Partial Annotation-based Video Moment Retrieval via Iterative Learning

Wei Ji
National University of Singapore
Singapore, Singapore
weiji0523@gmail.com

Renjie Liang
National University of Singapore
Singapore, Singapore
t0924327@u.nus.edu

Lizi Liao
Singapore Management University
Singapore, Singapore
lzliao@smu.edu.sg

Hao Fei*
National University of Singapore
Singapore, Singapore
haofei37@nus.edu.sg

Fuli Feng
University of Science and Technology of China
Hefei, China
fulifeng93@gmail.com

ABSTRACT

Given a descriptive language query, Video Moment Retrieval (VMR) aims to seek the corresponding semantic-consistent moment clip in the video, which is represented as a pair of the start and end timestamps. Although current methods have achieved satisfying performance, training these models heavily relies on the fully-annotated VMR datasets. Nonetheless, precise video temporal annotations are extremely labor-intensive and ambiguous due to the diverse preferences of different annotators.

Although there are several works trying to explore weakly supervised VMR tasks with scattered annotated frames as labels, there is still much room to improve in terms of accuracy. Therefore, we design a new setting of VMR where users can easily point to small segments of non-controversy video moments and our proposed method can automatically fill in the remaining parts based on the video and query semantics. To support this, we propose a new framework named Video Moment Retrieval via Iterative Learning (VMRIL). It treats the partial temporal region as the seed, then expands the pseudo label by iterative training. In order to restrict the expansion with reasonable boundaries, we utilize a pretrained video action localization model to provide coarse guidance of potential video segments. Compared with other VMR methods, our VMRIL achieves a trade-off between satisfying performance and annotation efficiency. Experimental results show that our proposed method can achieve the SOTA performance in the weakly supervised VMR setting, and are even comparable with some fully-supervised VMR methods but with much less annotation cost.

CCS CONCEPTS

• **Computing methodologies → Visual content-based indexing and retrieval**.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10.
https://doi.org/10.1145/3581783.3612088

KEYWORDS

video moment retrieval, labor-intensive, pseudo label, coarse guidance

ACM Reference Format:
Wei Ji, Renjie Liang, Lizi Liao, Hao Fei, and Fuli Feng. 2023. Partial Annotation-based Video Moment Retrieval via Iterative Learning. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3612088

1 INTRODUCTION

Video Moment Retrieval (VMR) (a.k.a., Natural Language Video Localization, Video Sentence Grounding) aims to localize the temporal moment from an untrimmed video corresponding to a descriptive query, which is represented as a pair of the start and end timestamps. As a multi-modal task that bridges computer vision and natural language processing, VMR is beneficial to a series of downstream tasks, such as video question answering [18, 19, 44], video relationship detection [36, 37], and video dialog [4, 34], *etc.* Thus the task has been fundamental in the topic of video understanding.

While VMR performance has achieved huge gains in recent years, most of the existing methods are limited to the fully supervised setting which rely on large-scale annotation corpus, such as Charades-STA [13], ActivityNet [38], *etc.* Manually annotating data, especially for the video modality, can be time-consuming and labor-intensive, which greatly hinders the application of VMR in real-world scenarios. Besides, as cast in [32], different annotators may have different preferences when annotating a video moment corresponding to the same query (*e.g.*, in Figure 1) , it might be hard for annotators to agree upon *whether it starts at the moment when the guy first holds the saxophone or blows into the saxophone.* Such disagreements in label supervision will inevitably mislead and harm the model training.

One direct solution to ease the above conundrums is to alleviate the reliance on annotation supervision. Some previous works have explored weakly-supervised VMR with only <*query, video*> pairs [29, 30] without fine-grained temporal boundary labels, as illustrated in Figure 1 (b). Due to the lack of location information in training, nevertheless, these practices can lead to much poorer performance compared with the fully-supervised methods.

*Corresponding Author

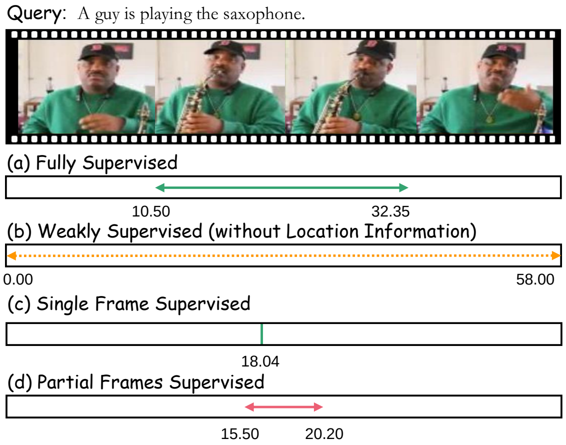

Query: A guy is playing the saxophone.
(a) Fully Supervised
10.50
32.35
(b) Weakly Supervised (without Location Information)
0.00
58.00
(c) Single Frame Supervised
18.04
(d) Partial Frames Supervised
15.50
20.20

Figure 1: Illustration of Video Moment Retrieval tasks in different settings. The green arrow represents the supervision. (a) is fully supervised setting with precise temporal labels, such as timestamp of (10.50, 32.35); (b) is weakly supervised setting without any location information; (c) is single frame annotation as supervision, such as timestamp of (18.04). Compared with (a), (b), and (c), our proposed weak-supervised setting (d) provides partial location signals that are without controversy.

Recently, [5] propose a different weakly-supervised VMR setting, where one single frame within fully-supervised ground truth is taken as "glance" annotation (as shown in Figure 1 (c)), so as to effectively reduce the burden of annotation. Despite achieving better task performance, this method suffers from the problem that a single frame of annotation is insufficient when facing queries with multiple semantics. Also, using Gaussian distribution with a single frame as a peak to simulate the pseudo label for video clips is heuristic, and thus limits the label area and confines the expanding possibility.

In view of the labor-saving potential as well as current limitations of existing weakly-supervised VMR methods, in this work, we explore a more effective and practical setting of weakly-supervised VMR. The core idea is characterized as *seed annotation*, i.e., encouraging the system to learn to gradually expand to more high-confidence regions. This naturally corresponds to the fact that different annotators can select a small segment of the non-controversy part (denoted as "seed", as shown in Figure 1 (d)) and the model learns to fill in the remaining parts. To realize this, two main challenges need to be properly addressed: (1) How to properly solicit evidence from the query and video content for area expansion, given the seed as supervision; (2) How to determine the boundaries for the iterative expansion process based on query and video content semantics.

Hence, given the partial temporal regions as weak supervision, we consider a new pipeline named Video Moment Retrieval via Iterative Learning (VMRIL) to train a reliable VMR model. Our proposed pipeline can be adaptive to any other fully-supervised VMR models. First, we train a VMR model under the supervision of the seed area, which automatically learns the semantic associations between the video segments and textual queries. Such knowledge is further utilized to gradually expand possible regions as pseudo labels for training data. In order to restrict the expansion of pseudo labels, we utilize a pretrained action localization model and filtering function to provide coarse-grained guidance of possible temporal boundaries. By iteratively training the VMR model and generating pseudo labels, it learns to incorporate various shreds of evidence to find the proper region.

Our contributions are summarized as follows:

- We propose a new setting of weakly-supervised VMR task with partial labels, release reorganized datasets of three public datasets, and introduce a corresponding solution named novel iterative learning based pipeline (VMRIL), which can be adaptive to any other fully-supervised VMR baselines.
- To restrict the expansion with reasonable boundaries, we utilize a pretrained video action model and filtering function to provide coarse guidance of potential video segments. We also propose a multi-label training strategy to make the model more robust.
- Experimental results show that our proposed methods achieve the SOTA performance in weakly supervised VMR, and are even comparable with some fully-supervised methods.

2 RELATED WORK

Fully supervised Video Moment Retrieval. Video Moment Retrieval (VMR) is defined as retrieving video segments with consistent semantics of query [1, 22–26, 45], which is also relevant to a series of visual retrieval tasks [6–10]. The fully supervised VMR methods can mainly be classified into three categories: Proposal-based method, Proposal-free method, and Reinforcement Learning-based method. In early works, some proposal-based methods [15, 27] treat this task as a ranking problem and follow the propose-and-rank pipeline. These methods first generate proposals in various lengths by sliding window [13] or Segment Proposal Network (SPN) [46], then they calculate the multi-modal semantic matching to find the best matching proposal for the query. The drawbacks of proposal-based methods lie in densely sampling video moment proposals to achieve good performance, which leads to large computation costs.

To overcome the above-mentioned drawbacks, Yuan et al.[51] propose a proposal-free method, which utilizes a Bi-LSTM to encode visual and sentence features, and a co-attention interaction module to fuse multi-modal features. The model treats the video as a whole and directly predicts the temporal coordinates according to the sentence queries. Lu et al.[28] propose a dense bottom-up framework, which treats all frames corresponding to the language query as foreground, and then regresses the unique distances of each frame in the foreground to bi-directional ground-truth boundaries, finally fuses appropriate temporal candidates as the final result. To mine the relationship of sentence semantics with diverse video contents, Yuan et al.[50] also propose a semantic conditioned dynamic modulation, which leverages sentence semantic information to modulate the temporal convolution processes in a hierarchical temporal convolutional network and establishes a precise matching relationship between sentence and video. Recently, Zhang et al. [54] propose a 2D temporal map to model the temporal relations of different moments with variant lengths, in which the two dimensions indicate

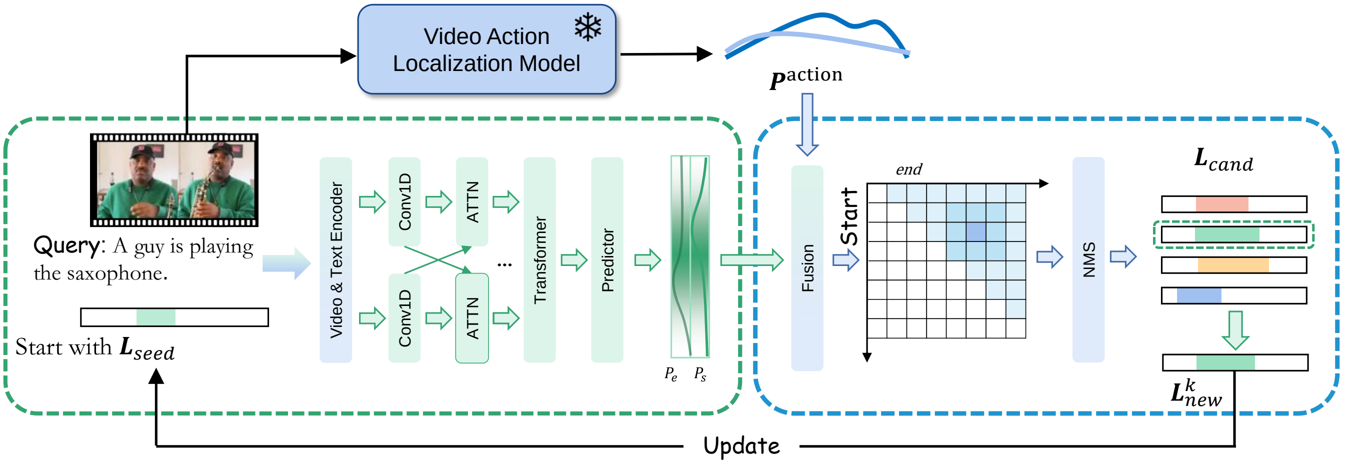


Figure 2: The proposed VMRIL pipeline. Our VMR model (in the green box) is first trained with the seed area as supervision. It then generates pseudo labels via iterative expansion. In each iteration, the pseudo labels are adjusted with the temporal guidance of the pretrained action location model. The VMR model will be updated with the pseudo labels.

the start and end timestamps, respectively. Apart from these top-down and bottom-up methods, there are also works [14, 41] which adopt reinforcement learning to make decisions on the action space of candidate segments, such as the start or end boundaries move left or right, corresponding to the language query matching result. The performance of all these methods mentioned above heavily relies on well-annotated datasets. Different from these, we explore weakly supervised Video Moment Retrieval tasks with trivial annotation costs.

Weakly Supervised Video Moment Retrieval. Without the annotation of precise temporal labels, weakly supervised Video Moment Retrieval methods can be mainly grouped into two categories: multi-instance learning and reconstruction-based methods. For the multi-instance learning methods, TGA [30] first deals with the Video Moment Retrieval task by treating the video and corresponding query as positive pairs, while video with other queries and query with other videos as negative pairs. For the reconstruction-based methods, Duan et al.[12] treat the moment localization and event captioning as dual tasks, and constrain the reconstructed caption and raw query with cycle consistency.

Different from these weakly supervised VMR works that only use <*query, video*> pairs as supervision, there are also some works that try to explore the higher performance of weakly supervised VMR with few localization annotations. For example, Cui et al. [5] propose to use one random frame with the temporal region of fully supervised labels as weak supervision and then propose the ViGA method based on contrastive learning. In this paper, we further explore the weakly supervised VMR task with few temporal annotations in an iterative learning framework, which yields better performance.

3 APPROACH

In this section, we first define the problem of weakly supervised VMR with seed annotation. Then we introduce each component of our VMRIL model in Section 3.3 and 3.4. The whole architecture of VMRIL is illustrated in Figure 2. Finally, the inference process is introduced in Section 3.5.

3.1 Dataset Collection

The motivation of our work is annotating small segments of non-controversy video moments for the weakly-supervised Video Moment Retrieval tasks. Compared with annotating a single frame for each video, our setting provides more information but at minor cost. The annotators only need to label the temporal region with the highest confidence, rather than locating the precise temporal boundaries, which is a quick annotating method for a new dataset. We provide a set of new partial annotations and provide ablation studies of different labeling methods in Section 4.5.1.

3.2 Problem Formulation

Given an untrimmed video $V = \{f_t\}_{t=1}^{T}$ and the language query $Q = \{q_j\}_{j=1}^{m}$, where T and m are the numbers of frames and words, respectively, our goal is to predict the start and end timestamp (τ^s, τ^e) in the video corresponding to query Q. In a fully supervised setting, VMR models are trained with the ground truth of $L_{GT} = (\tau^s, \tau^e)$. In this paper, we propose a new setting of VMR with partial temporal regions (ϕ^s, ϕ^e) which are randomly selected from the segment (τ^s, τ^e), where $\tau^s \leq \phi^s \leq \phi^e \leq \tau^e$. And the IoU of ($\phi^s$, ϕ^e) compared with (τ^s, τ^e) is defined as λ. We refer the partial temporal regions $L_{seed} = (\phi^s, \phi^e)$ as seed area. In each iteration, we calculate the IoU of the pseudo label compared with (τ^s, τ^e) as ω.

3.3 Multi-modal Feature Encoder Module

For each video $\boldsymbol{V}$, we extract its visual features $\boldsymbol{V} = \{\boldsymbol{v}_i\}_{i=1}^{n} \in \mathbb{R}^{n \times d_v}$ with a pretrained 3D ConvNet[3], where n is the length of extracted features. Each feature $\boldsymbol{v}_i$ here is a video feature vector. For each query Q, we initialize the word features $\{\boldsymbol{Q} = \boldsymbol{q}_j\}_{j=1}^{m} \in \mathbb{R}^{m \times d_q}$ by using the popular GloVe embeddings [33].

We first project $\boldsymbol{V}$ and $\boldsymbol{Q}$ into the same dimension d using projection matrices, then we feed them into the `VisualEncoder` and `QueryEncoder` respectively:

$$\widetilde{\boldsymbol{V}} = \texttt{VisualEncoder}(\boldsymbol{V}\boldsymbol{W}^{v}), \widetilde{\boldsymbol{Q}} \quad = \texttt{QueryEncoder}(\boldsymbol{Q}\boldsymbol{W}^{q}) \qquad (1)$$

where $\boldsymbol{W}^{v} \in \mathbb{R}^{d_v \times d}$ and $\boldsymbol{W}^{q} \in \mathbb{R}^{d_q \times d}$ are projection matrices to keep the dimension consistent between two modalities. Inspired by QANet [49], the `VisualEncoder` and `QueryEncoder` are composed of four convolution layers and a multi-head attention layer to capture in-depth semantics in the corresponding modality.

With the encoded visual and query features $\widetilde{V}$ and $\widetilde{Q}$, we calculate the similarity between the two modal features and fuse the multi-modal features by a multi-modal attention mechanism. Similar to [28], we first calculate the similarity scores, $\mathcal{S} \in \mathbb{R}^{n\times m}$, between each visual feature and query feature. Then the attention weights of visual-to-query ($\mathcal{A}$) and query-to-visual ($\mathcal{B}$) are computed as:

$$\begin{aligned} \mathcal{A} &= \mathcal{S}_r \cdot \widetilde{Q} \in \mathbb{R}^{n\times d}, \\ \mathcal{B} &= \mathcal{S}_r \cdot \mathcal{S}_c^T \cdot \widetilde{V} \in \mathbb{R}^{n\times d}, \end{aligned} \tag{2}$$

where $\mathcal{S}_r$ and $\mathcal{S}_c$ are the row-wise and column-wise normalization of $\mathcal{S}$ by softmax operation, respectively. Finally, the output of visual-query attention is written as:

$$V^q = \mathtt{FFN}\big([\widetilde{V}; \mathcal{A}; \widetilde{V} \odot \mathcal{A}; \widetilde{V} \odot \mathcal{B}]\big), \tag{3}$$

where $V^q \in \mathbb{R}^{n\times d}$; `FFN` is a single feed-forward layer; ⊙ denotes element-wise multiplication. V^q is the fused multi-modal semantic features with visual and query attention.

Then we follow [52] and calculate P_s and P_e, which represent $logit_{start}$ and $logit_{end}$ in Figure 2. Hence, the whole network of VMR model can be defined as:

$$(P_s, P_e) = F(V, Q; \theta), \tag{4}$$

where θ refers to the network parameters of F.

3.4 Iterative Learning with Pseudo Label

In the initial iteration, the start and end boundary curve (P_s, P_e) are supervised with seed area $L_{seed} = (\phi^s, \phi^e)$. After training with a cross entropy loss between L_{seed} and (P_s, P_e), the parameters of VMR model θ_0 are updated. In the following iterations, the labels of training data will be updated with training data as input of the VMR model.

Since the seed area is partial annotation, we need to expand the seed area by iterative learning and constrain the temporal boundary to avoid over-expansion. Then, we borrow the knowledge of pre-trained action localization model, BMN [20], to provide guidance on pseudo label expansion. Given the video as input, we extract Boundary Probability Curve P_s^{action} andP_e^{action}, which are the outputs of Temporal Evaluation Module in BMN model. P_s^{action} andP_e^{action} contain the temporal localization information of pretrained action labels.

Then, in order to fuse information of BMN model and output of VMR model (P_s, P_e), we define the following operations to generate the pseudo label based on (P_s, P_e) and Action Boundary Probability $(P_s^{action}, P_e^{action})$:

$$\begin{aligned} S &= matmul(P_s \times P_s^{action}, P_e \times P_e^{action}), \\ L_{cand} &= NMS(S) \end{aligned} \tag{5}$$

where (P_s, P_e) represents the start/end possibility in the video sequence, *matmul* means matrix multiply. Here, $S \in \mathbb{R}^{n\times n}$ is the scores map, *NMS* means the Non-Maximum Suppression, which is widely used in the object detection tasks, as shown in Figure 3.

Actually, we want candidate labels to be both diverse and accurate. NMS is an ideal method to generate candidate labels. After fusing the proposals from (P_s, P_e) and $(P_s^{action}, P_e^{action})$, we select the most confident one L_{cand} as the candidates. Then, we select pseudo label from L_{cand} and L_{seed}:

$$L_{new}^k = argmax(soft\text{-}mIoU(L_{cand}, L_{seed}) > \alpha), \tag{6}$$

where α is the hyper-parameter. More ablation studies about hyper-parameter and selection operations can be found in Section. 4.5.2. L_k^{new} is the new pseudo label corresponding to the language query Q, where k is the number of iteration.

Then, we take a new pseudo label L_k^{new} of training data to retrain the VMR model. The loss function is cross-entropy as [54] described. After k iterations, the parameters of VMR model θ_k are updated.

3.5 Training and Inference.

Multi-label Training. In the training stage, the output of SeqPAN [52] model is the probability distributions of start/end boundaries $P_{s/e}$. The training objective is:

$$\mathcal{L}_{\text{loc}} = \frac{1}{2} \times \left[f_{\text{CE}}(P_s, Y_s) + f_{\text{CE}}(P_e, Y_e) \right] \tag{7}$$

where f_{CE} is the cross-entropy function, $Y_{s/e}$ is the one-hot labels for start/end (i^s/i^e) boundaries.

Since we only have pseudo labels to train the SeqPAN, which are not precise as ground truth, we propose soft label to replace hard label in the seed region: For the start/end frame, we use a Gaussian distribution to model the labels of surrounding frames, where the peak position of Gaussian is the seed region, σ represents the width of Gaussian curve.

Besides, to improve the generalization ability of our model, we generate multi labels by adjusting the start and end boundaries of pseudo labels with a fixed offset. More ablation studies can be found in Section 4.5.3.

The overall training loss of SeqPAN is to minimize the combined loss of $\mathcal{L}_{\text{loc}}$ and supervision to intermediate features during the training process. Considering the multi-label regularization, the rectified loss is:

$$\begin{aligned} std &= var(P_s) + var(P_e), \\ Loss &= \frac{\mathcal{L}_{\text{loc}}}{std} + std, \end{aligned} \tag{8}$$

where $var(P_s)$ and $var(P_e)$ are the variance of P_s and P_e, we use $Loss$ to train the whole VMRIL model.

Inference. When testing, with the trained model $F(V, Q; \theta_k)$, the predicted start and end boundaries of the given video-query

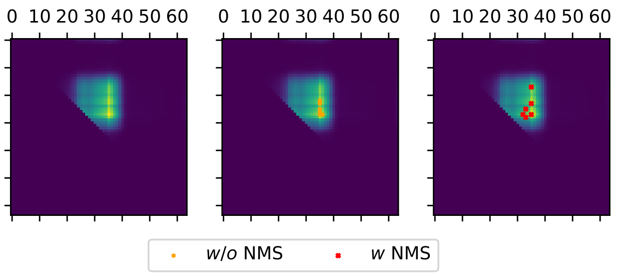


Figure 3: Illustration of NMS on temporal map. Samples with NMS are sparse with diversity.

pair (V, Q) are generated by maximizing the joint probability as:

$$(\hat{i}^s, \hat{i}^e) = \arg\max_{\hat{a}^s, \hat{a}^e} P_s(\hat{a}^s) \times P_e(\hat{a}^e) \\ \text{s.t.: } 0 \le \hat{i}^s \le \hat{i}^e \le N-1 \tag{9}$$

where $\hat{i}^s$ and $\hat{i}^e$ are the best start and end boundaries of the predicted moment for the given video-query pair. And the predicted start/end time is computed by $\hat{t}^{s(e)} = \hat{i}^{s(e)}/(N-1) \times \mathcal{T}$, where $\mathcal{T}$ is the duration of the given video.

4 EXPERIMENTS

4.1 Datasets

To evaluate the performance of our proposed VMRIL, we conduct experiments on three challenging Video Moment Retrieval datasets, all the queries in these datasets are in English:

Charades-STA [13] is composed of daily indoor activities videos, which is based on Charades dataset [38]. This dataset contains 6672 videos, 16,128 annotations, and 11,767 moments. The average length of each video is 30 seconds. $12,408$ and $3,720$ moment annotations are labeled for training and testing, respectively; **ActivityNet Caption** [2] is originally constructed for dense video captioning, which contains about 20k YouTube videos with an average length of 120 seconds. As a dual task of dense video captioning, Video Moment Retrieval utilize the sentence description as a query and outputs the temporal boundary of each sentence description. **TACoS** [35] is collected from MPII Cooking dataset [35], which has 127 videos with an average length of 286.59 seconds. TACoS has 18,818 query-moment pairs, which are all about cooking scenes. We follow the same splits in [13], where $10,146$, $4,589$, and $4,083$ annotations are used for training, validation, and testing, respectively.

It is worth noting that some methods make minor changes to the dataset when evaluating the experimental performance. For example, CMIN [55] uses val_1 as the validation set and val_2 as the testing set in the ActivityNet Captions dataset, while other methods [54] combine the val_1 and val_2 together as the testing set. And for the TACoS dataset, SeqPAN [54] utilize a modified TACoS dataset for evaluation. To make a fair comparison, we follow the setting of dataset splitting to report in their original papers when evaluating the performance of our method.

4.2 Evaluation Metrics

Following existing video grounding works, we evaluate the performance on two main metrics: **mIoU:** "mIoU" is the average predicted Intersection over Union in all testing samples. The mIoU metric is particularly challenging for short video moments; **Recall:** We adopt "R@n, IoU $= \mu$" as the evaluation metrics, following [13]. The "R@n, IoU $= \mu$" represents the percentage of language queries having at least one result whose IoU between top-n predictions with ground-truth is larger than μ. In our experiments, we reported the results of $n = 1$ and $\mu \in \{0.3, 0.5, 0.7\}$.

4.3 Implementation Details

For language query Q, we use the 300-D GloVe [33] vectors to initialize each lowercase word, and these word embeddings are fixed during training. For video V, we downsample frames and extracted RGB visual features using the 3D ConvNet which was pre-trained on the Kinetics dataset. We set the dimension of all the hidden layers in the model as 128, the kernel size of the convolutional layer as 7, and the head size of multi-head attention as 8. For all datasets, models were trained for 50 epochs. The batch size was set to 64. Dropout and an early stopping strategies were adopted to prevent overfitting. The whole framework was trained by Adam optimizer with an initial learning rate 0.0002. We use the weakly datasets with $\lambda = 0.3$ in the following experiments. The α is set at 0.05, 0.20, 0.20 in Charades-STA, ActivityNet and TACoS respectively. More ablation studies can be found in Section 4.5. All experiments are conducted on a NVIDIA RTX A5000 GPU with 24GB memory. The parameters of VMRIL keeps the same as SeqPAN and VSLNet.

4.4 Comparison with State-of-the-Arts

4.4.1 Experimental Settings. We compare our proposed VMRIL with state-of-the-art Video Moment Retrieval methods on three public datasets. These methods can be grouped into two categories according to the viewpoints of fully-supervised, weakly-supervised approaches:

1) For the fully-supervised VMR models: **CTRL** [13] produces proposals in various length via sliding window; **QSPN** [47] is a representative proposal generation-based method; **2D-TAN** [54] models proposals with 2D Temporal Map; **LGI** [31] model the visual and textual feature via effective local-global interaction in hierarchical levels; **VSLNet** [53] is a span-based method which aims to predict the start/end probability of each frame; **SeqPAN** [52] proposes a parallel attention network with sequence matching to deal with multi-modal representation and target moment boundary prediction; **EAMAT** [48] designs an entity-aware and motion-aware Transformers to detect the actions in the video sequences globally and refines the temporal boundaries locally; **BANet-APR** [11] proposes boundary-aware feature aggregation module to fuse boundary features and propose a proposal-level contrastive learning method to learn query-related content features; **EMB** [16] establishes an explicit association between the content information of each segment and the boundary information of each frame, and synergistically complements them via a novel guided attention mechanism.

2) In the weakly-supervised VMR models: **TGA** [30], **BAR** [43], **LoGAN** [40], **CRM** [17], as well as **VLANet**[29] are all methods based on multi-instance learning, which regards the input video as bag of instances with bag-level annotations; **SCN** [21] and **MARN** [39] are reconstruction-based methods, which select a certain number proposals as input to reconstruct masked query, and compute rewards based on reconstruction loss. **RTBPN** [56] designs the shared two-branch proposal module to generate positive proposals from the enhanced stream and plausible negative proposals from the suppressed one; **WSTAN** [42] learns cross-modal semantic alignment by exploiting temporal adjacent network in a multiple-instance learning (MIL) paradigm, with a whole description paragraph as input; **CNM** [57] introduces intra-video contrastive negative sample mining to deal with weakly supervised Video Moment Retrieval task; **CPL** [58] proposes a controllable Easy to Hard Negative sample mining strategy to collect negative proposals within the video and ease the model optimization;

Table 1: Performance comparison with the state-of-the-art methods under different supervision settings. VSLNet (baseline) and SeqPAN (baseline) represents retraining two baseline models with seed supervision and without iterations. VMRIL (VSLNet) and VMRIL (SeqPAN) represents VMRIL with different baselines.

Supervision	Method	Charades-STA				ActivityNet Captions				TACoS			
		R1@0.3	R1@0.5	R1@0.7	mIoU	R1@0.3	R1@0.5	R1@0.7	mIoU	R1@0.3	R1@0.5	R1@0.7	mIoU
Full Supervision	CTRL [13]	-	23.63	8.89	-	-	-	-	-	18.32	13.3	-	-
	QSPN [47]	54.7	35.6	15.8	-	45.3	27.7	13.6	-	-	-	-	-
	2D-TAN [54]	-	39.7	23.31	-	59.45	44.51	26.54	-	37.29	25.32	-	-
	LGI [31]	72.96	59.46	35.48	51.38	58.52	41.51	23.07	41.13	-	-	-	-
	VSLNet [53]	70.46	54.19	35.22	50.02	63.16	43.22	26.16	43.19	29.61	24.27	20.03	24.11
	SeqPAN [52]	73.84	60.86	41.34	53.92	61.65	45.50	28.37	45.11	48.64	39.64	28.07	37.17
	EAMAT [48]	74.19	61.69	41.96	54.45	55.33	38.07	22.87	40.12	50.11	38.16	26.82	36.43
	BANet-APR [11]	74.05	63.09	42.12	54.15	65.11	46.8	26.70	45.87	48.24	33.74	-	-
	EMB [16]	72.50	58.33	39.25	53.09	64.13	44.81	26.07	45.59	50.46	37.82	22.54	35.49
Weak Supervision	TGA [30]	32.14	19.94	8.84	-	-	-	-	-	-	-	-	-
	SCN [21]	42.96	23.58	9.97	-	47.23	29.22	-	-	-	-	-	-
	BAR [43]	44.97	27.04	12.23	-	49.03	30.73	-	-	-	-	-	-
	RTBPN [56]	60.04	32.36	13.24	-	49.77	29.63	-	-	-	-	-	-
	VLANet [29]	45.24	31.83	14.17	-	-	-	-	-	-	-	-	-
	MARN [39]	48.55	31.94	14.81	-	47.01	29.95	-	-	-	-	-	-
	LoGAN [40]	51.67	34.68	14.54	-	-	-	-	-	-	-	-	-
	CRM [17]	53.66	34.76	16.37	-	55.26	32.19	-	-	-	-	-	-
	WSTAN[42]	43.39	29.35	12.28	-	52.45	30.01	-	-	-	-	-	-
	CNM[57]	60.04	35.15	14.95	-	55.68	33.33	-	-	-	-	-	-
	CPL [58]	66.4	49.24	22.39	43.48	55.73	31.37	12.32	36.82	-	-	-	-
Glance	ViGA [5]	71.21	45.05	20.27	44.57	59.61	35.79	16.96	40.12	19.62	8.85	3.22	15.47
Partial	VSLNet (baseline)	25.48	7.77	2.07	19.13	28.72	12.84	3.78	20.08	21.69	9.12	2.17	15.16
	SeqPAN (baseline)	39.35	9.49	1.26	23.52	43.22	23.37	11.52	32.58	27.12	11.07	3.12	17.69
	VMRIL (VSLNet)	52.12	22.88	12.9	37.13	52.66	34.1	19.44	38.50	32.67	19.85	7.22	22.09
	VMRIL (SeqPAN)	**72.19**	**55.16**	**34.19**	**50.18**	**59.35**	**41.10**	**23.94**	**42.94**	**47.66**	**35.04**	**19.05**	**33.00**

ViGA [5] is recently proposed which brings in single frame label as glance supervision.

We choose VSLNet [53] and SeqPAN [54] as baseline networks, which are known as typical proposal-free models with published source codes. For the implementation based on each baseline method, our VMRIL (VSLNet) shares the same architecture as VSLNet in Feature Encoder Module. Specifically, we implement the VSLNet with C3D feature followed the settings they reported in original paper. For the VMRIL (SeqPAN), we follow the same settings according to original papers.

4.4.2 Quantitative Results. Table 1 summarizes the experimental results on Charades-STA, TACoS, and ActivityNet Captions dataset. SeqPAN (baseline) and VSLNet (baseline) represent directly training SeqPAN and VSLNet with seed area L_{seed}. VMRIL (SeqPAN) and VMRIL (VSLNet) mean iteratively training based on the two baselines. From the results we observe that our VMRIL can effectively improve the performance of baseline networks over all metrics and benchmarks. For Charades-STA dataset, we can see that VMRIL (SeqPAN) works well in even stricter metrics. Compared with ViGA [5], VMRIL (SeqPAN) achieves a significant 5.61% absolute improvement in mIoU, which demonstrates the effectiveness of proposed model. Moreover, VMRIL benefits from iterative learning and pretrained video localization model, and thus achieves state-of-the-art results in weakly-supervised setup on this dataset.

We further compare the results on TACoS and ActivityNet Captions dataset. Note that videos in TACoS have longer averaged lengths, and the ground-truth video segments in ActivityNet Captions have longer averaged lengths. Although different baseline methods show different performances on two datasets, for example, SeqPAN achieves better results on ActivityNet, while VSLNet performs better on TACoS, the performance gain of VMRIL is stable.

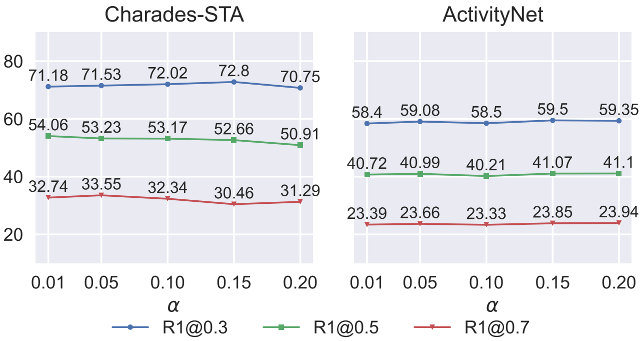


Figure 4: Qualitative results of mIoU with different hyperparameter α on two datasets.

As our proposed VMRIL is a general iterative learning based-method, it is model-agnostic and can be well adapted to any other fully-supervised Video Moment Retrieval method. Although there is no special network design in VMRIL compared with ViGA [5], VMRIL can achieve consistent performance gain based on SeqPAN [54] and VSLNet [53]. Above results and comparisons directly proves the efficacy of our proposed method.

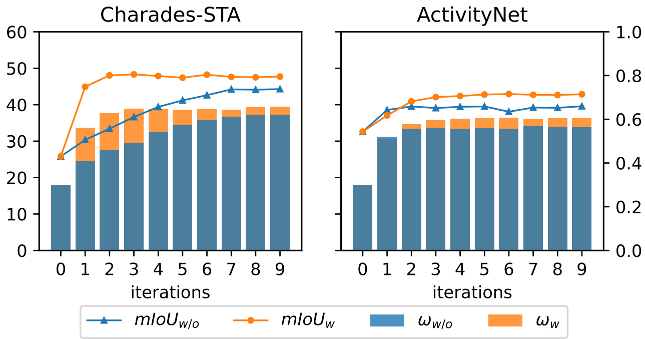
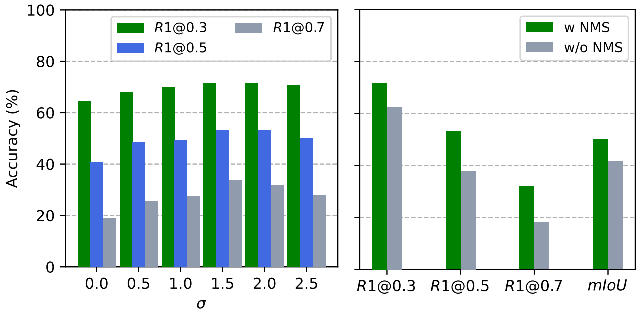


Figure 5: Performance comparison (%) of VMRIL (SeqPAN) w/ and w/o pretrained action model in different iterations on two datasets.

4.4.3 Qualitative Results. We further perform qualitative analysis on our method so as to enable a better understanding of its strength. The qualitative results of VMRIL (SeqPAN) on Charades-STA dataset are reported in Figure 7. According to the patterns and visualizations of three examples, the localized moments generated by VMRIL are very close to ground-truth, i.e., more accurate predictions, which also verifies the effectiveness of iterative learning and pretrained action localization model. Thanks to BMN model, the start and end boundaries of pseudo labels are roughly constrained. And starting from partial temporal labels are meaningful, we achieve better performance at low cost of annotation.

4.5 Ablative Studies

We finally conduct ablative experiments to analyze the effectiveness of different components in our approach. The ablation experiments presented below are based on three dataset with VMRIL (SeqPAN).



(a) Effect of σ values

(b) Effect of w/ and w/o NMS

Figure 6: The ablation studies of VMRIL (SeqPAN) on Charades-STA with different σ values and under the settings of w/ and w/o NMS.

Table 2: Performance comparison (%) of seed label in different distributions on Charades-STA dataset.

Distribution	Mean	R1@0.3	R1@0.5	R1@0.7	mIoU
fix	30%	72.19	55.16	34.19	50.18
	20%	71.02	47.23	24.49	46.6
	10%	68.04	44.17	24.25	45.73
beta	30%	71.64	53.41	33.06	49.9
	20%	68.12	44.7	25.24	46.69
	10%	66.8	44.81	23.09	44.76

4.5.1 Dataset Collection. As described in Section. 3.1, our reorganized dataset is labeled with fixed duration (fix). For the setting of λ, we conduct experiments with different values. As shown in Table 2, VMRIL with setting of $\lambda = 0.3$ achieves a trade-off between few label cost and satisfying performance. To benefit the VMR model in real application and relieve the restriction of fixed duration (the annotators can label the video clip with various lengths), we propose a more general situation (beta), where λ fits a Gaussian distribution of $\mu = 0.3$. We conduct experiments with this setting as shown in Table 2, our VMRIL can achieve comparable performance.

4.5.2 Selection of New Pseudo Labels . As shown in Figure 4, setting the hyper-parameter α as 0.15 helps achieve the best performance on two datasets. With the higher value of α, the performance of VMRIL will decrease sharply, since it requires generated pseudo labels with more overlapping with seed area L_{seed}.

In Figure 5, We also report the IoU curve of pseudo label and groundtruth, and IoU of pseudo label and seed area in different iterations. According to the curves, after 5 iterations, the performance results on three datasets all reach their saturation points in the end.

4.5.3 Effectiveness of VMRIL Components. We also carry out ablation studies on different components of proposed VMRIL method: (1) For the iterative learning part, we report the performance of

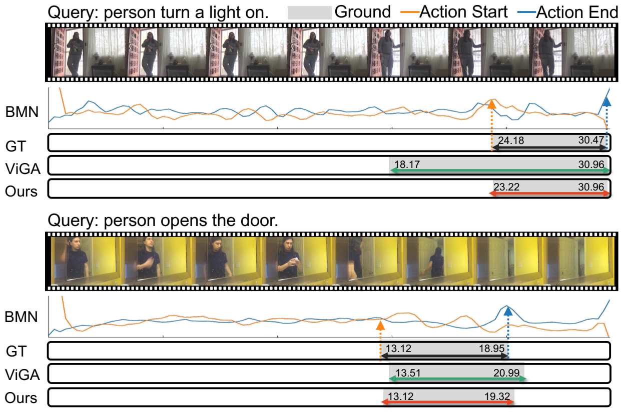


Figure 7: Qualitative results of VMRIL (SeqPAN) on Charades-STA dataset. BMN represents the start and end curve of predefined action. Compared with ViGA, our proposed VMRIL can generate more accurate temporal regions.

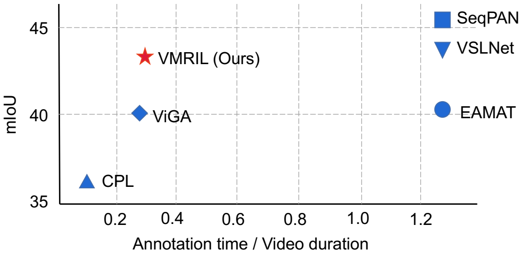


Figure 8: Annotation efficiency v.s. performance of different methods on the ActivityNet dataset.

VMRIL in different iterations. As shown in Figure 5, the performance of VMRIL on different datasets keeps steadily increasing and achieves saturation in around 5 iterations. In iteration 0, the VMRIL model is only trained with partial temporal labels. (2) For the utilization of pretrained action localization model BMN, the performance of VMRIL (SeqPAN) on three datasets will drop 3% ∼ 8% in mIoU without the pretrained model BMN, as shown in Figure 5.

To verify the effectiveness of multi-label training, we provide the ablation studies on w/ and w/o the multi-label training, as shown in Table 3 below. The performance of VMRIL (VSLNet) on three datasets will drop 0.64% ∼ 1.33% in R1@0.7 without the multi-label training. There are also ablation studies of σ in Figure 6 (a), which shows the robustness of our method in different value of σ. In Figure 6 (b), we provide abaltion studies of ourv method with and without NMS, which shows that it can achieve better performance with NMS.

5 IN-DEPTH ANALYSIS

Q1: What is the advantage of seed annotation compared with other supervisions? There are mainly four different kinds of annotation: (1) Fully supervised annotation, which requires precise start and end boundary of each video clip corresponding to the language query; (2) Weakly supervised annotation, which only provide the pair relation of video and language query without any location information; (3) Single-frame annotation, which randomly select one frame within the groundtruth; (4) Our proposed setting of partial temporal annotation as seed area.

Overall, the annotation cost of (1) is largest because the temporal boundary is not so discriminative and needs more time to localize the start and end timestamp. (2) requires smallest annotation cost but can not provide enough information, which result in the huge performance gap between fully-supervised VMR model and weakly-supervised VMR model. (3) is a compromise solution between (1) and (2), annotating one frame is at tiny cost compared with precise start and end boundaries. However, from our point of view, one frame is not enough for better VMR performance. Finally, (4) is an evolutionary version of weak label. Since annotating "seed" is not much harder than annotating a single frame but is much easier than full-annotation that requires precise "start" and "end". As shown in Figure 8, our method VMRIL (SeqPAN) can achieve comparable performance but requires much fewer annotation costs against fully-supervised methods.

Table 3: Performance comparison of (%) w and w/o multi-label training.

Dataset	multi-label	R1@0.3	R1@0.5	R1@0.7	mIoU
Charades-STA		71.53	53.23	33.55	50.59
	✓	71.29	55.16	34.19	50.18
ActivityNet		59.35	41.10	23.94	42.94
	✓	57.83	40.24	23.46	42.08
TACoS		43.56	31.14	17.72	30.68
	✓	47.66	35.04	19.05	33.00

Q2: Efficiency analysis of the proposed VMRIL. Our proposed VMRIL is based on the VMR model (such as SeqPAN, VSLNet). Our proposed VMRIL which utilizes pretrained action localization model and iterative learning runs with more parameters during training. When inferencing, the model size and running speed of VMRIL keep the same as the baseline VMR model.

Q3: Where does the performance gain come from? New setting or proposed framework? We think both new setting and proposed method benefit the performance gain. To validate the effectiveness of proposed framework, we run our VMRIL with the same single frame annotation as ViGA. Experimental results show that our proposed VMRIL can still outperform ViGA. Besides, we conduct experiments of ViGA with our seed label, with more annotated labels as supervision, ViGA can achieve better performance compared with single frame labels, which is also reasonable.

Q4: What about the generalization ability of the proposed framework, especially for other relevant datasets? Our proposed VMRIL uses the action localization pretrained model to provide guidance for the expansion of pseudo label. Action is the atomic element of complex event. Since the action localization pretrained model focus on the action related temporal span, it will generate response for complex event in video sequence. So our proposed VMRIL will also work on event-based dataset.

6 CONCLUSION

This paper focuses on Video Moment Retrieval task in a new label-efficient setting. We propose a new pipeline named Video Moment Retrieval via Iterative Learning (VMRIL). It starts training from the partial temporal region, which forms a striking contrast to those methods requiring fully-supervised ground truth. Specifically, we treat the partial temporal region as seed, and expand the pseudo label by iterative training. In order to restrict the expansion with reasonable boundaries, we utilize a pretrained video action model to provide coarse guidance of video segments. Experimental results demonstrate the effectiveness of our proposed method, which is even comparable with some fully-supervised methods but with fewer annotation costs. In the future, we will explore more label-efficient methods for VMR and make it closer to real application.

7 ACKNOWLEDGE

This work is supported by the Advanced Research and Technology Innovation Centre (ARTIC), the National University of Singapore under Grant (project number: A-8000969-00-00). This research is partially supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant.

REFERENCES

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *ICCV*. 5803–5812.

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*. 961–970.

[3] Joao Carreira and Andrew Zisserman. 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.

[4] Yun-Wei Chu, Kuan-Yen Lin, Chao-Chun Hsu, and Lun-Wei Ku. 2021. End-to-end Recurrent Cross-Modality Attention for Video Dialogue. *TASLP* (2021).

[5] Ran Cui, Tianwen Qian, Pai Peng, Elena Daskalaki, Jingjing Chen, Xiaowei Guo, Huyang Sun, and Yu-Gang Jiang. 2022. Video Moment Retrieval from Text Queries via Single Frame Annotation. *arXiv preprint arXiv:2204.09409* (2022).

[6] Jianfeng Dong, Xianke Chen, Minsong Zhang, Xun Yang, Shujie Chen, Xirong Li, and Xun Wang. 2022. Partially Relevant Video Retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*. 246–257.

[7] Jianfeng Dong, Xirong Li, and Cees GM Snoek. 2018. Predicting visual features from text for image and video caption retrieval. *IEEE Transactions on Multimedia* 20, 12 (2018), 3377–3388.

[8] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. 2022. Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 8 (2022), 4065–4080.

[9] Jianfeng Dong, Xiaoman Peng, Zhe Ma, Daizong Liu, Xiaoye Qu, Xun Yang, Jixiang Zhu, and Baolong Liu. 2023. From Region to Patch: Attribute-Aware Foreground-Background Contrastive Learning for Fine-Grained Fashion Retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1273–1282.

[10] Jianfeng Dong, Shengkai Sun, Zhonglin Liu, Shujie Chen, Baolong Liu, and Xun Wang. 2023. Hierarchical contrast for unsupervised skeleton-based action representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 525–533.

[11] Jianxiang Dong and Zhaozheng Yin. 2022. Boundary-aware Temporal Sentence Grounding with Adaptive Proposal Refinement. In *Proceedings of the Asian Conference on Computer Vision*. 3943–3959.

[12] Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. 2018. Weakly supervised dense event captioning in videos. *NeurIPS* 31 (2018).

[13] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *ICCV*. 5267–5275.

[14] Dongliang He, Xiang Zhao, Jizhou Huang, Fu Li, Xiao Liu, and Shilei Wen. 2019. Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos. In *AAAI*, Vol. 33. 8393–8400.

[15] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2018. Localizing moments in video with temporal language. *arXiv preprint arXiv:1809.01337* (2018).

[16] Jiabo Huang, Hailin Jin, Shaogang Gong, and Yang Liu. 2022. Video Activity Localisation with Uncertainties in Temporal Boundary. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIV*. Springer, 724–740.

[17] Jiabo Huang, Yang Liu, Shaogang Gong, and Hailin Jin. 2021. Cross-sentence temporal and semantic relations in video activity localisation. In *ICCV*. 7199–7208.

[18] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. 2018. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696* (2018).

[19] Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. 2022. Invariant grounding for video question answering. In *CVPR*. 2928–2937.

[20] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. 2019. Bmn: Boundary-matching network for temporal action proposal generation. In *ICCV*. 3889–3898.

[21] Zhijie Lin, Zhou Zhao, Zhu Zhang, Qi Wang, and Huasheng Liu. 2020. Weakly-supervised video moment retrieval via semantic completion network. In *AAAI*, Vol. 34. 11539–11546.

[22] Daizong Liu, Xiang Fang, Wei Hu, and Pan Zhou. 2023. Exploring optical-flow-guided motion and detection-based appearance for temporal sentence grounding. *IEEE Transactions on Multimedia* (2023).

[23] Daizong Liu and Wei Hu. 2022. Skimming, locating, then perusing: A human-like framework for natural language video localization. In *Proceedings of the 30th ACM International Conference on Multimedia*. 4536–4545.

[24] Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. 2021. Context-aware biaffine localizing network for temporal sentence grounding. In *CVPR*. 11235–11244.

[25] Daizong Liu, Xiaoye Qu, and Wei Hu. 2022. Reducing the vision and language bias for temporal sentence grounding. In *Proceedings of the 30th ACM International Conference on Multimedia*. 4092–4101.

[26] Daizong Liu, Xiaoye Qu, Xiao-Yang Liu, Jianfeng Dong, Pan Zhou, and Zichuan Xu. 2020. Jointly cross-and self-modal graph attention network for query-based moment localization. In *ACM Multimedia*. 4070–4078.

[27] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. 2018. Attentive moment retrieval in videos. In *SIGIR*. 15–24.

[28] Chujie Lu, Long Chen, Chilie Tan, Xiaolin Li, and Jun Xiao. 2019. DEBUG: A dense bottom-up grounding approach for natural language video localization. In *EMNLP*. 5147–5156.

[29] Minuk Ma, Sunjae Yoon, Junyeong Kim, Youngjoon Lee, Sunghun Kang, and Chang D Yoo. 2020. VLANet: Video-Language Alignment Network for Weakly-Supervised Video Moment Retrieval. In *ECCV*. 156–171.

[30] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K Roy-Chowdhury. 2019. Weakly supervised video moment retrieval from text queries. In *CVPR*. 11592–11601.

[31] Jonghwan Mun, Minsu Cho, and Bohyung Han. 2020. Local-Global Video-Text Interactions for Temporal Grounding. In *CVPR*. 10810–10819.

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

[33] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[34] Hoang-Anh Pham, Thao Minh Le, Vuong Le, Tu Minh Phuong, and Truyen Tran. 2022. Video Dialog as Conversation About Objects Living in Space-Time. In *ECCV*. Springer, 710–726.

[35] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. *ACL* 1 (2013), 25–36.

[36] Xindi Shang, Yicong Li, Junbin Xiao, Wei Ji, and Tat-Seng Chua. 2021. Video visual relation detection via iterative inference. In *Proceedings of the 29th ACM international conference on Multimedia*. 3654–3663.

[37] Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. 2017. Video visual relation detection. In *ACM Multimedia*. 1300–1308.

[38] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*. 510–526.

[39] Yijun Song, Jingwen Wang, Lin Ma, Zhou Yu, and Jun Yu. 2020. Weakly-supervised multi-level attentional reconstruction network for grounding textual queries in videos. *arXiv preprint arXiv:2003.07048* (2020).

[40] Reuben Tan, Huijuan Xu, Kate Saenko, and Bryan A Plummer. 2021. Logan: Latent graph co-attention network for weakly-supervised video moment retrieval. In *WACV*. 2083–2092.

[41] Weining Wang, Yan Huang, and Liang Wang. 2019. Language-driven temporal activity localization: A semantic matching reinforcement learning model. In *CVPR*. 334–343.

[42] Yuechen Wang, Jiajun Deng, Wengang Zhou, and Houqiang Li. 2021. Weakly supervised temporal adjacent network for language grounding. *IEEE Transactions on Multimedia* 24 (2021), 3276–3286.

[43] Jie Wu, Guanbin Li, Xiaoguang Han, and Liang Lin. 2020. Reinforcement learning for weakly supervised temporal grounding of natural language in untrimmed videos. In *ACM Multimedia*. 1283–1291.

[44] Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. 2022. Video as Conditional Graph Hierarchy for Multi-Granular Question Answering. AAAI.

[45] Shaoning Xiao, Long Chen, Songyang Zhang, Wei Ji, Jian Shao, Lu Ye, and Jun Xiao. 2021. Boundary Proposal Network for Two-Stage Natural Language Video Localization. In *AAAI*, Vol. 35. 2986–2994.

[46] Huijuan Xu, Abir Das, and Kate Saenko. 2017. R-c3d: Region convolutional 3d network for temporal activity detection. In *ICCV*. 5783–5792.

[47] Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. 2019. Multilevel language and vision integration for text-to-clip retrieval. In *AAAI*, Vol. 33. 9062–9069.

[48] Shuo Yang and Xinxiao Wu. 2022. Entity-aware and motion-aware transformers for language-driven action localization in videos. *arXiv preprint arXiv:2205.05854* (2022).

[49] Adams Wei Yu, David Dohan, Quoc Le, Thang Luong, Rui Zhao, and Kai Chen. 2018. Fast and accurate reading comprehension by combining self-attention and convolution. In *ICLR*.

[50] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. 2022. Semantic Conditioned Dynamic Modulation for Temporal Sentence Grounding in Videos. *TPAMI* 44, 05 (2022), 2725–2741.

[51] Yitian Yuan, Tao Mei, and Wenwu Zhu. 2019. To find where you talk: Temporal sentence localization in video with attention based location regression. In *AAAI*, Vol. 33. 9159–9166.

[52] Hao Zhang, Aixin Sun, Wei Jing, Liangli Zhen, Joey Tianyi Zhou, and Rick Siow Mong Goh. 2021. Parallel Attention Network with Sequence Matching for Video Grounding. *arXiv preprint arXiv:2105.08481* (2021).

[53] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. 2020. Span-based localizing network for natural language video localization. In *ACL*. 6543–6554.

[54] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. 2020. Learning 2d temporal adjacent networks for moment localization with natural language. In *AAAI*, Vol. 34. 12870–12877.

[55] Zhu Zhang, Zhijie Lin, Zhou Zhao, and Zhenxin Xiao. 2019. Cross-modal interaction networks for query-based moment retrieval in videos. In *SIGIR*. 655–664.

[56] Zhu Zhang, Zhijie Lin, Zhou Zhao, Jieming Zhu, and Xiuqiang He. 2020. Regularized two-branch proposal networks for weakly-supervised moment retrieval in videos. In *ACM Multimedia*. 4098–4106.

[57] Minghang Zheng, Yanjie Huang, Qingchao Chen, and Yang Liu. 2022. Weakly supervised video moment localization with contrastive negative sample mining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 3517–3525.

[58] Minghang Zheng, Yanjie Huang, Qingchao Chen, Yuxin Peng, and Yang Liu. 2022. Weakly Supervised Temporal Sentence Grounding with Gaussian-based Contrastive Proposal Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15555–15564.