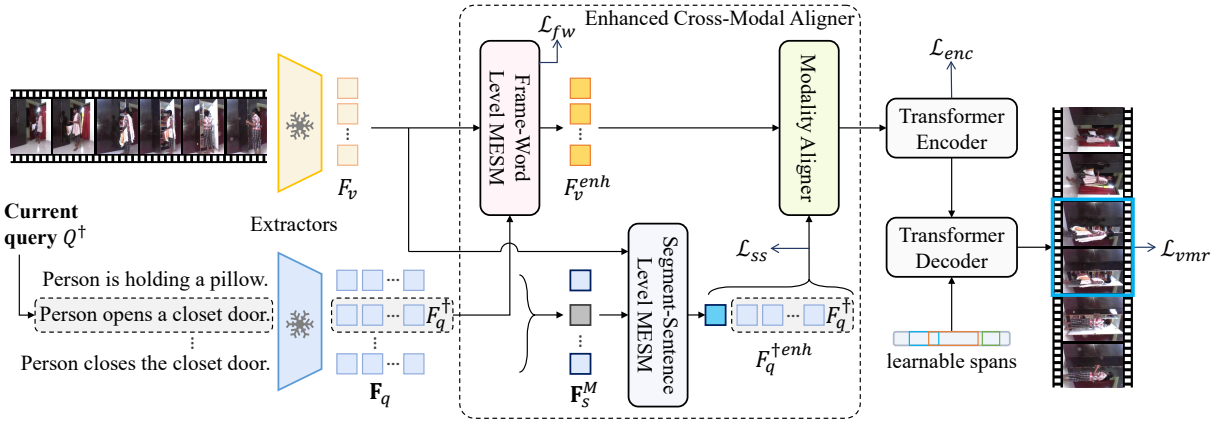

Figure 2: An overview of our MESM, which includes the feature extractors, the proposed Enhanced Cross-Modal Aligner (ECMA), and a transformer encoder-decoder network. We model the enhanced semantic information at two levels in ECMA, which consists of the Frame-Word level MESM (FW-MESM) and the Segment-Sentence level MESM (SS-MESM).

word-level features. Then, we design an Enhanced Cross-Modal Aligner (ECMA) to alleviate the modality imbalance problem and complete a more balanced alignment. Last, a transformer encoder-decoder network is utilized to encode the aligned feature and decode the moments from learnable spans. Different from many methods that directly align the features of different modalities, we focus on balancing the alignment through modal-enhanced semantic modeling from both frame-word and segment-sentence levels in the proposed ECMA, bridging the modality gap.

Feature Extractors

Feature extractors are necessary for downstream tasks (Du et al. 2022; Zheng et al. 2023; Zhang et al. 2023). Followed by most VMR methods (Zhang et al. 2020b; Wang et al. 2023a), we use offline feature extractors to get pre-obtained features from the raw data of the video and text. Generally, given a video extractor and a text extractor, we use trainable MLPs to map the extracted video feature and text feature to a common space. Given a set of language queries $\mathbf{Q} = \{Q^i | i = 1, ..., K\}$ belonging to the same video V, the mapped video and text feature can be represented as $F_v \in \mathbb{R}^{L_v \times D}$ and $\mathbf{F}_q = \{F_q^i \in \mathbb{R}^{L_w \times D} | i = 1, ..., K\}$, respectively, where K is the number of sentences in the video, D is the dimension of the common space. L_v and L_w are the lengths of the features. We use $F_q^\dagger \in \mathbf{F}_q$ to represent the feature of current query $Q^\dagger$ for moment retrieval, and thus F_v and $F_q^\dagger$ are frame-level and word-level features, respectively.

Enhanced Cross-Modal Aligner

This section presents our proposed Enhanced Cross-Modal Aligner, comprising three sub-modules: Frame-Word level MESM (FW-MESM), Segment-Sentence level MESM (SS-MESM), and the Modality Aligner (MA). FW-MESM enhances the video modality at the frame-word level by emphasizing the query-relevant portions of frame-level features and suppressing irrelevant ones. SS-MESM enhances the text modality at the segment-sentence level by generating a complementary token derived from both the query set and the ground-truth segment. Given that FW-MESM and SS-MESM generate enhanced features to address the modality imbalance issue, we subsequently implement MA to achieve the ultimate cross-modal alignment.

Frame-Word Level MESM. Since words often refer to specific parts within frames (Ge et al. 2021, 2022), we enhance the frame-level feature to filter out redundant parts and design an efficient semantic modeling strategy based on a weight-shared cross-attention mechanism. It is proven that weight-shared self-attention can process data from different modalities (Bao et al. 2022; Wang et al. 2023b), and we expand it to the case of cross-attention. As shown in Figure 3, the output of cross-attention F_v^{enh} (the left branch) symbolizes the video feature targeted for enhancement. The enhancement necessitates the acquisition of fine-grained discrimination ability to emphasize word-relevant parts within the frames, and we implement it by an auxiliary masked language modeling (MLM) task with the weight-shared cross attention (the right branch). Once the ability is obtained, the shared weights provide a bridge to enhance the output.

Specifically, we first treat the projection of frame-level feature $\mathcal{Q} = W^v F_v$ as *query*, the projection of word-level feature $\mathcal{K} = W^k F_q^\dagger$ and $\mathcal{V} = W^v F_q^\dagger$ as *key* and *value*, where W^q, W^k, W^v are linear projection matrices. Therefore, the output feature $F_v^{enh} \in \mathbb{R}^{L_v \times D}$ can be formulated as:

$$F_v^{enh} = F_v + \text{MLP}\left(\text{softmax}\left(\frac{\mathcal{Q}\mathcal{K}^\top}{\sqrt{d}}\right)\mathcal{V}\right), \quad (1)$$

where d is the dimension of the *query*, *key* and *value*. To enhance F_v^{enh}, we exchange the modality of the input and employ MLM. During the MLM, 1/3 of the words are randomly masked. If we denote the masked word-level feature as $F_q^{\dagger m}$, and the modal-exchanged inputs are $\mathcal{Q}^* = W^q F_q^{\dagger m}$ for *query*, $\mathcal{K}^* = W^k F_v$ and $\mathcal{V}^* = W^v F_v$ for *key* and *value*, the reconstructed feature of words $F_q^{\dagger r} \in \mathbb{R}^{L_w \times D}$ can be calculated similar to Equation 1. Then a fully connected layer and softmax operation is utilized to get the probability distribution $P(F_q^{\dagger r}) \in \mathbb{R}^{L_w \times N_{\text{vocab}}}$ of the words, where

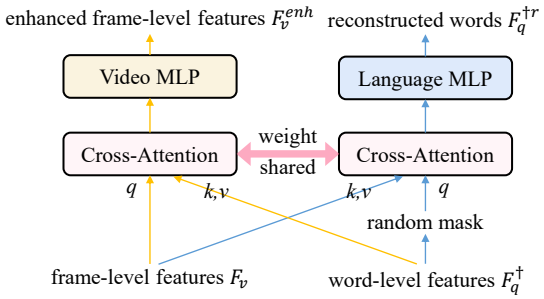

Figure 3: The pipeline of FW-MESM. The weights of the cross-attention are shared, inputs are exchanged for MLM.

N_{vocab} is the vocabulary size. We use the cross-entropy loss to measure the similarity between the reconstructed words and the original words, which can be formulated as:

$$\mathcal{L}_{fw} = -\frac{1}{L_w}\sum_{j=1}^{L_w} z_j^{\dagger} \log P_j(F_q^{\dagger r}), \tag{2}$$

where z_j is the label of the j-th word in a sentence.

Due to the shared weights, the obtained ability from the MLM task applies to the original output as well. Therefore, F_v^{enh} is enhanced to highlight the semantically relevant portions in F_v while filtering out irrelevant portions, making it more balanced with the textual modality.

Segment-Sentence Level MESM. Since the sentence can not fully cover the segment, we enhance the textual modality at the segment-sentence level by generating a complementary token from context sentences and the ground-truth segment, then the token is concatenated to the given query. To supervise the learning of the complementary knowledge, we construct a positive set for contrastive learning. The positive set collects the existing neighborhoods of the ground-truth segment to perform a soft supervision since they own similar semantic information, which we introduce later.

As we have extracted the word-level feature $\mathbf{F}_q$ of the K sentences, we simply average F_q^i for i-th sentence to get the sentence-level feature $F_s^i = \frac{1}{L_w}\sum_{j=1}^{L_w}(F_q^i)_j$ and thus $F_s^i \in \mathbb{R}^D$. For the current sentence-level feature $F_s^{\dagger}$, we replace it with a learnable [MASK] token $F_s^{\dagger M} \in \mathbb{R}^D$, and the set of the sentence-level feature with the masked one can be written as $\mathbf{F}_s^M = \{F_s^1, ..., F_s^{\dagger M}, ..., F_s^K\} \in \mathbb{R}^{K\times D}$. Then the cross-attention layers are implemented on $\mathbf{F}_s^M$ (*query*) and F_v (*key* and *value*). The output of the cross-attention layers can be represented as:

$$\mathbf{F}_s^{gen} = \{F_s^{1gen}, ..., F_s^{\dagger gen}, ..., F_s^{Kgen}\} \in \mathbb{R}^{K\times D}. \tag{3}$$

Note that $F_s^{\dagger gen}$ is in the output $\mathbf{F}_s^{gen}$ and we take it out as the generated token for the complementary knowledge to the sentence. Then we concatenate the generated token $F_s^{\dagger gen}$ with the word-level feature $F_q^{\dagger}$ together to get the enhanced word-level feature $F_q^{\dagger enh} = [F_s^{\dagger gen}, F_q^{\dagger}] \in \mathbb{R}^{(L_w+1)\times D}$.

With the complementary knowledge, the query is more balanced with the segment. Thus, we use the segment-level feature to supervise $F_q^{\dagger enh}$ to obtain knowledge related to the segment. Given the ground truth of the video segment (l_s, l_e), where l_s and l_e denote the start and the end index of the frame-level feature F_v. We take the average as the segment-level feature $S \in \mathbb{R}^D$, which can be formulated as:

$$S = \frac{1}{l_e + 1 - l_s}\sum_{j=l_s}^{l_e}(F_v)_j. \tag{4}$$

Then we design a contrastive loss to supervise the knowledge learning. Since there may be some neighbor segments with similar moments in the video, we build a positive set S_{pos} in a batch based on the IoU among the segments. We take the segments as positive when the IoU between two of them is larger than γ, and the corresponding knowledge should be similar. The contrastive loss can be formulated as:

$$\mathcal{L}_{ss} = -\log\frac{\sum_{j\in S_{pos}}\exp(\sum_{k=1}^{L_w+1} F_q^{\dagger enh}\cdot S/\tau)}{\sum_{j=1}^{N_b}\exp(\sum_{k=1}^{L_w+1} F_q^{\dagger enh}\cdot S/\tau)}, \tag{5}$$

where N_b denotes the batch size, τ is the temperature coefficient. Supervised by the segment-level feature S, the enhanced word-level $F_s^{\dagger enh}$ thus contains the complementary semantic information within the whole segment and is thus more balanced with the video modality.

Modality Aligner. Since we have gotten the enhanced frame-level feature F_v^{enh} and enhanced word-level feature $F_q^{\dagger enh}$, we finally employ cross-attention layers between F_v^{enh} (*query*) and $F_q^{\dagger enh}$ (*key* and *value*) to do the modality interaction and alignment. The final aligned feature $F \in \mathbb{R}^{L_v\times D}$ can be calculated as the standard cross-attention.

Transformer Encoder-Decoder

After the modal-aligned feature F is obtained from ECMA, a DETR (Carion et al. 2020) network is utilized to complete the VMR, which consists of a transformer encoder and decoder. The transformer encoder encodes F to a fusion representation F_{enc}, helping the model better understand the sequence relations. The encoding process follows the standard self-attention and the loss can be calculated as:

$$\mathcal{L}_{enc} = -\frac{1}{L_v}\sum_{j=1}^{L_v} y_j\log(s_j) + (1-y_j)\log(1-s_j), \tag{6}$$

where $s \in \mathbb{R}^{L_v}$ is the similarity vector which represents the attention of the model to focus, and is obtained by an MLP from F_{enc}. $y \in \mathbb{R}^{L_v}$ is the similarity label, where y_j=1 if the j-th frame is within the ground-truth and y_j=0 otherwise.

As for the transformer decoder, inspired by DAB-DETR (Liu et al. 2021b), we follow QD-DETR (Moon et al. 2023) to design learnable spans, representing the center coordinate and the window. The decoder calculates the standard cross-attention between learnable spans and the pooled features, refining the result of spans continually.

Inspired by (Carion et al. 2020; Lei, Berg, and Bansal 2021), the moment retrieval loss consists of three parts:

$$\mathcal{L}_{vmr} = \lambda_{L1}\|m - \hat{m}\|_1 + \lambda_{iou}\mathcal{L}_{iou}(m, \hat{m}) + \lambda_{ce}\mathcal{L}_{ce}, \tag{7}$$

where m and $\hat{m}$ are the predicted and ground-truth moments, $\lambda_{(L1,iou,ce)}$ are the hyper-parameters, $\mathcal{L}_{iou}$ is the generalized IoU loss (Union 2019), $\mathcal{L}_{ce}$ is the cross-entropy loss to classify the foreground or background (Carion et al. 2020).

Methods	Extractors	Charades-STA				
		R1		mAP		
		@0.5	@0.7	@0.5	@0.75	avg
2D-TAN*	VGG, GloVe	41.34	23.91	54.68	24.15	29.26
CBLN		47.94	28.22	-	-	-
RaNet*		42.91	25.82	53.28	24.41	28.55
DCM		47.80	28.00	-	-	-
MMN*		46.93	27.07	58.85	28.16	31.58
UMT†		48.44	29.76	58.03	27.46	30.37
QD-DETR*		51.51	32.69	62.88	32.60	34.46
MESM(Ours)		**56.69**	**35.99**	**67.94**	**33.64**	**37.33**
VDI	C, C	52.32	31.37	-	-	-
M-DETR*	C+SF, C	53.22	30.87	58.86	26.43	30.43
QD-DETR*		56.89	32.50	66.49	32.00	35.39
MESM(Ours)		**61.24**	**38.04**	**70.31**	**36.36**	**38.57**

Table 1: Performance comparison (%) on the Charades-STA dataset. "†" means the method uses the audio data. "*" denotes that we re-implement the method under the same training scheme. M-DETR is short for MomentDETR.

Methods	TACoS			
	R1@0.1	R1@0.3	R1@0.5	mIoU
VSLNet	-	29.61	24.27	24.11
2D-TAN	47.59	37.29	25.32	-
CBLN	49.16	38.98	27.65	-
RaNet	-	43.34	33.54	-
SeqPAN	-	31.72	27.19	25.86
SMIN	-	48.01	35.24	-
MMN	51.39	39.24	26.17	-
MS-DETR	-	47.66	37.36	35.09
MESM(Ours)	**65.03**	**52.69**	**39.52**	**36.94**

Table 2: Performance comparison (%) on TACoS. All the listed methods use C3D and GloVe as their extractors.

As a result, the final loss is:

$$\mathcal{L} = \lambda_{fw}\mathcal{L}_{fw} + \lambda_{ss}\mathcal{L}_{ss} + \lambda_{enc}\mathcal{L}_{enc} + \mathcal{L}_{vmr}, \quad (8)$$

where λ_{fw}, λ_{ss} and λ_{enc} are the hyper-parameters.

Experiments

Experimental Settings

Datasets. We evaluate the proposed method on three widely used datasets, which are Charades-STA (Gao et al. 2017), TACoS(Regneri et al. 2013), and QVHighlights (Lei, Berg, and Bansal 2021). We also experiment on Charades-CG (Li et al. 2022a), which proposes out-of-distribution (OOD) settings for Charades-STA. Charades-STA is built upon the Charades dataset (Sigurdsson et al. 2016), which consists of daily indoor activities. TACoS includes long-term videos about cooking activities. videos in QVHighlights range from daily vlog, travel vlog, and news. Charades-CG is proposed to evaluate the generalization ability by constructing new splits. These datasets cover videos from different domains, which are suitable for our evaluation in multiple scenes.

Metrics. We calculate R1@μ, mAP@μ, mIoU, and mAP_{avg} as used in previous methods (Lei, Berg, and Bansal 2021; Zhang et al. 2020b,a). R1@μ and mAP@μ are the recall and mean average precision with IoU thresholds μ within the top-1 results. mIoU denotes the average IoU and mAP_{avg} indicates the average mAP with μ=[0.5:0.05:0.95].

Methods	QVHighlights				
	R1		mAP		
	@0.5	@0.7	@0.5	@0.75	avg
MCN	11.41	2.72	24.94	8.22	10.67
CAL	25.49	11.54	23.40	7.65	9.89
XML	41.83	30.35	44.63	31.73	32.14
XML+	46.69	33.46	47.89	34.67	34.90
MomentDETR	52.89	33.02	54.82	29.40	30.73
UMT†	56.23	41.18	53.38	37.01	36.12
QD-DETR	62.40	44.98	62.52	39.88	39.86
MESM(Ours)	**62.78**	**45.20**	**62.64**	**41.45**	**40.68**

Table 3: Performance comparison (%) on QVHighlights *test* split. All the listed methods use C+SF and C as their extractors. "†" denotes they use audio data.

Implementation Details. We use different offline feature extractors for a fair comparison. VGG (Simonyan and Zisserman 2014), I3D (Carreira and Zisserman 2017), C3D (Tran et al. 2015) and C+SF (short for CLIP+SlowFast) (Radford et al. 2021; Feichtenhofer et al. 2019) are utilized as video extractors, GloVe (Pennington, Socher, and Manning 2014) and C (short for CLIP) are used as text extractors. We set γ as 0.9, the hidden dimension of the transformer layers as 256, the layers of FW-MESM, MA, transformer encoder, and decoder as 2. We build our model upon QD-DETR (Moon et al. 2023) with some optimizations, and train our model with Adam optimizer (Kingma and Ba 2014) on a single NVIDIA RTX 3090.

Performance Comparisons

We compare our MESM with the following state-of-the-art methods: MCN (Anne Hendricks et al. 2017), CAL (Escorcia et al. 2019), XML (Lei et al. 2020), 2D-TAN (Zhang et al. 2020b), VSLNet (Zhang et al. 2020a), LGI (Mun, Cho, and Han 2020), CBLN (Liu et al. 2021a), RaNet (Gao et al. 2021), MomentDETR (Lei, Berg, and Bansal 2021), SeqPAN (Zhang et al. 2021), SMIN (Wang et al. 2021), VISA (Li et al. 2022a), MMN (Wang et al. 2022), UMT (Liu et al. 2022b), QD-DETR (Moon et al. 2023), VDI (Luo et al. 2023), MS-DETR (Wang et al. 2023a).

Charades-STA. As Table 1 shows, our MESM performs the best with a large margin both on the uni-modal feature extractor (VGG, GloVe) and the multi-modal extractor (C+SF, C). Compared with the strong baseline QD-DETR, our MESM obtains 3.03% average gains in mAP_{avg} and 4.42% in R1@0.7 on two types of extractors. Though multi-modal pre-trained extractors perform better than separated ones, the modality imbalance problem still makes them hard to achieve comprehensive alignment. When our MESM models more balanced semantics, it reasonably outperforms the existing state-of-the-art methods. Though VDI does not use SlowFast, they employ a sequence model to capture the temporal relationship based on CLIP features with spatial information in $\mathbb{R}^{L_v \times H \times W \times D}$, incurring much more compu-

Methods	Year	Extractors	Novel-composition			Novel-word		
			R1@0.5	R1@0.7	mIoU	R1@0.5	R1@0.7	mIoU
LGI	2020	I3D, GloVe	29.42	12.73	30.09	26.48	12.47	27.62
VSLNet	2020	I3D, GloVe	24.25	11.54	31.43	25.60	10.07	30.21
VISA	2022	I3D, GloVe	45.41	22.71	**42.03**	42.35	20.88	40.18
MESM(Ours)	2024	I3D, GloVe	**46.19**	**26.00**	41.40	**50.50**	**33.67**	**46.20**
MomentDETR*	2021	C+SF, C	37.65	18.91	36.17	43.45	21.73	38.37
QD-DETR*	2023	C+SF, C	40.62	19.96	36.64	48.2	26.19	43.22
VDI	2023	C, C	-	-	-	46.47	28.63	41.60
MESM(Ours)	2024	C+SF, C	**44.39**	**23.27**	**39.89**	**52.66**	**31.22**	**46.38**

Table 4: Performance comparison (%) on Charades-CG, which contains two types of OOD settings on Charades-STA: novel-composition and novel-word. "*" denotes the result we re-implement under the same training scheme.

FW	SS	$\mathcal{L}_{enc}$	R1@0.5	R1@0.7	mIoU	mAP_{avg}
			53.82	30.78	46.75	33.99
✓			54.76	31.51	47.15	34.18
	✓		54.60	33.17	47.60	34.24
		✓	57.66	35.00	49.91	36.82
✓		✓	59.62	36.26	50.91	37.83
	✓	✓	60.19	37.39	51.15	38.31
✓	✓	✓	**61.24**	**38.04**	**52.14**	**38.57**

Table 5: Main ablation study (%) of module FW-MESM (FW) and SS-MESM (SS), loss $\mathcal{L}_{enc}$ on Charades-STA.

cross-attention layers		R1@0.5	R1@0.7	mIoU	mAP_{avg}
w/o SS	2×MA w/o FW	57.66	35.00	49.91	36.82
	4×MA w/o FW	55.03	33.05	47.63	36.79
	2×FW+2×MA	**59.62**	**36.26**	**50.91**	**37.83**
+SS	all w/o MLM	57.39	36.37	50.10	37.49
	all	**61.24**	**38.04**	**52.14**	**38.57**

Table 6: Ablation study (%) of MLM. FW and SS denote FW-MESM and SS-MESM, w/o means without.

tational cost. On the contrary, we use the global frame-level feature in $\mathbb{R}^{L_v \times D}$ and filter out some semantically irrelevant components for more balanced alignment, getting better results than VDI with much less computational cost.

TACoS. Different from Charades-STA, there are much fewer but longer videos in TACoS with much more sentences within a video. Table 2 shows the comparison with the state-of-the-art methods. We achieve the best in all metrics. Note MS-DETR uses multi-scale video features, which is beneficial for results, but it still suffers from the modality imbalance problem and is thus sub-optimal, we obtain 5.03% and 2.16% gain in R1@0.3 and R1@0.5, respectively.

QVHighlights. QVHighlights is a special dataset as each video only contains one sentence, which is quite challenging for our SS-MESM, making it only learn the complementary knowledge from the video modality. As shown in Table 3, we also obtain gains in all metrics on the *test* split compared with QD-DETR. When FW-MESM works normally, we analyze that the masked sentence in SS-MESM tends to perform as a prompt to learn from various videos.

Charades-CG. When the MESM bridges the modality gap, the model should be more generalizable to understand the relationship between videos and language queries. To validate it, we conduct experiments on the OOD settings of Charades-STA. Table 4 shows the comparison. The novel-composition set contains the unseen combination of observed constituents, and the novel-word set contains novel words for a sentence. For the novel-composition set, compared with VISA, we gain considerably (3.29%) in R1@0.7, which means we can get answers with higher quality due to better modality alignment. For the novel-word set, we obtain inspiring gains (12.79% in R1@0.7). The reason can be concluded from both FW-MESM and SS-MESM. When FW-MESM provides fine-grained discrimination ability to understand novel words from frames and other words, SS-MESM also supplements additional semantics for better understanding. Compared with QD-DETR, we also gain 5.03% in R1@0.7, demonstrating better generalization ability.

SS-MESM layers	R1@0.5	R1@0.7	mIoU	mAP_{avg}
2	57.10	35.32	49.75	36.26
3	59.38	35.22	50.68	37.90
4	**61.24**	38.04	**52.14**	**38.57**
5	60.97	**38.39**	52.09	38.50
6	58.25	35.70	50.31	37.60

Table 7: Ablation study (%) on the layers of SS-MESM.

Ablation Study

To validate the effectiveness of each component, we conduct ablation studies on Charades-STA with C+SF as the video extractor and C as the text extractor.

Main Ablation. The key components of our MESM are the two different levels of semantic modeling, FW-MESM and SS-MESM, while we also add a loss function $\mathcal{L}_{enc}$ for the transformer encoder. As shown in Table 5, each component is beneficial for VMR and $\mathcal{L}_{enc}$ makes the most of it since it provides a supervised signal for modal-aligned features to figure out the correct segment. Without FW-MESM and SS-MESM, the framework is similar to QD-DETR, so is the performance. Based on the result of adding $\mathcal{L}_{enc}$, FW-MESM and SS-MESM achieve gains of 1% and 1.24% in mIoU, respectively. When we use all of them, the result comes to the best, which gains 3.04% in R1@0.7 and 2.23% in mIoU,

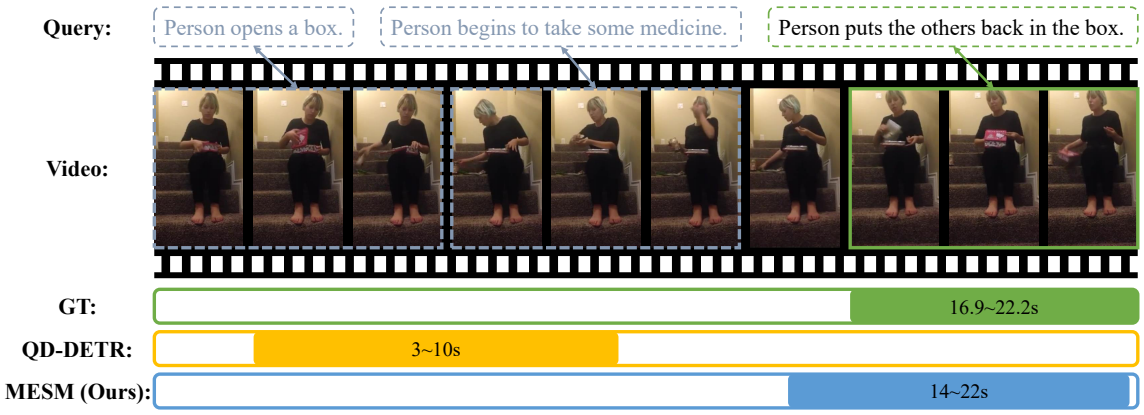

Figure 4: Visualization of prediction on Charades-STA. Model needs to understand what the *others* are, which is challenging.

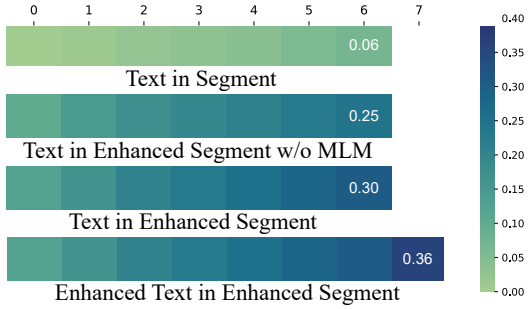

Figure 5: Visualization of the level of text modality within the video modality, a reference to evaluate alignment.

demonstrating that both of them are beneficial to the results. This is because they both make the alignment more balanced and are compatible with each other.

The Weight-Shared MLM. To make sure it is the weight-shared MLM works for FW-MESM, instead of the extra cross-attention layers, we conduct the ablation study as shown in Table 6. When SS-MESM is not implemented, we set 2 layers of cross-attention (2×MA) as baseline (line 1). If we simply add the layers of cross-attention to 4 without MLM (line 2), the scores drop, which may be caused by overfitting. 2 layers of FW-MESM and 2 layers of MA together achieve the best (line 3). When SS-MESM is implemented, all metrics drop with a margin without MLM. These results demonstrate the effectiveness of the MLM on the weight-shared cross-attention, instead of the extra layers.

The Layers of SS-MESM. We also conduct the ablation study on the SS-MESM to figure out the suitable number of layers. As shown in Table 7, the mIoU and mAP_{avg} come to the best when implementing 4 layers of SS-MESM, too few layers can not provide enough power to learn the semantics while too many layers may cause overfitting.

Qualitative Analysis

In Figure 4, we show an example of prediction. For the given query *Person puts the others back in the box*, the model should understand what the *others* are, which is challenging. Therefore, QD-DETR may only simply catch the words *puts* and *box*, then gives a wrong answer to similar scenes. When we supplement semantics and enhance both modalities, MESM understands the *others* stand for the things except for *medicine*, and thus gives a more accurate answer.

In Figure 5, we answer the question of how much of the text modality is contained in the video modality. We randomly select a query and calculate the subspace similarity based on the singular value decomposition (Hamm and Lee 2008; Hu et al. 2021). The calculation is implemented between the top-i singular vectors of the query and all singular vectors of the segment. The similarity is quite low between the original text $F_q^{\dagger}$ and original segment $F_v[l_e{:}l_s]$ (line 1), and becomes much higher when it comes to $F_q^{\dagger}$ and the enhanced segment $F_v^{enh}[l_e{:}l_s]$ (line 2&3) due to both the cross-modal interaction and the suppression of semantically irrelevant parts by the MLM task. The similarity between the enhanced text $F_q^{\dagger enh}$ and $F_v^{enh}[l_e{:}l_s]$ is the highest (line 4) owing to the complementary knowledge added to the textual modality. As the similarity stands for the level of containing, the results demonstrate both the enhanced video and textual modalities are more balanced than before.

Conclusion

In this paper, we address the modality imbalance problem, which means the inherently richer semantic information in the video modality than the textual modality. The imbalance makes the direct alignment sub-optimal but most methods ignore it. Therefore, we propose a novel framework MESM to tackle this problem from two levels. At the frame-word level, we enhance the video modality to filter out the redundant query-irrelevant semantics, making it more balanced with the texts. At the segment-sentence level, we enhance the textual modality to capture more segment-relevant semantics, making it more balanced with the videos. In the future, we aim to design a framework to model the semantic information progressively from the frame-word level to the segment-sentence level to achieve robust alignment, and we believe the issue and solution introduced in this work can provide fundamental insights to related fields.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (2022YFB3104700), the National Nature Science Foundation of China (62121002, U23B2028, 62232006).

References

Anne Hendricks, L.; Wang, O.; Shechtman, E.; Sivic, J.; Darrell, T.; and Russell, B. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, 5803–5812.

Bao, H.; Wang, W.; Dong, L.; Liu, Q.; Mohammed, O. K.; Aggarwal, K.; Som, S.; Piao, S.; and Wei, F. 2022. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35: 32897–32912.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

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

Chen, S.; and Jiang, Y.-G. 2019. Semantic proposal for activity localization in videos via sentence query. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 8199–8206.

Ding, X.; Wang, N.; Zhang, S.; Cheng, D.; Li, X.; Huang, Z.; Tang, M.; and Gao, X. 2021. Support-set based cross-supervision for video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11573–11582.

Du, Y.; Chen, Z.; Jia, C.; Yin, X.; Zheng, T.; Li, C.; Du, Y.; and Jiang, Y.-G. 2022. Svtr: Scene text recognition with a single visual model. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence*.

Escorcia, V.; Soldan, M.; Sivic, J.; Ghanem, B.; and Russell, B. 2019. Temporal localization of moments in video collections with natural language. *arXiv preprint arXiv:1907.12763*.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6202–6211.

Gao, J.; Sun, C.; Yang, Z.; and Nevatia, R. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, 5267–5275.

Gao, J.; Sun, X.; Xu, M.; Zhou, X.; and Ghanem, B. 2021. Relation-aware Video Reading Comprehension for Temporal Language Grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 3978–3988.

Ge, J.; Xie, H.; Min, S.; Li, P.; and Zhang, Y. 2022. Dual Part Discovery Network for Zero-Shot Learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, 3244–3252.

Ge, J.; Xie, H.; Min, S.; and Zhang, Y. 2021. Semantic-guided reinforced region embedding for generalized zero-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 1406–1414.

Hamm, J.; and Lee, D. D. 2008. Grassmann discriminant analysis: a unifying view on subspace-based learning. In *Proceedings of the 25th international conference on Machine learning*, 376–383.

Hu, E. J.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ko, D.; Choi, J.; Ko, J.; Noh, S.; On, K.-W.; Kim, E.-S.; and Kim, H. J. 2022. Video-text representation learning via differentiable weak temporal alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5016–5025.

Lei, J.; Berg, T. L.; and Bansal, M. 2021. QVHighlights: Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34: 11846–11858.

Lei, J.; Yu, L.; Berg, T. L.; and Bansal, M. 2020. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, 447–463. Springer.

Li, J.; Xie, J.; Qian, L.; Zhu, L.; Tang, S.; Wu, F.; Yang, Y.; Zhuang, Y.; and Wang, X. E. 2022a. Compositional temporal grounding with structured variational cross-graph correspondence learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3032–3041.

Li, K.; Guo, D.; and Wang, M. 2021. Proposal-free video grounding with contextual pyramid network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 1902–1910.

Li, P.; Xie, C.-W.; Xie, H.; Zhao, L.; Zhang, L.; Zheng, Y.; Zhao, D.; and Zhang, Y. 2023a. MomentDiff: Generative Video Moment Retrieval from Random to Real. In *Advances in neural information processing systems*.

Li, P.; Xie, C.-W.; Zhao, L.; Xie, H.; Ge, J.; Zheng, Y.; Zhao, D.; and Zhang, Y. 2023b. Progressive Spatio-Temporal Prototype Matching for Text-Video Retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4100–4110.

Li, P.; Xie, H.; Ge, J.; Zhang, L.; Min, S.; and Zhang, Y. 2022b. Dual-Stream Knowledge-Preserving Hashing for Unsupervised Video Retrieval. In *ECCV*, 181–197. Springer Nature Switzerland.

Liu, D.; Qu, X.; Di, X.; Cheng, Y.; Xu, Z.; and Zhou, P. 2022a. Memory-guided semantic learning network for temporal sentence grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1665–1673.

Liu, D.; Qu, X.; Dong, J.; Zhou, P.; Cheng, Y.; Wei, W.; Xu, Z.; and Xie, Y. 2021a. Context-aware biaffine localizing network for temporal sentence grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11235–11244.

Liu, S.; Li, F.; Zhang, H.; Yang, X.; Qi, X.; Su, H.; Zhu, J.; and Zhang, L. 2021b. DAB-DETR: Dynamic Anchor Boxes are Better Queries for DETR. In *International Conference on Learning Representations*.

Liu, Y.; Li, S.; Wu, Y.; Chen, C.-W.; Shan, Y.; and Qie, X. 2022b. 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*, 3042–3051.

Luo, D.; Huang, J.; Gong, S.; Jin, H.; and Liu, Y. 2023. Towards Generalisable Video Moment Retrieval: Visual-Dynamic Injection to Image-Text Pre-Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23045–23055.

Moon, W.; Hyun, S.; Park, S.; Park, D.; and Heo, J.-P. 2023. Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23023–23033.

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

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

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Regneri, M.; Rohrbach, M.; Wetzel, D.; Thater, S.; Schiele, B.; and Pinkal, M. 2013. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1: 25–36.

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

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

Sun, X.; Wang, X.; Gao, J.; Liu, Q.; and Zhou, X. 2022. You need to read again: Multi-granularity perception network for moment retrieval in videos. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1022–1032.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 4489–4497.

Union, G. I. O. 2019. A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 658–666.

Wang, H.; Zha, Z.-J.; Li, L.; Liu, D.; and Luo, J. 2021. 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*, 7026–7035.

Wang, J.; Sun, A.; Zhang, H.; and Li, X. 2023a. MS-DETR: Natural Language Video Localization with Sampling Moment-Moment Interaction. *arXiv preprint arXiv:2305.18969*.

Wang, W.; Bao, H.; Dong, L.; Bjorck, J.; Peng, Z.; Liu, Q.; Aggarwal, K.; Mohammed, O. K.; Singhal, S.; Som, S.; et al. 2023b. Image as a Foreign Language: BEiT Pretraining for Vision and Vision-Language Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19175–19186.

Wang, Z.; Wang, L.; Wu, T.; Li, T.; and Wu, G. 2022. Negative sample matters: A renaissance of metric learning for temporal grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2613–2623.

Wu, W.; Luo, H.; Fang, B.; Wang, J.; and Ouyang, W. 2023. Cap4Video: What Can Auxiliary Captions Do for Text-Video Retrieval? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10704–10713.

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

Zhang, B.; Xie, H.; Wang, Y.; Xu, J.; and Zhang, Y. 2023. Linguistic More: Taking a Further Step toward Efficient and Accurate Scene Text Recognition. *arXiv preprint arXiv:2305.05140*.

Zhang, H.; Sun, A.; Jing, W.; Zhen, L.; Zhou, J. T.; and Goh, S. M. R. 2021. Parallel Attention Network with Sequence Matching for Video Grounding. In *Findings of the Association for Computational Linguistics*, 776–790.

Zhang, H.; Sun, A.; Jing, W.; and Zhou, J. T. 2020a. Span-based Localizing Network for Natural Language Video Localization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6543–6554.

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

Zheng, T.; Chen, Z.; Fang, S.; Xie, H.; and Jiang, Y.-G. 2023. Cdistnet: Perceiving multi-domain character distance for robust text recognition. *International Journal of Computer Vision*, 1–19.