Aligning Moments in Time using Video Queries

Yogesh Kumar[1*] Uday Agarwal[1*] Manish Gupta[2] Anand Mishra[1]
[1]Indian Institute of Technology Jodhpur [2]Microsoft
{kumar.204, agarwaluday, mishra}@iitj.ac.in, gmanish@microsoft.com

Abstract

*Video-to-video moment retrieval (*Vid2VidMR*) is the task of localizing unseen events or moments in a target video using a query video. This task poses several challenges, such as the need for semantic frame-level alignment and modeling complex dependencies between query and target videos. To tackle this challenging problem, we introduce MATR (Moment Alignment TRansformer), a transformer-based model designed to capture semantic context as well as the temporal details necessary for precise moment localization. MATR conditions target video representations on query video features using dual-stage sequence alignment that encodes the required correlations and dependencies. These representations are then used to guide foreground/background classification and boundary prediction heads, enabling the model to accurately identify moments in the target video that semantically match with the query video. Additionally, to provide a strong task-specific initialization for MATR, we propose a self-supervised pre-training technique that involves training the model to localize random clips within videos. Extensive experiments demonstrate that MATR achieves notable performance improvements of 13.1% in R@1 and 8.1% in mIoU on an absolute scale compared to state-of-the-art methods on the popular ActivityNet-VRL dataset. Additionally, on our newly proposed dataset, SportsMoments, MATR shows a 14.7% gain in R@1 and a 14.4% gain in mIoU on an absolute scale over strong baselines. We make the dataset and code public at: https://github.com/vl2g/MATR.*

1. Introduction

Video moment retrieval is the task of temporally localizing the start and end times of a moment[1] in a target video described by a given query. Although text-based video moment retrieval has been extensively explored [25, 28, 32, 33, 57], it often poses challenges for users attempting to describe specific moments verbally. For example, consider a spectacular *bicycle kick* in soccer (Fig. 1 top). Although an expert can search by naming or describing it in detail, a beginner in soccer may struggle to describe this impressive move accurately. They might say something like "The player kicked the ball while in the air", which lacks the nuance needed to convey the athleticism and artistry involved. Further, such a description may lead to poor results for someone searching for similar moments. In contrast, if a soccer trainee shows a short video clip of a *bicycle kick*, the input is rich and crisp, making it easier for a retriever to locate similar moments in a target soccer match video. This approach aligns with how users naturally would prefer to search content within the same (video) modality, making it a more effective method for retrieving specific moments. Fig. 1 (bottom) shows a video *bicycle kick* query in soccer that is localized in a target video.

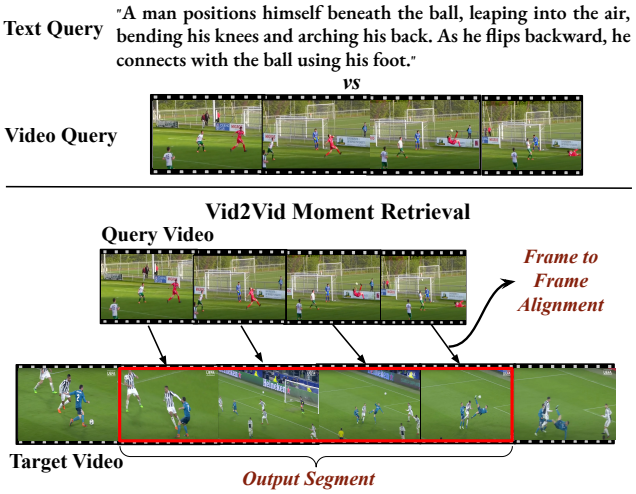

Figure 1. (**Top**): Activity where text falls short to explain complex action or events, such as *bicycle kick*, necessitating the need for a more intuitive query modality for localizing semantically matching moments. (**Bottom**): Shows the Vid2Vid moment retrieval setting, which is the goal of this work.

Therefore, in this work, we study video-to-video moment retrieval (*Vid2VidMR*) where the aim is to temporally localize a moment in the target video with a high se-

*Equal Contribution

[1]A moment is a continuous segment of frames within a target video that best represents the actions, events, or interactions described by the query.

mantic match between the moment and the query video. *Vid2VidMR*, a task formally introduced by Feng et al. [14], has many potential application areas, such as sports video analytics, educational content creation and e-learning, and surveillance systems. This task is challenging and requires semantic frame-level alignment and modeling complex dependencies between query and target videos. The need for a semantic understanding of video content and variety in video length, context, and action speed calls for adaptive models capable of generalizing across diverse scenarios. Addressing these challenges demands advanced temporal video modeling techniques.

To address the aforementioned challenges, we propose MATR (Moment Alignment TRansformer) – a method that uses explicit 'dual-stage sequence alignment' to capture the required correlation and temporal details essential for accurate moment localization. By conditioning target video representations on query features, MATR produces query-aligned representations that encode the necessary correlations and temporal dependencies between the two videos. We use differentiable dynamic time warping loss [9] for aligning the query and target videos, and represent the target video by conditioning it on the query video to focus on correlated temporal features. These representations guide a classification head to discriminate relevant moments from the irrelevant background and a boundary prediction head to mark the start and end of the identified moment in the target video. Further, to enhance our model's generalization using unlabeled videos, we introduce a self-supervised pre-training strategy which involves training MATR to localize randomly sampled clips within the same video, enabling it to learn the moment localization skill in a self-supervised manner.

We evaluate MATR on public ActivityNet-VRL [14] benchmark and on our newly introduced dataset on sports domain, viz. SportsMoments, covering two of the most popular sports, namely soccer and cricket. Our approach achieves significant performance gains, with an improvement of 13.1% and 8.1% in R@1 and mIoU, respectively, on ActivityNet-VRL outperforming the state-of-the-art methods. Furthermore, MATR outperforms the implemented strong baselines on our proposed dataset with 14.7% gains in R@1 and 14.4% gains in mIoU, all on an absolute scale.

Our contributions are as follows: (i) We introduce MATR, which uses explicit dual-stage sequence alignment within a transformer framework between target and query video to capture temporal correlations and dependencies for accurate moment localization. (ii) We propose a self-supervised pre-training objective that enhances model initialization by understanding rich video structure without requiring any labeled data. (iii) We conduct extensive experiments and ablations to study the efficacy of our framework against competitive baselines and state-of-the-art methods. Our findings offer valuable insights into our design choices, and our approach advances the state-of-the-art on *Vid2VidMR*.

2. Related Work

Video Moment Retrieval (VMR): VMR has recently gained significant interest in the research community [1, 15, 18, 24, 25, 28, 29, 32–34, 53, 55, 57]. Unlike video action understanding tasks such as action classification [6, 13, 27, 44, 47, 54, 56, 60] or temporal action localization [43, 59, 64], VMR focuses on identifying segments that semantically align with a broader range of queries, which may describe complex and context-specific moments. Based on the query modality, existing VMR methods can be broadly grouped into (i) textual query-based approaches like Moment-DETR [25], QD-DETR [33], UniVTG [28], and (ii) video query-based approaches like GDP [7], FFI+SRM [19], SRL [51]. Our work falls under video query-based VMR (*Vid2VidMR*). However, in addition to developing a new approach tailored for video query, we also adapt several text-query-based methods to make them suitable for *Vid2VidMR* and compare our method against both text and video query-based methods.

Alignment in Videos: Alignment has been a thrust area in the video understanding community. It has been applied to a wide range of video understanding tasks, including video retrieval [52], procedural steps alignment [11], action recognition [10], anomaly detection [12], movie understanding [2] and video synchronization [35, 40]. Researchers have also explored sequence alignment for the text-VMR task, e.g., Mithun et al. [31] used sequence alignment between CNN representation of frames and GRU representation of text query to perform text-VMR. Jung et al. [20] used alignment to enhance semantic understanding between query and target video at the abstract level for the text-VMR task. However, their alignment is not at the sequence level.

Compared to the existing literature, MATR goes beyond traditional sequence alignment by introducing a dual-stage alignment mechanism, leveraging transformer-based feature fusion, and maintaining flexibility with multiple alignment strategies. These innovations enable it to achieve more accurate and context-aware video moment retrieval compared to existing approaches.

3. The MATR Model

Our objective is to temporally localize a moment in a target video V_t using a query video V_q. We represent the target and query videos as sequences of M and N uniformly sampled frames, respectively. We refer to this problem as *Vid2VidMR*. In this work, we present Moment Alignment

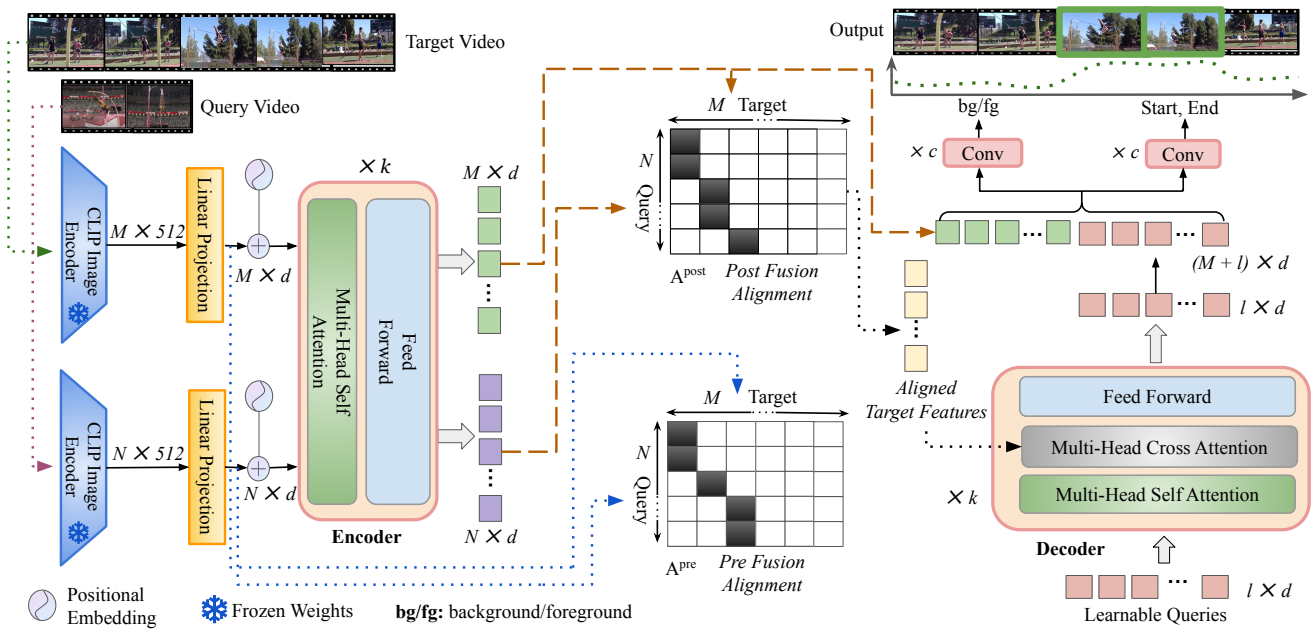

Figure 2. Moment Alignment TRansformer (MATR) Architecture. We represent the target video as a query-aligned representation capturing the required correlation and dependencies with query video at the abstract level (output of encoder) and finer level (output of decoder). These features are used to extract foreground target frames, as well as localize the relevant moment in terms of start and end times via two prediction heads. The alignment is computed using representations both before and after the encoder (shown using blue dotted and red dashed lines, respectively). **[Best viewed in color]**.

TRansformer (MATR) – a transformer-based architecture that leverages explicit alignment for precise moment localization. MATR transforms the target video into a query-aligned representation via a transformer framework to exploit finer-level correlation and dependencies for precise moment localization.

3.1. Motivation behind MATR Architecture

Video moments vary significantly in duration, motion patterns, and visual appearance, making their retrieval inherently challenging. Additionally, preserving the correct temporal order of events is crucial for accurate localization, as moments often involve complex interactions that unfold over time. A robust *Vid2VidMR* model must, therefore, capture both high-level semantic relationships and fine-grained frame dependencies to ensure precise alignment between the query and target video.

To address these challenges, MATR incorporates an explicit *dual-stage sequence alignment* strategy within a transformer-based encoder-decoder framework. This strategy enables MATR to learn a *query-aligned representation* of the target video by combining the abstract representation from the encoder with the refined fine-grained features from the decoder. In doing so, it facilitates precise moment localization by leveraging both global semantic alignment and fine-grained temporal dependencies between the query and target videos. Figure 2 illustrates the overall architecture of MATR, which we discuss in detail in the following.

3.2. Architecture Details

Input Representation. We uniformly sample frames every 2 seconds to obtain M and N frames from the target and query videos, respectively. We encode these frames using a frozen CLIP [39] (ViT-B/32) encoder along with a linear projection module. CLIP is applied separately to both videos to obtain target video embeddings $\mathrm{E}_t \in \mathbb{R}^{M\times d}$ and query video embeddings $\mathrm{E}_q \in \mathbb{R}^{N\times d}$. The linear projection module is a two-layer perceptron, each with layer normalization and dropout. It maps 512-dimensional CLIP embeddings to d-dimensional outputs. The resulting projected features (E_t and E_q) are concatenated along the sequence length dimension to form the input to the transformer encoder, denoted as $\mathrm{E}_c = [\mathrm{E}_t; \mathrm{E}_q] \in \mathbb{R}^{(M+N)\times d}$.

Encoder. The combined target and query video representation (E_c) is processed through a series of k standard Transformer [48] encoder layers, each comprising a multi-head self-attention mechanism and a feed-forward network. Further, following the prior works [3, 5, 36], fixed positional encodings are added to the input of each attention layer to preserve temporal order. The encoder generates a fused representation $[\mathrm{E}_\mathrm{t}^\mathrm{g}; \mathrm{E}_\mathrm{q}^\mathrm{g}]$ of the target video conditioned on the query video. Here $\mathrm{E}_\mathrm{t}^\mathrm{g} \in \mathbb{R}^{M\times d}$ and $\mathrm{E}_\mathrm{q}^\mathrm{g} \in \mathbb{R}^{N\times d}$ represent the target and query parts respectively.

Dual-stage Sequence Alignment. The transformer encoder computes effective features by performing a joint understanding (or fusion) of target and query video frames.

The final goal of *Vid2VidMR* is to localize a moment in the target video that semantically matches the content of the query video. In other words, we would prefer the features of the moment in the target to align strongly with the features of the query video. Therefore, we perform dual-stage sequence alignment, i.e, before and after the encoder. For alignement, we choose soft-DTW [9][2] and perform the alignment as follows.

Pre-fusion alignment: Before encoder fusion, given the target and query video feature sequences, $\mathrm{E_t} = [\boldsymbol{e}_1^t, \ldots, \boldsymbol{e}_M^t] \in \mathbb{R}^{M \times d}$ and $\mathrm{E_q} = [\boldsymbol{e}_1^q, \ldots, \boldsymbol{e}_N^q] \in \mathbb{R}^{N \times d}$, soft-DTW outputs a binary alignment matrix $\mathrm{A^{pre}}$ and an alignment cost matrix $\mathrm{C^{pre}}$. $\mathrm{A^{pre}} \in \{0,1\}^{M \times N}$ is an alignment matrix such that:

$$\mathrm{A}_{i,j}^{\mathrm{pre}} = \begin{cases} 1 & \text{if } \boldsymbol{e}_i^t \text{ is matched to } \boldsymbol{e}_j^q, \\ 0 & \text{otherwise.} \end{cases}$$

further, soft-DTW finds the optimal alignment by minimizing the pre-fusion alignment loss $\mathcal{L}_{\text{align}}^{\text{pre}}$, defined as:

$$\mathcal{L}_{\text{align}}^{\text{pre}} = \text{soft-DTW}_\gamma(\mathrm{A}_{i,j}^{\text{pre}}, \mathrm{C}_{i,j}^{\text{pre}}),$$

such that aligned frames in the target are contiguous. Here γ is a smoothing factor for soft-min operator. The value of γ in Soft-DTW is selected empirically to balance smoothness and alignment fidelity. Note that $\mathrm{C^{pre}}$ is the alignment cost matrix with elements defined by cosine similarity:

$$\mathrm{C}_{i,j} = 1 - \frac{\langle \boldsymbol{e}_i^t, \boldsymbol{e}_j^q \rangle}{\|\boldsymbol{e}_i^t\| \|\boldsymbol{e}_j^q\|}.$$

Pre-fusion alignment enhances the semantic representation of target and query video. These enhanced representations are further processed by the encoder. The encoder outputs the fused representation of the target and query.

Post-fusion alignment: Given the post-fusion target and query video feature sequences, $\mathrm{E_t^g} \in \mathbb{R}^{M \times d}$ and $\mathrm{E_q^g} \in \mathbb{R}^{N \times d}$, soft-DTW outputs a binary alignment matrix $\mathrm{A^{post}}$ and an alignment cost matrix $\mathrm{C^{post}}$, soft-DTW achieves optimal alignment between $\mathrm{E_t^g}$ and $\mathrm{E_q^g}$ by minimizing the post-fusion alignment loss $\mathcal{L}_{\text{align}}^{\text{post}}$, defined as:

$$\mathcal{L}_{\text{align}}^{\text{post}} = \text{soft-DTW}_\gamma(\mathrm{A}_{i,j}^{\text{post}}, \mathrm{C}_{i,j}^{\text{post}}).$$

Post-fusion alignment refines target video features by leveraging fused query-target representations, ensuring fine-grained semantic matching for precise moment localization.

Decoder. The decoder in MATR further refines the target video representation by processing aligned query-target features, enabling fine-grained temporal matching and enhancing frame-level precision for accurate moment localization. To this end, the post-fusion alignment matrix $\mathrm{A^{post}}$ defines a contiguous sub-sequence $\mathrm{E_t^g}[s : e]$ which aligns best with the query video. These aligned target features, capturing query-aligned fine-grained information, are passed as input to the decoder for further refinement. The decoder consists of k Transformer layers, each containing a multi-head self-attention layer, a multi-head cross-attention layer, and a feed-forward network. The input to the decoder comprises fixed size l learnable query vectors denoted by $\mathrm{Q} \in [q_1, \ldots, q_l] \in \mathbb{R}^{l \times d}$. These queries guide the extraction of refined features relevant to moment localization. As these queries are processed through each decoder layer, they are refined to capture intricate temporal dependencies, with positional encodings applied at each attention layer. The cross-attention layers enable interaction between $\mathrm{E_t^g}[s : e]$ from the encoder (which serve as keys and values) and representations of moment queries (which serve as queries). The final decoder output, $\mathrm{E_t^l} \in \mathbb{R}^{l \times d}$, provides a refined fine-grained set of target video features conditioned on the query video.

Finally, the encoder and decoder representations are combined as $\mathrm{E_f} = [\mathrm{E_t^g}; \mathrm{E_t^l}]$, making the final target video representation query-aligned. The final *query-aligned representation*, combining the abstract and fine-grained semantics, serves as input for the prediction heads.

Prediction Heads. The prediction heads consist of three ($c = 3$) sequential convolutional layers each having d, 1×3 kernels followed by a ReLU activation. Finally, a sigmoid activation function is applied to produce the foreground predictions $\hat{f}_i$. The model is trained using the following binary cross-entropy loss $\mathcal{L}_{fg}$ to distinguish foreground (relevant moment) from background predictions:

$$\mathcal{L}_{fg} = -\left(f_i \log \hat{f}_i + (1 - f_i) \log(1 - \hat{f}_i)\right),$$

where f_i is the ground truth label, with $f_i = 1$ and $f_i = 0$ indicating foreground and background, respectively.

In addition to foreground classification, the boundary prediction head is designed to estimate the start and end boundaries for the moment at every position. This head shares the same initial structure as the classification head but differs in the final output, producing two channels for left and right boundary offsets relative to the current position. Specifically, given $\mathrm{E_f}$, the boundary prediction head outputs predicted offsets $\hat{d}_i = \{\hat{d}_i^L, \hat{d}_i^R\}$ for each position $i \in [0, M-1]$, where $\hat{d}_i^L$ and $\hat{d}_i^R$ represent the left and right boundary offsets, respectively. The boundary prediction head is trained with a combination of smooth $L1$ and generalized intersection over union (IoU) loss [42], specifically applied to foreground positions ($f_i = 1$), defined as:

$$\mathcal{L}_{\text{seg}} = \mathbb{1}_{f_i=1} \left[\lambda_{\text{L1}} \mathcal{L}_{\text{L1}}\left(\hat{d}_i, d_i\right) + \lambda_{\text{IoU}} \mathcal{L}_{\text{IoU}}\left(\hat{d}_i, d_i\right)\right].$$

Here, d_i denotes the ground truth offset. The parameters

[2]Soft-DTW allows non-linear alignments between sequences of different lengths, is differentiable, robust to variations in speed, and can handle noise and outliers, though other alignment algorithms, such as TCC [12] and DropDTW [11] can also be used in our model.

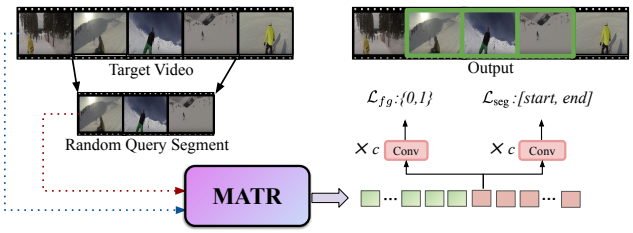

Figure 3. Proposed self-supervised pre-training strategy: Given a target video, a query clip is randomly sampled from it and processed by the model. The model then predicts the boundaries of the selected clip within the target video, highlighting the corresponding frames with a green border. **(Best viewed in color)**

λ_{L1} and λ_{IoU} weigh the contributions of the smooth L1 and IoU losses, ensuring that boundary localization focuses on the predictions identified as foreground.

Convolutional layers operate along the temporal axis over the concatenated encoder query-conditioned and decoder-aligned target features. This preserves temporal continuity, enabling fine-grained predictions without boundary artifacts.

Overall Loss. The overall loss function $\mathcal{L}$ is computed as a combination of multiple objectives, accounting for alignment costs (pre- and post-fusion alignment), foreground/background classification, and boundary localization. For S training samples, this multi-objective overall loss is defined as:

$$\mathcal{L} = \frac{1}{S}\sum_{i=1}^{S}\left(\lambda_{fg}\mathcal{L}_{fg} + \lambda_{\text{seg}}\mathcal{L}_{\text{seg}} + \lambda_{\text{align}}^{\text{pre}}\mathcal{L}_{\text{align}}^{\text{pre}} + \lambda_{\text{align}}^{\text{post}}\mathcal{L}_{\text{align}}^{\text{post}}\right),$$

where the λs control the importance of each component.

Inference. During inference, given a target video V_t and a query video V_q, we pass both through the model to generate foreground probabilities $\{\hat{f}_i\}_{i=0}^{M-1}$ and boundary predictions $\{\hat{d}_i\}_{i=0}^{M-1}$ using the two prediction heads. To handle the densely generated boundaries, while predicting $\hat{d}_i$, we apply 1-dimensional non-maximal suppression with a threshold of 0.7, filtering out highly overlapping boundaries and producing the final set of predictions. We choose $\hat{d}_i$ corresponding to position i with highest $\hat{f}_i$.

3.3. Self-supervised Pre-training Strategy

The video representation learning community has leveraged self-supervised pre-training techniques designed for training effective encoders [22, 23, 37, 38, 45, 49, 50], which can be finetuned for enhanced performance on downstream tasks. In this work, we introduce a self-supervised pre-training objective designed to improve temporal localization capabilities without relying on labeled data. Specifically, given a target video V_t, a query clip V_q is randomly sampled from V_t as shown in Fig. 3. The model is then trained to localize this query clip in V_t. This pre-training objective closely aligns with the task of *Vid2VidMR*. To enhance generalization capabilities and encourage the model to become robust to variations in both appearance and timing, we apply one of the following random augmentations to each of the query clips: reversing frames, adding gaussian noise, slowing down or speeding up the action, thereby doubling the number of pre-training samples. These augmentations introduce temporal and spatial variations, enabling the model to learn diverse representations.

The overall pre-training loss, $\mathcal{L}_{pt}$, combines the foreground-background classification loss, the boundary prediction loss, pre-fusion and post-fusion alignment costs. The total loss is averaged over P pre-training samples as:

$$\mathcal{L}_{pt} = \frac{1}{P}\sum_{i=1}^{P}\left(\lambda_{fg}^{p}\mathcal{L}_{\text{fg}} + \lambda_{seg}^{p}\mathcal{L}_{\text{seg}} + \lambda_{\text{align}}^{\text{pre}}\mathcal{L}_{\text{align}}^{\text{pre}} + \lambda_{\text{align}}^{\text{post}}\mathcal{L}_{\text{align}}^{\text{post}}\right),$$

where the λs control the importance of each component.

4. Datasets

We use the following two datasets to compare our proposed approach with existing *Vid2VidMR* methods and other strong baselines:

(i) ActivityNet-VRL Dataset [14]: ActivityNet-VRL is the popular benchmark dataset shared by Feng et al. [14], based on the ActivityNet video understanding benchmark [17] consisting of 200 action classes. The dataset is split into disjoint 160 classes for training and 20 classes each for the validation and test splits. Furthermore, the training set comprises of $\sim$ 463K query-target video pairs, while the validation set and the test set have 829 and 978 query-target video pairs, respectively.

(ii) Our proposed SportsMoments Dataset[3]: One of the most promising applications for *Vid2VidMR* lies in the area of sports analytics. Although there exist large-scale sports datasets such as Sports-1M [21], they are primarily tailored for broader sports video classification tasks and do not contain fine-grained sports actions and moments, such as "Cover Drive," "Ducking a Bouncer," "Goal Kick," or "Penalty." Secondly, although ActivityNet-VRL covers a wide range of action categories (including sports actions), it only encompasses broader categories such as "Playing Polo" and does not focus on capturing higher-granularity events within a sport. Towards filling this gap, we introduce the SportsMoments dataset, which consists of $\sim$770K query-target pairs annotated from 176.6 hours of complete match footage of two of the most popular sports, viz. soccer and cricket. We obtain a total of 80 full-length cricket and soccer full-length match videos from YouTube. We then curate a list of 29 action classes comprising 13 soccer and 16 cricket actions, respectively. Given this list, we employed two annotators, each with strong knowledge of these sports,

[3] https://github.com/vl2g/MATR/tree/main/sportsmoments

to mark the start and end timestamps for the specified actions in the videos. We split SportsMoments into training, validation and test sets. The training split consists of approximately 750K pairs spanning 16 classes, while the validation and test splits each contain 10K pairs, covering four and nine classes, respectively. Action classes are disjoint across the train, validation, and test sets, ensuring no overlap. Additionally, each split includes cricket and soccer actions for a well-rounded distribution.

Additionally, we leverage the unlabeled videos from the Kinetics700 [6] dataset for pre-training.

5. Experiments and Results

5.1. Baselines

We experiment with an extensive set of strong baselines grouped into four categories as follows.

(i) **Fully-supervised *Vid2VidMR* methods:** We compare with *Vid2VidMR* methods like CGBM [14], GDP [7], SRL [51], FFI+SRM [19], SST [4] and Video-level match [14]. Further, Huo et al. [19] adapt text-VMR methods like VSLNet [61], MABAN [46] and 2D-TAN [63] by replacing their text encoders with a C3D [47] feature extractor to extract query video features. All of these models have been trained on ActivityNet-VRL.

(ii) **Vision-Language Models (VLMs):** VLMs like Video-LLaMa [62], Video-LLaVa [26] and TimeChat [41] have shown promising results for multimodal text-video tasks like Video VQA, captioning, etc. Unfortunately, they have not been pre-trained with video-video aligned data. First we represent the query video using a state-of-the-art captioner, i.e., mPLUG-OWL [58]. Next, the VLMs are zero-shot prompted to generate start and end moment timestamps given query caption and target video tokens.

(iii) **Text-VMR methods:** Moment-DETR [25], QD-DETR [33], CG-DETR [32], and UniVTG [28] have been originally proposed for text-VMR. We compare with five variants of these methods. The zero-shot variant (a) uses caption from mPLUG-OWL [58] to represent query video, and leverages the pretrained checkpoints. Variants (b) and (c) both use caption from mPLUG-OWL [58] to represent query video, but train a randomly initialized checkpoint (variant b) or finetune the pre-trained checkpoint (variant c), respectively. Variants (d) and (e) are equivalent to (b) and (c) where their CLIP text encoder is replaced by CLIP ViT/B-32 vision encoder, and therefore take query video directly as input along with the target video.

(iv) **Image-VMR methods:** Our work focuses on video moment retrieval using 'video' queries. As a single image query may not effectively capture the temporal aspects of a video query, this comparison may not be appropriate. However, we still design baselines by representing the video by its key-frame (more precisely, the middle frame) using Text-VMR methods, namely Moment-DETR, QD-DETR and UniVTG.

5.2. Implementation Details

We choose hidden dimension of 1024 with $k = 4$ layers in both encoders and apply a dropout rate of 0.1 and 0.5 within the transformer and linear projection layers, respectively. Model weights are initialized using Xavier initialization [16]. To optimize the model parameters, we utilize AdamW optimizer [30] with an initial learning rate of 1e-4 and a weight decay of 1e-4. Training was done on two NVIDIA A6000 GPUs. We trained our model for 200 epochs with a batch size of 1200 while using ActivityNet-VRL. For SportsMoments, we used 40 epochs with a batch size of 40. We set number of learnable queries $l = 10$ and all λs in both pre-training as well as finetuning losses to 1. We make our implementation and checkpoints available at: https://github.com/vl2g/MATR.

5.3. Results and Discussion

Main Results: Table 1 shows our main results where we compare MATR with fully-supervised *Vid2VidMR* methods, VLMs, image-VMR methods, and five variants of each of the Text-VMR methods. As mentioned before these variants correspond to zero-shot, finetuned or trained setup. As usual, trained setup implies random initialization while finetuned setup implies initialization using a pre-trained checkpoint (different for each architecture). Query video can be represented using text (T) captions obtained using mPLUG-OWL [58] or as video (V) itself. Since MATR is inherently designed to take video query as input, experimenting with text (T) captions and image query is not needed. Following previous work, we report the standard mean Intersection over Union (mIoU) and Recall@1 IoU=0.5 metrics.

Comparison with fully-supervised *Vid2VidMR* methods: For a fair comparison, we directly use the reported results from original papers for fully-supervised methods on ActivityNet-VRL. Consequently, we do not report their results on SportsMoments. As shown in Table 1, MATR outperforms the best method, FFI+SRM, with an 8.1% gain in mIoU and achieves a 13.1% improvement in R@1 over GDP, which had the highest recall. These results underscore MATR's superior ability to accurately localize moments using video queries.

Comparison with VLMs: Table 1 shows that our method MATR outperforms all four VLM-based baselines by a significant margin. In comparison, the best-performing baseline, TimeChat [41], achieves 26.4 mIoU and 23.8 R@1 on ActivityNet-VRL, and 22.6 mIoU and 21.3 R@1 on SportsMoments. Other methods, such as Video-LLaVA [26], Video-LLaMA [62], and Video-LLaMA2 [8], perform significantly worse. These results show that task-specific models for *Vid2VidMR* are significantly better.

	Methods	Variants	ActivityNet-VRL mIoU	ActivityNet-VRL R@1	SportsMoments mIoU	SportsMoments R@1
Fully-supervised Methods	Random [14]	-	7.3	16.2	-	-
	Video Match [14]	-	12.4	24.3	-	-
	SST [4]	-	17.1	33.2	-	-
	CGBM [14]	-	25.7	43.5	-	-
	GDP [7]	-	27.8	44.0	-	-
	SRL [51]	-	40.6	29.3	-	-
	2D-TAN [63]	-	45.3	39.6	-	-
	MABAN [46]	-	42.8	37.5	-	-
	VSLNet [61]	-	27.2	43.8	-	-
	FFI+SRM [19]	-	48.7	40.6	-	-
VLMs	Video-LLaVA [26]	-	15.1	14.7	13.8	11.9
	Video-LLaMA [62]	-	14.7	13.9	12.5	11.2
	Video-LLaMA2 [8]	-	17.6	15.2	14.7	13.4
	TimeChat [41]	-	26.4	23.8	22.6	21.3
I-VMR	Moment-DETR [25]	-	35.6	32.5	25.2	18.5
	QD-DETR [33]	-	37.2	38.1	27.7	20.7
	UniVTG [28]	-	38.8	42.0	34.6	37.2
Text-VMR Methods	Moment-DETR [25]	(a) ZS+T	28.2	22.8	3.7	0.9
		(b) Trained+T	37.1	35.8	31.4	29.2
		(c) Finetuned+T	40.0	38.9	35.8	34.0
		(d) Trained+V	35.9	34.7	28.7	24.2
		(e) Finetuned+V	40.0	39.9	30.4	25.1
	QD-DETR [33]	(a) ZS+T	25.4	22.2	6.3	2.6
		(b) Trained+T	38.4	39.0	29.8	27.2
		(c) Finetuned+T	41.6	45.2	36.0	33.5
		(d) Trained+V	39.7	41.0	27.2	25.1
		(e) Finetuned+V	42.5	42.7	35.4	30.6
	UniVTG [28]	(a) ZS+T	32.4	26.7	11.4	6.0
		(b) Trained+T	43.8	45.6	41.5	36.5
		(c) Finetuned+T	45.8	46.4	44.8	39.2
		(d) Trained+V	48.4	49.8	43.2	39.5
		(e) Finetuned+V	49.1	50.7	43.6	41.8
	CG-DETR [32]	(a) ZS+T	25.9	21.9	3.3	1.6
		(b) Trained+T	39.4	38.4	35.1	31.3
		(c) Finetuned+T	41.0	43.2	35.4	32.5
		(d) Trained+V	40.0	41.7	36.6	35.1
		(e) Finetuned+V	40.1	41.7	37.2	34.8
Ours	**MATR**	Zero-shot	32.6	30.1	31.8	30.7
		Trained+V	53.2	54.8	56.2	52.7
		Finetuned+V	**56.8**	**57.1**	**59.2**	**56.5**

Table 1. Comparing MATR with fully-supervised *Vid2VidMR* methods, VLMs, image-VMR methods, and variants of Text-VMR methods. Trained implies random initialization. Finetuned implies initialization using a pre-trained (respective to each architecture) checkpoint. Query video can be represented using text (T) captions or as video (V) itself. Since MATR is inherently designed to take video query as input, experimenting with text (T) captions is not needed. For more details of baselines and their variants, please refer to Section 5.1. Results for the overall best, best among implemented baselines and previously reported SOTA are highlighted in bold, underline and box, respectively.

Comparison with Text-VMR Methods: Table 1 shows that amongst all the *five variants*, the finetuned variants are typically better than the trained variants, i.e., variant (c) is better than (b), and (e) is better than (d). On both ActivityNet-VRL and SportsMoments, UniVTG achieves the best results among variants which use video queries. However, our MATR model outperforms UniVTG by 7.7% in mIoU and 6.4% in R@1 on ActivityNet-VRL. On SportsMoments, MATR shows a further improvement, surpassing UniVTG by 15.6% in mIoU and 14.7% in R@1. These gains highlight the superior performance of MATR compared to the text-VMR methods when used with video query input. When text captions are used to represent the query, we observe similar trends as in the video query input case. Among the baselines, UniVTG performs the best. However, MATR outperforms it by 11% in mIoU and 10.7% in R@1 on ActivityNet-VRL. On SportsMoments, MATR achieves 14.4% higher mIoU and 17.3% higher R@1 than UniVTG. These results demonstrate the superior performance of MATR compared to the text-VMR methods when used with text query input for *Vid2VidMR*.

Pre-Fusion	Post-Fusion	ActivityNet-VRL mIoU	ActivityNet-VRL R@1	SportsMoments mIoU	SportsMoments R@1
✗	✗	49.7	50.2	49.1	46.6
✗	✓	52.3	52.9	53.1	48.2
✓	✗	50.4	51.2	52.4	46.7
✓	✓	**53.2**	**54.8**	**56.2**	**52.7**

Table 2. Advantage of explicit dual-stage alignment of MATR. For this study, experiments were performed without pre-training.

Pre-training	Augmentation	ActivityNet-VRL mIoU	ActivityNet-VRL R@1	SportsMoments mIoU	SportsMoments R@1
✗	-	53.2	54.8	56.2	52.7
✓	✗	54.1	55.6	56.9	53.6
✓	✓	**56.8**	**57.1**	**59.2**	**56.5**

Table 3. Effect of pre-training and augmentation in MATR.

Comparison with Image-VMR: Table 1 shows that the best-performing approach, UniVTG [28], achieves a mIoU of 38.8 and 42.0 R@1 on ActivityNet-VRL and 34.6 mIoU and R@1 37.2 on SportsMoments. Although this is comparable to previous fully supervised methods, it still lags significantly behind our MATR. This further goes to show the importance of temporal dynamics captured by video queries as compared to static image queries.

Advantage of Dual Sequence Alignment: Table 2 shows the impact of different alignment strategies in MATR on both datasets. This study is done without any pre-training. We evaluate the effect of incorporating alignment at the pre-fusion and post-fusion stages. Our results indicate that the combined use of both pre-fusion and post-fusion alignment achieves the best performance, suggesting that alignment is important to ensure both before and after the encoder. Models with alignment either before (pre) or after (post) show moderate gains over neither. Removing post-fusion alignment hurts more than removing pre-fusion alignment, suggesting that post-fusion alignment is more important. Omitting alignment entirely results in the lowest scores. This highlights the importance of dual alignment for effectively capturing relevant temporal information for video query-based moment localization.

Effect of Pre-training: We analyze the impact of pre-training and data augmentation on model performance in

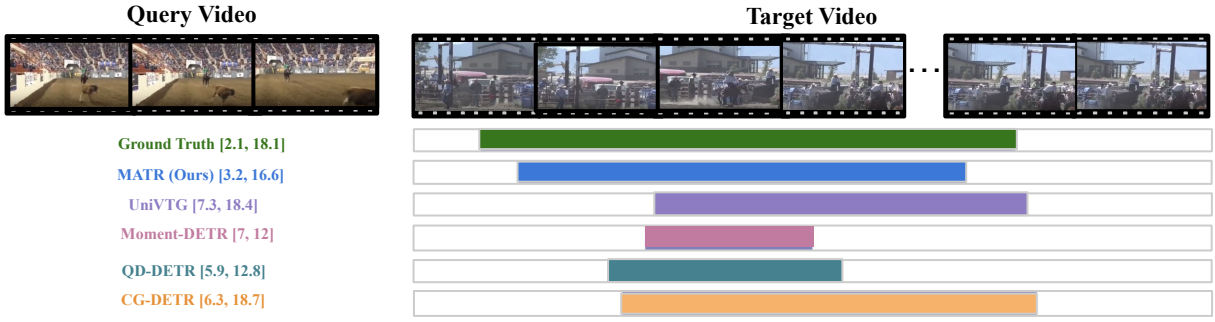

Figure 4. Visualization of *Vid2VidMR* on a sample from ActivityNet-VRL for *calf roping* action. Our proposed MATR model shows improved generalization capabilities over the best-performing baseline methods (variant (e), i.e., finetuned+V). Start and end times for the ground truth and predicted moments are shown in the brackets. Refer to supplementary material for the videos used in this figure.

	ActivityNet-VRL		SportsMoments	
	mIoU	R@1	mIoU	R@1
Pre-fusion alignment (A^{pre})	34.1	32.6	31.6	29.8
Post-fusion alignment (A^{post})	38.3	35.2	36.4	33.5
Prediction-heads	**56.8**	**57.1**	**59.2**	**56.5**

Table 4. Advantage of predicting using prediction heads on decoder. We observe that directly predicting from fusion matrices is inferior as it lacks fine-grained refinement.

	ActivityNet-VRL		SportsMoments	
Alignment	mIoU	R@1	mIoU	R@1
None	50.4	51.8	52.1	48.6
TCC	52.4	52.7	53.9	52.8
Drop-DTW	54.2	54.7	55.6	54.7
soft-DTW	**56.8**	**57.1**	**59.2**	**56.5**

Table 5. Ablation on alignment methods.

Table 3. On ActivityNet-VRL, without pre-training, the model achieves 53.2 mIoU and 54.8 R@1. Adding pre-training without augmentation yields 54.1 mIoU and 55.6 R@1. The best results are obtained by using pre-training with augmentation, leading to a significant boost to 56.8 mIoU and 57.1 R@1. We got similar observations for the SportsMoments dataset.

Predictions from Heads vs Alignment Matrices: The results in Table 4 show that prediction from heads outperforms the prediction from both pre-fusion alignment (A^{pre}) and post-fusion alignment (A^{post}) on both datasets. Notably, prediction heads improve mIoU and R@1 by a large margin, demonstrating their effectiveness in accurately retrieving temporal moments. Post-fusion alignment provides better results than pre-fusion alignment. However, the gains are modest compared to the substantial boost from prediction heads. This justifies the need for the decoder.

Ablation on Alignment Losses: MATR can incorporate different alignment methods. Table 5 compares the performance of different alignment methods on both ActivityNet-VRL and SportsMoments datasets. Without any alignment ("None"), the model achieves R@1 scores of 51.8% on ActivityNet-VRL and 48.6% on SportsMoments. TCC alignment yields an improvement of 0.9% and 4.2% on ActivityNet-VRL and SportsMoments, respectively. Drop-DTW is better than TCC. soft-DTW alignment consistently demonstrates superior performance across both datasets, achieving the highest gains of 5.3% on ActivityNet-VRL and 7.9% on SportsMoments.

Qualitative Results: We present moment retrieval results on a sample from ActivityNet-VRL in Fig. 4, where we compare our proposed approach to variant (e) (refer Sec. 5.1 for details) of the four best-performing baselines. The video query is a 5-second clip depicting *calf roping*[4]. The target video, with a duration of 24 seconds, showcases this event occurring between 2.1 and 18.1 seconds. Our proposed MATR model exhibits the highest overlap with the ground truth for the given query, accurately identifying the correct temporal boundaries. Among the baseline methods, CG-DETR and UniVTG achieve strong overlap, however, are still outperformed by our approach.

6. Conclusion

We introduced MATR, a robust approach for *Vid2VidMR* that combines abstract representation from encoder with fine-grained features from decoder conditioned on aligned query features. It captures semantic and temporal cues for precise moment localization. Our self-supervised pre-training enhances initialization and boosts performance. Extensive experiments on ActivityNet-VRL and our new SportsMoments dataset show that MATR outperforms strong baselines. Future directions of this work include exploring multimodal queries and developing scalable architectures to enhance broader applicability.

Acknowledgment: This work is supported by the Microsoft Academic Partnership Grant (MAPG) 2023. Yogesh Kumar is supported by a UGC fellowship, Govt. of India.

[4] A rodeo event where a rider on horseback attempts to catch and tie a calf within a timed competition.

References

[1] U. Agarwal, Y. Kumar, A. Shahid, P. Gatti, M. Gupta, and A. Mishra. CHAPVIDMR: Chapter-based video moment retrieval using natural language queries. In *ICVGIP*, 2024. 2

[2] D. M. Argaw, J.-Y. Lee, M. Woodson, I.-S. Kweon, and F. C. Heilbron. Long-range multimodal pretraining for movie understanding. *ICCV*, 2023. 2

[3] I. Bello, B. Zoph, Q. Le, A. Vaswani, and J. Shlens. Attention augmented convolutional networks. In *ICCV*, 2019. 3

[4] S. Buch, V. Escorcia, C. Shen, B. Ghanem, and J. C. Niebles. SST: single-stream temporal action proposals. In *CVPR*, 2017. 6, 7

[5] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 3

[6] J. Carreira and A. Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, 2017. 2, 6

[7] L. Chen, C. Lu, S. Tang, J. Xiao, D. Zhang, C. Tan, and X. Li. Rethinking the bottom-up framework for query-based video localization. In *AAAI*, 2020. 2, 6, 7

[8] Z. Cheng, S. Leng, H. Zhang, Y. Xin, X. Li, G. Chen, Y. Zhu, W. Zhang, Z. Luo, D. Zhao, and L. Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 6, 7

[9] M. Cuturi and M. Blondel. Soft-DTW: a differentiable loss function for time-series. In *ICML*, 2017. 2, 4

[10] P. Dogan, B. Li, L. Sigal, and M. Gross. A neural multi-sequence alignment technique (neumatch). In *CVPR*, 2018. 2

[11] N. Dvornik, I. Hadji, K. G. Derpanis, A. Garg, and A. D. Jepson. Drop-DTW: Aligning common signal between sequences while dropping outliers. In *NeurIPS*, 2021. 2, 4

[12] D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman. Temporal cycle-consistency learning. In *CVPR*, 2019. 2, 4

[13] C. Feichtenhofer, H. Fan, J. Malik, and K. He. Slowfast networks for video recognition. In *ICCV*, 2019. 2

[14] Y. Feng, L. Ma, W. Liu, T. Zhang, and J. Luo. Video re-localization. In *ECCV*, 2018. 2, 5, 6, 7

[15] J. Gao, C. Sun, Z. Yang, and R. Nevatia. TALL: temporal activity localization via language query. In *ICCV*, 2017. 2

[16] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics*, 2010. 6

[17] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015. 5

[18] L. A. Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. C. Russell. Localizing moments in video with natural language. In *ICCV*, 2017. 2

[19] S. Huo, Y. Zhou, R. Wang, W. Xiang, and S. Kung. Semantic relevance learning for video-query based video moment retrieval. *IEEE Trans. Multim.*, 25:9290–9301, 2023. 2, 6, 7

[20] M. Jung, Y. Jang, S. Choi, J. Kim, J.-H. Kim, and B.-T. Zhang. Background-aware moment detection for video moment retrieval. In *WACV*, 2025. 2

[21] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 5

[22] Y. Kumar, S. Mallick, A. Mishra, S. Rasipuram, A. Maitra, and R. Ramnani. Qdetrv: Query-guided detr for one-shot object localization in videos. In *AAAI*, 2024. 5

[23] H.-Y. Lee, J.-B. Huang, M. Singh, and M.-H. Yang. Unsupervised representation learning by sorting sequences. In *ICCV*, 2017. 5

[24] J. Lei, L. Yu, T. L. Berg, and M. Bansal. TVR: A large-scale dataset for video-subtitle moment retrieval. In *ECCV*, 2020. 2

[25] J. Lei, T. L. Berg, and M. Bansal. Detecting moments and highlights in videos via natural language queries. In *NeurIPS*, 2021. 1, 2, 6, 7

[26] B. Lin, B. Zhu, Y. Ye, M. Ning, P. Jin, and L. Yuan. Video-llava: Learning united visual representation by alignment before projection. In *EMNLP*, 2024. 6, 7

[27] J. Lin, C. Gan, and S. Han. TSM: temporal shift module for efficient video understanding. In *ICCV*, 2019. 2

[28] K. Q. Lin, P. Zhang, J. Chen, S. Pramanick, D. Gao, A. J. Wang, R. Yan, and M. Z. Shou. Univtg: Towards unified video-language temporal grounding. In *ICCV*, 2023. 1, 2, 6, 7

[29] Y. Liu, S. Li, Y. Wu, C. W. Chen, Y. Shan, and X. Qie. UMT: unified multi-modal transformers for joint video moment retrieval and highlight detection. In *CVPR*, 2022. 2

[30] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *ICLR*, 2017. 6

[31] N. C. Mithun, S. Paul, and A. K. Roy-Chowdhury. Weakly supervised video moment retrieval from text queries. In *CVPR*, 2019. 2

[32] W. Moon, S. Hyun, S. B. Lee, and J. Heo. Correlation-guided query-dependency calibration in video representation learning for temporal grounding. *CoRR*, abs/2311.08835, 2023. 1, 2, 6, 7

[33] W. Moon, S. Hyun, S. Park, D. Park, and J. Heo. Query - dependent video representation for moment retrieval and highlight detection. In *CVPR*, 2023. 1, 2, 6, 7

[34] J. Mun, M. Cho, , and B. Han. Local-Global Video-Text Interactions for Temporal Grounding. In *CVPR*, 2020. 2

[35] F. Padua, R. Carceroni, G. Santos, and K. Kutulakos. Linear sequence-to-sequence alignment. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32:304–320, 2008. 2

[36] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran. Image transformer. In *ICML*, 2018. 3

[37] A. Piergiovanni, A. Angelova, and M. S. Ryoo. Evolving losses for unsupervised video representation learning. In *CVPR*, 2020. 5

[38] R. Qian, T. Meng, B. Gong, M.-H. Yang, H. Wang, S. Belongie, and Y. Cui. Spatiotemporal contrastive video representation learning. In *CVPR*, 2021. 5

[39] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3

[40] Rao, Gritai, and Shah. View-invariant alignment and matching of video sequences. In *ICCV*, 2003. 2

[41] S. Ren, L. Yao, S. Li, X. Sun, and L. Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *CVPR*, 2024. 6, 7

[42] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, 2019. 4

[43] Z. Shou, D. Wang, and S.-F. Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *CVPR*, 2016. 2

[44] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NuerIPS*, 2014. 2

[45] N. Srivastava, E. Mansimov, and R. Salakhudinov. Unsupervised learning of video representations using lstms. In *ICML*, 2015. 5

[46] X. Sun, H. Wang, and B. He. Maban: Multi-agent boundary-aware network for natural language moment retrieval. *IEEE Trans. on Image Processing*, 30:5589–5599, 2021. 6, 7

[47] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 2, 6

[48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3

[49] G. Wang, Y. Zhou, C. Luo, W. Xie, W. Zeng, and Z. Xiong. Unsupervised visual representation learning by tracking patches in video. In *CVPR*, 2021. 5

[50] J. Wang, Y. Gao, K. Li, J. Hu, X. Jiang, X. Guo, R. Ji, and X. Sun. Enhancing unsupervised video representation learning by decoupling the scene and the motion. In *AAAI*, 2021. 5

[51] R. Wang and Y. Zhou. A feature pair fusion and hierarchical learning framework for video re-localization. In *ICIP*, 2020. 2, 6, 7

[52] X. Wang, L. Zhu, and Y. Yang. T2vlad: global-local sequence alignment for text-video retrieval. In *CVPR*, 2021. 2

[53] Y. Xiao, Z. Luo, Y. Liu, Y. Ma, H. Bian, Y. Ji, Y. Yang, and X. Li. Bridging the gap: A unified video comprehension framework for moment retrieval and highlight detection. In *CVPR*, 2024. 2

[54] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, 2018. 2

[55] S. Yan, X. Xiong, A. Nagrani, A. Arnab, Z. Wang, W. Ge, D. Ross, and C. Schmid. Unloc: A unified framework for video localization tasks. In *ICCV*, 2023. 2

[56] C. Yang, Y. Xu, J. Shi, B. Dai, and B. Zhou. Temporal pyramid network for action recognition. In *CVPR*, 2020. 2

[57] J. Yang, P. Wei, H. Li, and Z. Ren. Task-driven exploration: Decoupling and inter-task feedback for joint moment retrieval and highlight detection. In *CVPR*, 2024. 1, 2

[58] Q. Ye, H. Xu, J. Ye, M. Yan, A. Hu, H. Liu, Q. Qian, J. Zhang, F. Huang, and J. Zhou. mplug-owi2: Revolutionizing multi-modal large language model with modality collaboration. In *CVPR*, 2024. 6

[59] R. Zeng, W. Huang, M. Tan, Y. Rong, P. Zhao, J. Huang, and C. Gan. Graph convolutional networks for temporal action localization. In *CVPR*, 2019. 2

[60] C. Zhang, A. Gupta, and A. Zisserman. Temporal query networks for fine-grained video understanding. In *CVPR*, 2021. 2

[61] H. Zhang, A. Sun, W. Jing, L. Zhen, J. T. Zhou, and R. S. M. Goh. Natural language video localization: A revisit in span-based question answering framework. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022. 6, 7

[62] H. Zhang, X. Li, and L. Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *EMNLP*, 2023. 6, 7

[63] S. Zhang, H. Peng, J. Fu, and J. Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *AAAI*, 2020. 6, 7

[64] Z. Zhu, W. Tang, L. Wang, N. Zheng, and G. Hua. Enriching local and global contexts for temporal action localization. In *ICCV*, 2021. 2