Weakly Supervised Video Moment Retrieval via Location-irrelevant Proposal Learning

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

Ruiqi Shi
gzfoxie@sjtu.edu.cn
Shanghai Jiao Tong University
China

Yinwei Wei
e0962995@u.nus.edu
Monash University
Australia

Shanshan Zhao
sshan.zhao00@gmail.com
The University of Sydney
Australia

Roger Zimmermann
rogerz@comp.nus.edu.sg
National University of Singapore
Singapore

ABSTRACT

This paper deals with Video Moment Retrieval (VMR) in a weakly-supervised fashion, which aims to retrieve local video clips with only global video-level descriptions. Scrutinizing the recent advances in VMR, we find that the fully-supervised models achieve strong performance, but they are heavily relied on the precise temporal annotations. Weakly-supervised methods do not rely on temporal annotations, however, their performance is much weaker than the fully-supervised ones. To fill such gap, we propose to take advantage of a pretrained video-text model as hitchhiker to generate pseudo temporal labels. The pseudo temporal labels, together with the descriptive labels, are then utilized to guide the training of the proposed VMR model. The proposed Location-irrelevant Proposal Learning (LPL) model is based on a pretrained video-text model with cross-modal prompt learning, together with different strategies to generate reasonable proposals with various lengths. Despite the simplicity, we find that our method performs much better than the previous state-of-the-art methods on standard benchmarks, *e.g.*, +4.4% and +1.4% in mIoU on the Charades and ActivityNet-Caption datasets respectively, which benefits from training with fine-grained video-text pairs. Further experiments on two synthetic datasets with shuffled temporal location and longer video length demonstrate our model's robustness towards temporal localization bias as well as its strength in handling long video sequences.

CCS CONCEPTS

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

KEYWORDS

Video Moment Retrieval, Proposal Learning, Cross-modal

ACM Reference Format:
Wei Ji, Ruiqi Shi, Yinwei Wei, Shanshan Zhao, and Roger Zimmermann. 2024. Weakly Supervised Video Moment Retrieval via Location-irrelevant Proposal Learning. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion), May 13–17, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3589335.3651942

1 INTRODUCTION

Given a natural language sentence query, Video Moment Retrieval (VMR) aims to locate the temporal video segment of an untrimmed video corresponding to the query description. As an important cross-modal task which bridges the gap between natural language understanding and computer vision, VMR has been widely applied in a series of real-world applications, such as information retrieval [6], human-computer interaction [26], *etc.*. The core challenge of VMR lies in the cross-modal semantic alignment of video clip and language query.

Existing VMR models working in a fully-supervised setting heavily rely on well-annotated labels and suffer from annotation inconsistency among different annotators [25]. To relieve the burden on collecting precise annotations, some existing works focus on the weakly-supervised setting that treats video-text pair as supervision without the need on temporal information, as shown in Figure 1. Since the training data are query and untrimmed video pairs, it is unstable to train the weakly supervised VMR model due to the partial video clips irrelevant with query as noise, which is also harmful to the model performance. Hence, there exists a large performance gap between the fully supervised and weakly supervised VMR methods. We think trimming the background region in video sequence for fine-grained well-aligned video-query pairs seems a promising direction to pursue better performance for weakly supervised VMR methods.

Based on the observations above, we consider how to achieve satisfying performance while relieving the heavy burden of well-annotated labels. Recently, a series of large pretrained video-text models have been proposed, such as ClipBert [13], Frozen [2], *et al.*. We therefore consider how to deal with Video Moment Retrieval task by utilizing the knowledge of cross-modal semantic alignment from pretrained video-text models so as to keep a trade-off between accuracy and labor-intensive label collection.

WWW '24 Companion, May 13–17, 2024, Singapore, Singapore

ACM ISBN 979-8-4007-0172-6/24/05…$15.00
https://doi.org/10.1145/3589335.3651942

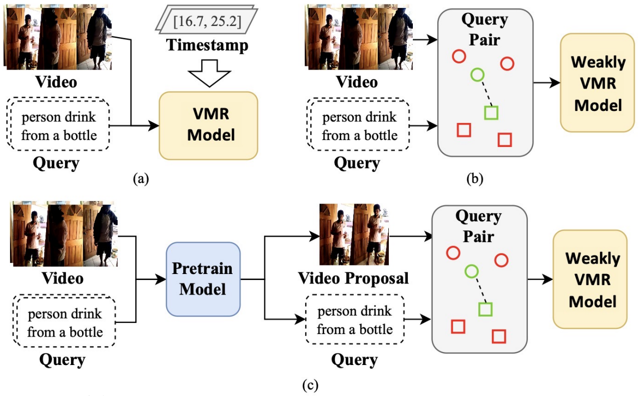


Figure 1: (a) Current fully supervised VMR methods relies on the precise annotation of start/end timestamps. (b) Most weakly supervised VMR methods are following the Multiple Instance Learning (MIL) pipeline, which treats paired video-text as positive samples (green circle/square represents paired video/text), unpaired video and text (red circles/squares) as negative samples. (c) Our proposed LPL method utilizes a pretrained video-text model to generate pseudo labels, and trim relevant video clip with query as positive samples. Our model is insensitive to location bias and can achieve better performance due to aligned video-text pairs.

Since the pretrained video-text models are trained with large amounts of video-text data, these models have good ability of semantic matching between language query and video. We can first divide the untrimmed video into several non-overlapping clips, and then feed the video clips into the pretrained model to calculate the similarity score of each video clip and query. We then select the video clip with the highest similarity as seed, and propose three different strategies to generate segments in various length to adjust the temporal boundary of pseudo label: (a) Naive: directly calculate the similarity of each clip and query; (b) Greedy: iteratively erase the irrelevant part according to similarity score; and (c) Anchor: select the video clip with the highest similarity score as seed, generate proposal in various lengths with different radiuses.

For the architecture of pretrained model, we design a prompt-based encoder to generate the visual embedding of video sequence with global context. The visual embedding is concatenated with the textual embedding from text encoder. Finally, the whole network with prompt encoder is finetuned with pseudo labels in conjunction with description labels. Since our proposed model is trained with splitted video clips, it is less sensitive to the location bias in videos [39], and can better deal with videos in a long sequence. To verify such strengths, we artificially synthesize two VMR datasets with longer video length and diverse temporal locations. Our extensive experiments on the synthesized datasets show that our model performs better towards perturbation of temporal locations compared with established methods.

Our main contributions are summarized as follows:

- We propose a novel prompt-based VMR model with Location-irrelevant Proposal Learning (LPL) on the basis of pretrained video-text model, which includes reasonable proposal generation and prompt-based cross-modal feature fusion. Besides, we propose three different strategies of generating reasonable proposals with various lengths.
- We synthesize two VMR datasets with shuffled temporal locations and longer video lengths, to validate the robustness and effectiveness of our proposed LPL model and SOTA methods.
- Experiments on two public VMR datasets and synthetic datasets demonstrate the effectiveness and consistent superiority of our methods compared to the current SoTAs.

2 RELATED WORK

2.1 Video Moment Retrieval

Video Moment Retrieval (VMR) is defined as retrieving video segments with consistent semantics of query [1, 15–17, 35, 36]. 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 [10, 18] 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 [8] and then calculate the multi-modal semantic matching to find the best matching proposal for the query. For proposal-free methods, Yuan *et al.* [42] learn cross-modal interactions between video and query, treat the video as a whole and directly predict the temporal coordinates. DEBUG [19] proposes 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 final result. Recently, Zhang *et al.* [45] propose a two-dimensional temporal map to model the temporal relations of moments with variant length, in which the two dimensions indicate the start and end timestamps, respectively. Apart from these, there are also works [9, 31] which adopt reinforcement learning to make decisions on the action space of candidate segments, such as the start/end boundaries move left/right, corresponding to the language query matching result. The performance of all these methods mentioned above are heavily relied on the well-annotated datasets.

In weakly supervised VMR, most works [14, 23, 28, 29, 47] treat the paired video-text as positive samples, video and unpaired text or text and unpaired video as negative samples, and then follow the multi-instance learning paradigm. [21] selects appropriate candidate proposals by calculating similarity between each proposal and the query and proposes a Cascaded Cross-modal Attention module to perform both intra-/inter-modality attention. [34] employs a novel boundary adaptive refinement process and reinforcement learning to deal with the wVMR task. [46] proposes a language-aware filter to generate the enhanced and suppressed video streams, and then generate positive and negative proposals, respectively. Although weakly supervised methods can lower the cost of temporal annotation, there exists noise in the untrimmed video, which makes the training unstable. In this paper, we utilize a pretrained

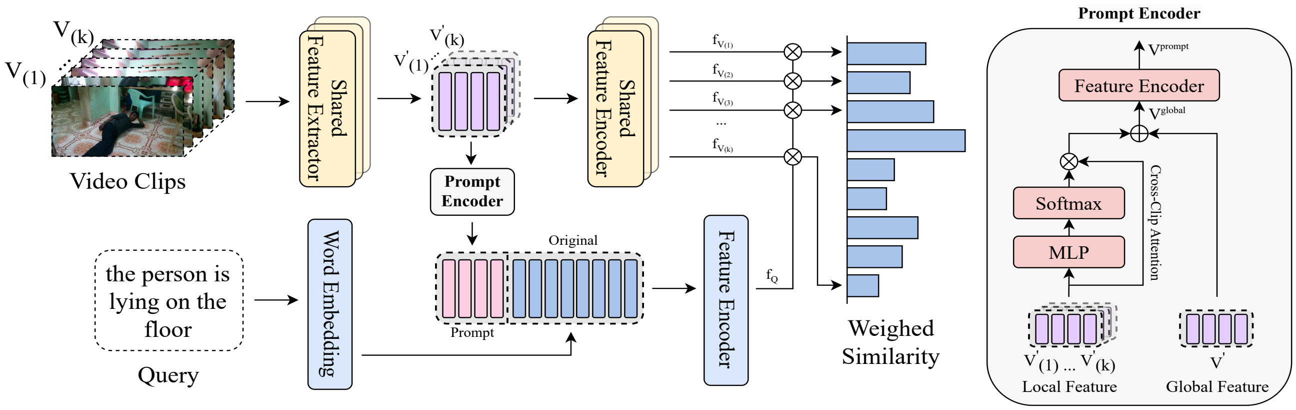


Figure 2: The overall framework of our LPL model. Based on the pretrained video-text model, we propose a new prompt encoder module to fuse the global video feature as visual embedding, and then concatenate with textual embedding. We first select the video clip with the highest similarity score computed by pretrained video-text model. Then after the selection of proposal generation, we obtain the pseudo label as supervision, we train the whole model with prompt encoder in an end-to-end manner.

video-text model to erase the video clip and preserve the responsive regions corresponding to the query, which can get rid of fully annotated labels and achieve well-aligned video-text pairs.

2.2 Prompt Learning

Prompt learning aims at designing reformat templates to transfer knowledge from pretrained models to downstream tasks, which is first proposed by GPT-3 [3]. Different from fine-tuning which is inefficient with enormous amounts of parameters, prompt learning has been proved effective in a series of multi-modal tasks, such as image captioning [33], visual question answering [40], visual grounding [41], *etc.*. In the zero-shot Video Moment Retrieval task, Jinwoo *et al.*[24] generate temporal event proposal and pseudo query via pretrained textual and action models to train the NLVL model. PZVMR [30] improves the zero-shot VMR performance by fusing visual features in the stage of proposal generation and pseudo query generation. Kim *et al.* [12] propose a language-free framework for zero-shot VMR, which skips doubtful sentence generation. Different from these recent works above, we introduce a visual feature prompt-based VMR model, which fuses the cross-modal features before calculating similarity in different domains.

2.3 Video-text Pretrained Model

With the release of large-scale video-text pretraining datasets, such as HowTo100M [22], WebVideo-2M [2], video-text pretraining has become a new trend with significant attention in the community to learn better video-text cross-modal representations. For example, ClipBert [13] reduces the pretraining cost via sparsely sampling frames from consecutive video sequences. Frozen [2] proposes a visual encoder based on ViT [7], and splits the image into sequential patches, which makes it possible to train image and video with text seamlessly. Different from LocVTP [5], which utilizes a pretrained video-text model to deal with video retrieval and location tasks in fully-supervised setting, our work focuses on the weakly supervised setting, which leverages a pretrained model to generate pseudo temporal labels, and thus relieves the burden on dense time span annotations in downstream tasks.

3 METHOD

We introduce a prompt-based VMR model named LPL as shown in Fig. 2. In this section, we first provide the problem definition in Sec. 3.1. Then we introduce detailed network architecture and prompt learning design in Sec. 3.2 and Sec. 3.4. Finally, we introduce the process of proposal generation in Sec. 3.3 and validate the robustness to temporal location bias and generalization ability to long videos. We also introduce two synthetic datasets in Sec. 3.5.

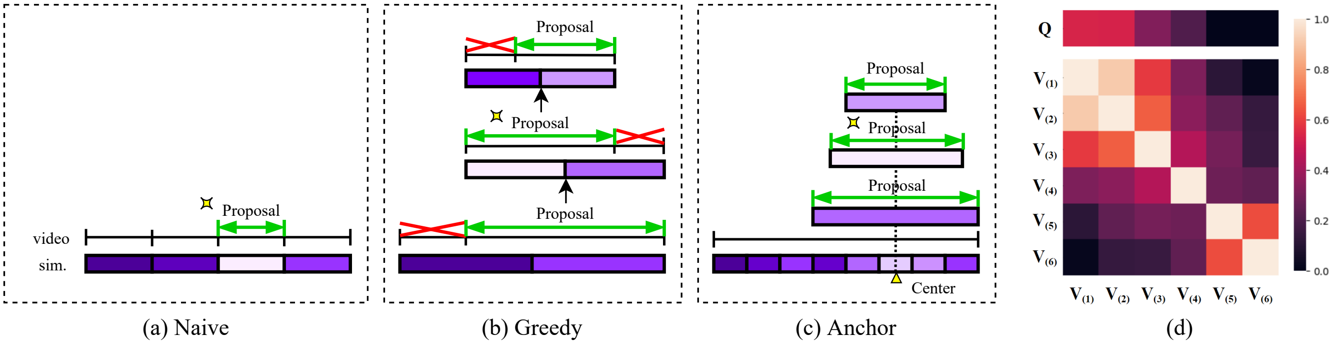
3.1 Problem Definition

Given an untrimmed video $V = \{v_t\}_{t=1}^{T}$ and the language query $Q = \{q_j\}_{j=1}^{M}$, where T and M are the numbers of frames and words, respectively, the goal of Video Moment Retrieval is to predict the start and end timestamps (τ^s, τ^e) in the video corresponding to query Q, where $(\tau^s, \tau^e) = f(V, Q)$. In the zero-shot setting, the ground truth of (τ^s, τ^e) is not accessible. Hence, we utilize the pretrained video-text model to provide supervision of the temporal location in this paper.

Since the pretrained video-text model has been trained with large amounts of video-text pairs, we believe the pretrained model excels at semantic matching between text and video sequence. Hence, for the temporal location of one single video, we consider erasing the irrelevant part or selecting the most matching part in the video.

3.2 Network Architecture

For the pretrained video-text model, we follow the structure of Frozen [2]: 1) For video $V \in \mathrm{R}^{T\times 3\times H\times W}$, the input video sequence is first divided into $M\times N$ non-overlapping patches of size $P\times P$, where $N = HW/P^2$. After fed into a 2D convolutional layer to reduce dimension and added spatial and temporal positional embeddings to ascertain the temporal and spatial position of patches, the video sequence is then fed into a stack of space-time transformer blocks.



(a) Naive (b) Greedy (c) Anchor (d)

Figure 3: Three different strategies of proposal generation.(a) Naive: directly calculate the similarity of each clip and query; (b) Greedy: iteratively erase the irrelevant part according to similarity score; (c) Anchor: select the video clip with the highest similarity score as seed, generate proposals in various lengths with different radiuses. (d) represents the similarity matrix in the Naive method, where adjacent video clips have relatively higher similarities.

2) For textual embedding, the text encoder architecture is a multi-layer bidirectional transformer encoder, For the final text encoding, we use the output of the '[CLS]' token.

Finally, we use single linear layers to project text and video embeddings into a common dimension. We compute the similarity between text and video by performing the dot product between the two projected embeddings.

3.3 Proposal Generation

Based on the architecture introduced above, we consider how to adapt it to temporal sentence localization. Towards this, we propose three different strategies to generate proposals as illustrated in Fig. 3.

Naive: The intuitive idea is to select the most relevant clip as pseudo label. We first split the video V in k clips. For each vlip $V_{(m)}$, we calculate the similarity score of $V_{(m)}$ and query Q first and update the similarity score by comparing with global context:

$$\begin{aligned} Sim(V_{(m)}, Q) &= \frac{\sum_{n=1}^{k} v_{sim}(V_{(m)}, V_{(n)}) \cdot t_{sim}(V_{(n)}, Q)}{\sum_{n=1}^{k} v_{sim}(V_{(m)}, V_{(n)})} \\ &= \frac{\sum_{n=1}^{k} (f(V_{(m)})^T f(V_{(n)})) \cdot (f(V_{(n)})^T f(Q))}{\sum_{n=1}^{k} (f(V_{(m)})^T f(V_{(n)})} \end{aligned} \tag{1}$$

where Q means language query, v_{sim} and t_{sim} refers dot product. The similarity of video clip $V_{(m)}$ and query Q is determined by the similarity of all video clips $V_{(n)}$ and query Q, depending on how similar $V_{(m)}$ and $V_{(n)}$ is. A visualized example of the similarity between video clips, and between video clips and the query, is shown in Figure 3 (d).

Then, the candidate proposal in video V corresponding to query Q is calculated by:

$$V^* = \arg\max_{p} \ Sim(V_{(m)}, Q) \tag{2}$$

Greedy: Besides selecting the most relevant clips, we consider erasing the most irrelevant clips from the raw video based on the similarity score, as shown in Figure 3 (b). We first divide the video V into two clips: V_A and V_B with length $[L/2]$, and calculate the similarity score of V_A and V_B with query. Then, we select the clip with lower similarity score, take V_A as example, and erase half part of V_A which is not adjacent to V_B. If the similarity score of V_A and V_B are the same, we random select the half part close to the video temporal boundaries. The procedure can be repeated until the similarity score is not changed compared with last round. In each turn, the proposal will be shrunk with $[3L/4]$ length compared with last turn.

Anchor: As shown in Figure 3 (c), we first divide the video V in k clips and calculate the similarity score S_k for each clip. Then, we select the video clip with the highest similarity score as the center point, and d as radius to extend, which generates the proposal in this round. To generate the proposals with different lengths, we set different radiuses value in the experiment. Finally, we select the proposal with largest similarity as the pseudo label.

3.4 Prompt Learning

Although the pretrained model has good performance in semantic matching of video and text, there is domain gap between pretrained data and samples in downstream tasks. Hence, we consider using prompt-based visual encoder to fuse the multi-granularity visual feature into text encoder, so as to bridge the pretrained model and downstream datasets with pseudo labels. To be specific, we extract the visual feature of each video clip as local feature, then we calculate the global feature via fusing local features. With global feature as input, prompt encoder outputs the visual prompt features, which are concatenated with textual embeddings. The detailed structure of prompt encoder can be described as:

$$V^{prompt} = Transformer(V^{global}) \tag{3}$$

where we follow the Transformer structure in Frozen [2], and the global feature V^{global} is computed through the cross-attention of video clips:

$$f_{(m)} = MLPs(ViT(V_{(m)})), \tag{4}$$

$$\alpha_{(m)} = \frac{exp(f_{(m)})}{\sum_{i=1}^{k} exp(f_{(i)})}, \tag{5}$$

$$V^{global} = \lambda \sum_{i=1}^{k} \alpha_{(m)} ViT(V_{(m)}) + ViT(V). \tag{6}$$

Rather than directly compute the cross-modal similarity of features in visual and text space, our prompt encoder fuse the global visual information into the text encoding.

Training Stage. In the training stage, the pretrained model with prompt encoder is trained with proposals generated in Section 3.3. The whole model is trained in an end-to-end manner.

We follow [2] and train the whole model in a retrieval setting. We crop each video according to the pseudo label. We treat the matched text-video pairs in the batch as positives, and one video and other unpaired texts or one text with unpaired videos can be treated as negative samples. video-to-text and text-to-video:

$$L_{v2t} = -\frac{1}{B}\sum_{i}^{B} \log \frac{\exp(x_i^{\top} y_i/\sigma)}{\sum_{j=1}^{B} \exp(x_i^{\top} y_j/\sigma)} \tag{7}$$

$$L_{t2v} = -\frac{1}{B}\sum_{i}^{B} \log \frac{\exp(y_i^{\top} x_i/\sigma)}{\sum_{j=1}^{B} \exp(y_i^{\top} x_j/\sigma)} \tag{8}$$

where x_i and y_j are the normalized embeddings of i-th cropped video within the temporal region of pseudo label and the j-th text respectively in a batch of size B. σ is the parameter of temperature. We minimize the sum of two losses until the model convergences.

Inference. When testing, each video is first segmented into k non-overlapping video clips, and then these video clips are fed into trained model, the prediction can be produced the same as the process of proposal generation. We select the proposal with largest similarity as the final result.

3.5 Synthetic Datasets

To evaluate the robustness of proposed method towards temporal location bias and the effectiveness of dealing with long video sequence, we propose two synthetic datasets: Take the Chardes dataset as example, for each video-query pair V_i with groundtruth gt_i, we select a video V_j from testing set of video corpus, and concatenate them together as V_{i+j}^{pseudo}. To be specific, we select V_j based on the similarity of Q and Q_j, which should be as small as possible so that there will not be regions with similar semantics of Q in the selected video V_j. Hence, the relative location of gt_i in the pseudo video sequence V_{i+j}^{pseudo} will change accordingly, and the whole length of video is enlarged.

Different from randomly generating video clips like [39] described, we follow a "select-by-retrieve" pipeline to concatenate videos with videos in real scenario from the video corpus, which can greatly enriched the semantic space of the synthetic datasets compared with [39].

We conduct experiments on pseudo testing videos to validate the robustness of current SOTA models and our proposed method. The method of generating synthetic data can also be used as data augmentation strategy when training.

4 EXPERIMENT

4.1 Datasets

To evaluate the performance of our proposed, we conduct experiments on two challenging Video Moment Retrieval datasets: **Charades-STA** [8] is composed of daily indoor activities videos, which is based on Charades dataset [27]. This dataset contains 6,672 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** [4] is originally constructed for dense video captioning, which contains about 20k YouTube videos with an average length of 120 seconds.

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 over all testing samples. The mIoU metric is particularly challenging for short video moments; **Recall:** We adopt "R@n, IoU = μ" as the evaluation metrics, following [8]. The "R@n, IoU = μ" 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 the structures of the original pretrained model, we follow the implementation in Frozen [2]. During the training process, we split the video into $\{1, 2, 4, 8\}$ clips for the naive proposal generation, and use anchors of $\{1/6, 1/5, 1/4\}$ for the anchor proposal generation. During the testing process, we split the video into 16 clips for the original datasets, and 24 clips for the synthetic datasets. We use 768 dimensions of hidden layers and set $\lambda = 1$ in the prompt encoder. All datasets are trained for 15 epochs. The learning rate is set at $3e-5$ for the prompt encoder and $1e-5$ for others. All experiments are conducted on 2 Nvidia Tesla V100 GPUs with 32GB memory. More details can be found in Supplementary Material.

4.4 Comparison with SOTA methods

Table 1 summarizes the experimental results on Charades-STA and ActivityNet Captions dataset. We mainly compare our HUAL with the following SOTA methods: **1) Fully-supervised method:** CTRL [8], QSPN [37], 2D-TAN [45], VSLNet [44], SeqPAN [43]; **2) Weakly-supervised method:** TGA [23], SCN [14], BAR [34], RTBPN [46], VLANet [21], MARN [28], LoGAN [29], CRM [11]; **3) Pretrained model-based method:** LTZVG [12], PZVMR [30]. LTZVG and PZVMR are two methods use pretrained model CLIP. Besides, LocVTP [5] is also a pretrained model-based method but in fully supervision, while our LPL method are focusing on utilizing pretrained model to relieve the burden of annotation in downstream tasks, which is in weakly supervised method.

In the Table 1, LPL (Baseline) means pretrained video-text model without proposal generation and prompt encoder. From the results we observe that our LPL can effectively improve the performance of baseline networks over all metrics and benchmarks. For Charades-STA dataset, we can see that LPL works well in even stricter metrics, such as R@0.7. Compared with LTZVG, LPL achieves a significant 4.35% absolute improvement in mIoU on Charades dataset, which demonstrates the effectiveness of proposed model. Our LPL is superior to most weakly supervised VMR methods, and has surpassed some fully supervised VMR models. We further compare the results on ActivityNet Captions dataset. Note that the ground-truth video segments in ActivityNet Captions dataset have a longer averaged duration but in various length, which is more challenging.

Table 1: Performance comparison of LPL and the state-of-the-art methods under different supervision settings on two public datasets. The best performance of using a pretrained model is in red, while the second-best is in blue. All fully supervised methods use temporal location information, and weakly supervised method and pretrained model-based method list above are proposed without temporal location labels. CRM* requires an additional paragraph description annotation. CNM† and CPL‡ use Gaussian Mask generation, controllable sample mining strategy, and fancy loss designs, which are also adaptive to our LPL method for better performance.

Supervision	Method	Charades-STA				ActivityNet Captions			
		R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU
Full Supervision	CTRL [8]	-	23.63	8.89	-	-	-	-	-
	QSPN [37]	54.7	35.6	15.8	-	45.3	27.7	13.6	-
	2D-TAN [45]	-	39.7	23.31	-	59.45	44.51	26.54	-
	VSLNet [44]	70.46	54.19	35.22	50.02	63.16	43.22	26.16	43.19
Weak Supervision	TGA [23]	32.14	19.94	8.84	-	-	-	-	-
	SCN [14]	42.96	23.58	9.97	-	47.23	29.22	-	-
	BAR [34]	44.97	27.04	12.23	-	49.03	30.73	-	-
	RTBPN [46]	60.04	32.36	13.24	-	49.77	29.63	-	-
	VLANet [21]	45.24	31.83	14.17	-	-	-	-	-
	MARN [28]	48.55	31.94	14.81	-	47.01	29.95	-	-
	LoGAN [29]	51.67	34.68	14.54	-	-	-	-	-
	CRM* [11]	53.66	34.76	16.37	-	55.26	32.19	-	-
	CNM† [47]	60.04	35.15	14.95	-	55.68	33.33	-	37.14
	CPL‡ [48]	66.40	49.24	22.39	43.48	55.73	31.37	-	-
Pretrained Model	LTZVG [12]	52.95	37.24	19.33	36.05	47.61	**32.59**	15.42	31.85
	PZVMR [30]	46.83	33.21	18.51	32.62	45.73	31.26	**17.84**	30.35
	LPL (Baseline)	51.86	33.31	16.75	34.09	48.24	30.44	10.16	31.83
	LPL	**62.34**	**41.30**	**19.54**	**40.40**	**48.60**	30.61	12.32	**33.25**

Compared with LTZVG, LPL achieves a significant 1.40% absolute improvement in mIoU.

Also, the qualitative results of LPL on Charades-STA dataset are reported in Figure 4 (a). Compared with other weakly supervised VMR methods, our method can better localize the interval corresponding to the query.

4.5 Ablation Studies

We mainly conduct the ablation studies and make comparison on the Charades-STA dataset. More analysis and results can be found in the Supplementary Material.

Effect of Prompt Learning. In the prompt encoder, we mainly compare the performance of visual feature concatenated in different locations. As shown in Table 2,"Front" denotes prepending the video feature in front of the text feature in the text encoder. "End" denotes appending the video feature behind the text feature. "Replace" means replacing '[CLS]' in the text feature with visual features. Experimental results show that "Replace + Global Feature" achieves the best performance compared with other combinations. In addition, methods using global features always perform better than those using local feature. Such results are reasonable and verify our hypothesis.

Table 2: Performance comparison (%) of LPL with different locations of visual prompt on Charades dataset.

Prompt Position	R@0.3	R@0.5	R@0.7	mIoU
Front + Global Feature	50.65	31.72	15.97	33.24
End + Global Feature	51.08	32.88	16.72	34.20
Replace + Global Feature	**55.83**	**35.32**	**16.94**	**36.26**
Replace + Local Feature	40.89	24.97	11.75	27.21
Replace + Cross Attention	51.43	33.20	17.88	34.24

Table 3: Performance comparison (%) of LPL with different numbers of split on Charades dataset.

Split	R@0.3	R@0.5	R@0.7	mIoU
1	58.74	34.17	16.34	37.47
2	**61.08**	34.17	15.40	38.30
4	59.65	37.82	**18.47**	**38.61**
8	56.75	**38.68**	16.40	36.82

Effect of Proposal Generation. We also compare the performance with different numbers of split, as shown in Table 3."Split 1" means taking the whole video sequence as input. In the setting of 4 splits, LPL achieves the best performance, which is a trade-off between accuracy and computation cost. As shown in Table 4, we analysis the performance of different strategies in the process of proposal generation. Among them, "Anchor" strategy achieves the best performance in mIoU, which is also proved effective in other tasks, such as object detection, action localization. As shown in Table 5, we further compare different settings of anchors on the Charades dataset. According to the overall performance, we select anchor of { 1/6, 1/5, 1/4 } as default.

Synthetic Dataset. To validate the robustness of our method towards temporal localization bias as well as its strength in handling long video sequence, we conduct experiments on the synthetic datasets. As described in Sec. 3.5, for each video-language pair,

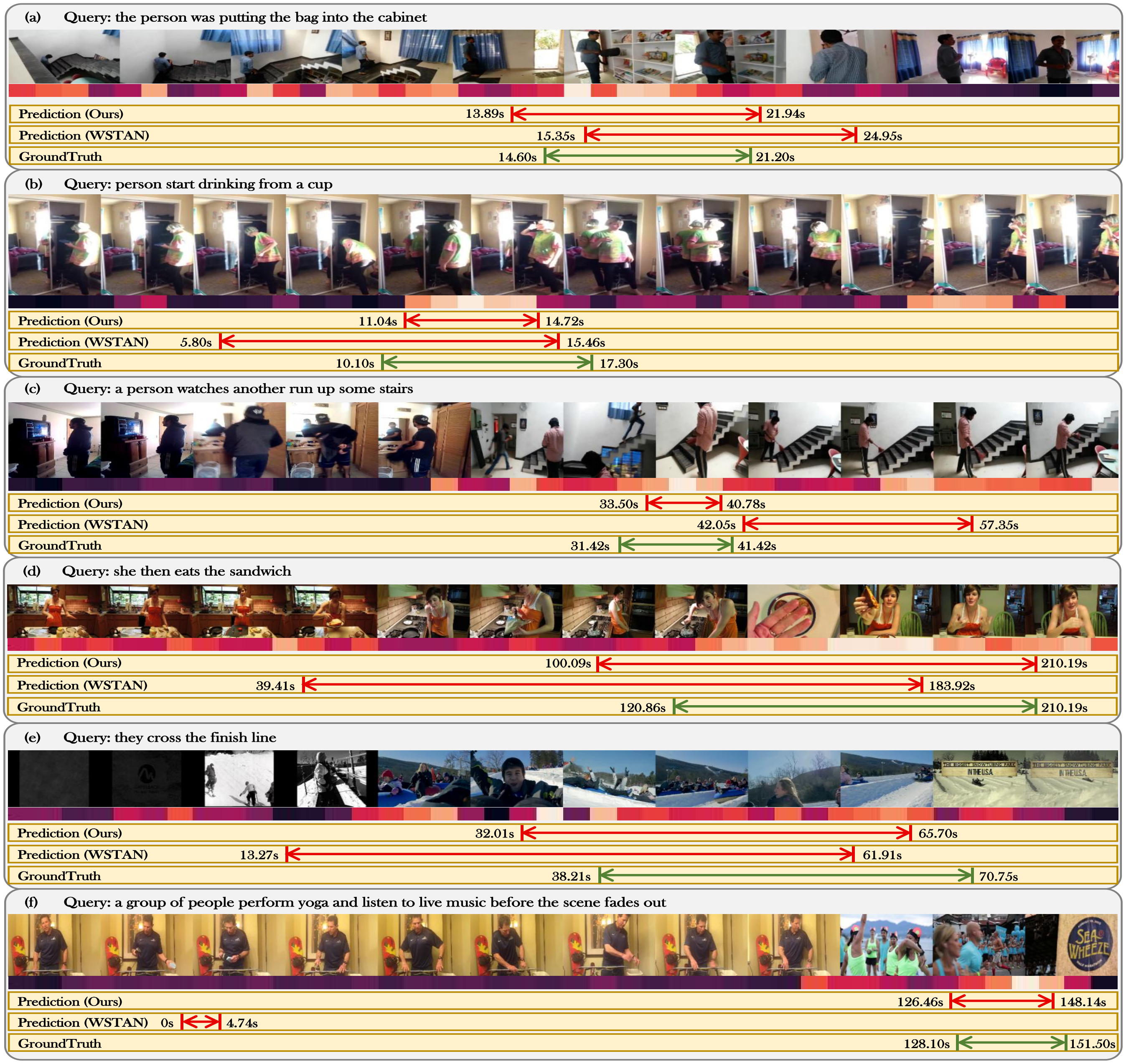

(a) Query: the person was putting the bag into the cabinet
Prediction (Ours) 13.89s 21.94s
Prediction (WSTAN) 15.35s 24.95s
GroundTruth 14.60s 21.20s
(b) Query: person start drinking from a cup
Prediction (Ours) 11.04s 14.72s
Prediction (WSTAN) 5.80s 15.46s
GroundTruth 10.10s 17.30s
(c) Query: a person watches another run up some stairs
Prediction (Ours) 33.50s 40.78s
Prediction (WSTAN) 42.05s 57.35s
GroundTruth 31.42s 41.42s
(d) Query: she then eats the sandwich
Prediction (Ours) 100.09s 210.19s
Prediction (WSTAN) 39.41s 183.92s
GroundTruth 120.86s 210.19s
(e) Query: they cross the finish line
Prediction (Ours) 32.01s 65.70s
Prediction (WSTAN) 13.27s 61.91s
GroundTruth 38.21s 70.75s
(f) Query: a group of people perform yoga and listen to live music before the scene fades out
Prediction (Ours) 126.46s 148.14s
Prediction (WSTAN) 0s 4.74s
GroundTruth 128.10s 151.50s

Figure 4: Qualitative results of LPL on Charades-STA (a)(b)(c), and ActivityNet (d)(e)(f). (a)(b)(d)(e) use the original datasets. Compared with other weakly supervised VMR methods, our method can better localize the interval corresponding to the query. (c)(f) use synthetic datasets. Our LPL is less sensitive to the location bias compared with the WSTAN method.

Table 4: Performance comparison (%) of LPL with different strategies of proposal generation on Charades dataset.

Proposal Generation	R@0.3	R@0.5	R@0.7	mIoU
Naive	**59.65**	37.82	18.47	38.61
Greedy	58.39	40.51	19.44	38.37
Anchor	58.28	**43.39**	**19.54**	**39.27**

we select another video from testing set and concatenate them together, so that the whole video length will be doubled and the relative location of ground truth will also be changed. As shown in Tab. 6, our LPL model shows the smallest performance degradation compared with other fully supervised and weakly supervised VMR methods. Compared with fully supervised VMR method, weakly supervised VMR methods are more vulnerable to temporal location change (*e.g.*, CPL and WSTAN will decrease to almost 0 in the

Table 5: Performance comparison (%) of LPL with different settings of anchors on Charades dataset.

Anchor	R@0.3	R@0.5	R@0.7	mIoU
1/6, 1/5, 1/4, 1/3, 1/2, 1	56.24	36.72	15.27	36.14
1/6, 1/5, 1/4	58.28	**43.39**	**19.54**	**39.27**
1/3, 1/2, 1	**59.25**	42.45	19.06	39.05
1/8, 1/4, 1/2, 1	57.42	38.82	16.34	37.42
1/8, 2/8, 3/8, 4/8	58.49	39.27	16.61	37.59
5/8, 6/8, 7/8, 1	58.33	40.08	19.33	38.45

Table 6: Performance comparison (%) of LPL and other methods on synthetic Charades dataset (Up), and synthetic ActivityNet dataset (Down). 2D-TAN, VSLNet, and EAMAT are fully supervised methods. LPL, WSTAN, and CPL are weakly supervised methods. "raw" means original testing data on Charades, "syn." represents the synthetic dataset of Charades.

Method	R@0.3	R@0.5	R@0.7	mIoU
LPL (raw)	51.86	33.31	16.75	34.09
LPL (syn.)	30.67	18.98	8.04	20.10
Δ (↓)	**-40.85%**	**-43.02%**	**-52.00%**	**-41.04%**
2D-TAN [45] (raw)	57.31	42.80	23.23	39.22
2D-TAN [45] (syn.)	33.44	11.88	2.61	19.75
Δ (↓)	-41.65%	-72.24%	-88.76%	-49.64%
VSLNet [44] (raw)	71.45	54.57	35.27	50.44
VSLNet [44] (syn.)	30.08	13.04	4.09	20.26
Δ (↓)	-57.90%	-76.10%	-88.40%	-59.83%
EAMAT [38] (raw)	70.86	55.78	35.54	50.61
EAMAT [38] (syn.)	34.73	19.51	7.12	22.59
Δ (↓)	-50.99%	-65.02%	-79.97%	-55.36%
WSTAN [32] (raw)	43.39	29.35	12.28	27.83
WSTAN [32] (syn.)	0.30	0.03	0.00	0.24
Δ (↓)	-99.31%	-99.90%	-100.00%	-99.14%
CPL [48] (raw)	66.40	49.24	22.39	43.48
CPL [48] (syn.)	7.00	1.55	0.25	4.81
Δ (↓)	-89.46%	-96.85%	-98.88%	-88.94%

Method	R@0.3	R@0.5	R@0.7	mIoU
LPL (raw)	48.98	31.11	8.81	32.00
LPL (syn.)	38.47	19.13	7.04	25.02
Δ (↓)	**-21.46%**	**-38.53%**	**-20.13%**	**-21.82%**
2D-TAN [45] (raw)	58.75	44.05	27.38	43.29
2D-TAN [45] (syn.)	24.72	11.52	4.35	26.24
Δ (↓)	-57.92%	-73.85%	-84.11%	-39.39%
VSLNet [44] (raw)	57.25	40.79	25.36	41.86
VSLNet [44] (syn.)	15.94	6.16	2.19	13.99
Δ (↓)	-72.16%	-84.90%	-91.36%	-66.58%
WSTAN [32] (raw)	42.13	25.64	10.01	29.83
WSTAN [32] (syn.)	15.88	5.37	1.67	12.84
Δ (↓)	-62.31%	-79.06%	-83.32%	-56.96%
CPL (raw)	53.97	31.66	13.52	36.63
CPL (syn.)	16.80	4.90	1.28	14.95
Δ (↓)	-68.87%	-84.52%	-90.53%	-59.19%

metric of R@0.7, which is a highly destructive performance drop).

We also provide the visualization result in Fig. 4 (b), which shows that our LPL can locate the moment in long videos.

5 IN-DEPTH ANALYSIS

Q1: Does the selection of pretrained models affect the results? We provide the comparison with Clip4Clip [20], which shows a comparable performance as Frozen (41.67 in R1@0.3, 37.50 in R1@0.5, 20.83 in R1@0.7). If given pretrained models with better performance, Our LPL will be further improved correspondingly.

Q2: Why adding knowledge from another dataset to solve the wVMR task resulting from the lack of dataset/annotation issue? From our perspective, precise temporal labels in the VMR task are cost-expensive and labor-intensive, while a bunch of video-text pretrained models without temporal information are recently proposed with popularity. Then, we consider how to utilize the knowledge from pretrained video-text models to relieve the annotation reliance on precise temporal labels in the wVMR task. Compared with directly using a pretrained video-text model to finetune on limited VMR datasets (in a fully-supervised setting), our proposed scheme can still work without extra temporal labels when dealing with a new VMR dataset.

Q3: Are there any other advantages of utilizing a pretrained model? Our proposed framework with a pretrained model is effective for video data in different domains. With much fewer learnable parameters, training our model will have a faster convergence speed and can be well adapted to small datasets.

Q4: Computation Complexity. Parameters of LPL model are summarized as follows: Model Parameters: 195M; GFlops: 960.015G; GPU memory: 1.3GB. Most VMR methods use pre-extracted features, but ours includes the calculation of feature extraction.

6 CONCLUSION

This paper proposes a prompt-based VMR model with Location-irrelevant Proposal Learning (LPL), which exploits a pretrained video-text model and prompt learning for better performance. Our LPL model includes reasonable proposal generation and prompt-based cross-modal feature fusion. Besides, we propose three different strategies of generating reasonable proposals with various lengths. Moreover, we propose two synthetic datasets with shuffled temporal location and longer video length to validate the robustness of proposed model to temporal localization bias and effectiveness of dealing with long video sequences. Extensive experiments on two public VMR datasets and synthetic datasets show the effectiveness of our proposed methods with consistency. Moving forward, we are going to: 1) extend our framework to other more general and challenging tasks, such as video corpus moment retrieval, *et al.*; 2) design more effective multi-modal prompt-based fusion structures or prompt-encoders, which can be easily adapted to other pretrained visual-language models.

7 ACKNOWLEDGEMENT

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).

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] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*. 1728–1738.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *NeurIPS* 33 (2020), 1877–1901.

[4] 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.

[5] Meng Cao, Tianyu Yang, Junwu Weng, Can Zhang, Jue Wang, and Yuexian Zou. 2022. Locvtp: Video-text pre-training for temporal localization. In *ECCV*. Springer, 38–56.

[6] Jianfeng Dong, Xianke Chen, Minsong Zhang, Xun Yang, Shujie Chen, Xirong Li, and Xun Wang. 2022. Partially Relevant Video Retrieval. In *ACM Multimedia*. 246–257.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

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

[9] 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.

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

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

[12] Dahye Kim, Jungin Park, Jiyoung Lee, Seongheon Park, and Kwanghoon Sohn. 2023. Language-free Training for Zero-shot Video Grounding. *WACV* (2023).

[13] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. In *CVPR*. 7331–7341.

[14] 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.

[15] Daizong Liu, Xiaoye Qu, Xing Di, Yu Cheng, Zichuan Xu, and Pan Zhou. 2022. Memory-Guided Semantic Learning Network for Temporal Sentence Grounding. *AAAI* (2022).

[16] 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.

[17] 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.

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

[19] 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.

[20] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2022. CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning. *Neurocomputing* 508 (2022), 293–304.

[21] 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.

[22] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*. 2630–2640.

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

[24] Jinwoo Nam, Daechul Ahn, Dongyeop Kang, Seong Jong Ha, and Jonghyun Choi. 2021. Zero-shot natural language video localization. In *ICCV*. 1470–1479.

[25] Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkilä. 2020. Uncovering hidden challenges in query-based video moment retrieval. *BMVC* (2020).

[26] T Prathiba and RSSP Kumari. 2021. Content based video retrieval system based on multimodal feature grouping by KFCM clustering algorithm to promote human–computer interaction. *Journal of Ambient Intelligence and Humanized Computing* 12, 6 (2021), 6215–6229.

[27] 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.

[28] 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).

[29] 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.

[30] Guolong Wang, Xun Wu, Zhaoyuan Liu, and Junchi Yan. 2022. Prompt-based Zero-shot Video Moment Retrieval. In *ACM Multimedia*. 413–421.

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

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

[33] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2022. Simvlm: Simple visual language model pretraining with weak supervision. *ICLR* (2022).

[34] 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.

[35] Junbin Xiao, Xindi Shang, Xun Yang, Sheng Tang, and Tat-Seng Chua. 2020. Visual relation grounding in videos. In *ECCV*. Springer, 447–464.

[36] 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.

[37] 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.

[38] Shuo Yang and Xinxiao Wu. 2022. Entity-aware and Motion-aware Transformers for Language-driven Action Localization in Videos. *IJCAI* (2022).

[39] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. 2021. Deconfounded Video Moment Retrieval with Causal Intervention. *SIGIR* (2021).

[40] Yuan Yao, Qianyu Chen, Ao Zhang, Wei Ji, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2022. PEVL: Position-enhanced Pre-training and Prompt Tuning for Vision-language Models. *EMNLP* (2022).

[41] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2021. Cpt: Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797* (2021).

[42] 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.

[43] 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. *ACL Findings* (2021).

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

[45] 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.

[46] 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.

[47] Minghang Zheng, Yanjie Huang, Qingchao Chen, and Yang Liu. 2022. Weakly supervised video moment localization with contrastive negative sample mining. In *AAAI*, Vol. 1. 3.

[48] Minghang Zheng, Yanjie Huang, Qingchao Chen, Yuxin Peng, and Yang Liu. 2022. Weakly Supervised Temporal Sentence Grounding With Gaussian-Based Contrastive Proposal Learning. In *CVPR*. 15555–15564.