Transferable Video Moment Localization by Moment-Guided Query Prompting

Hao Jiang, Yizhang Yang, Yadong Mu*

Wangxuan Institute of Computer Technology, Peking University
jianghao@stu.pku.edu.cn, myd@pku.edu.cn

Abstract

Video moment localization stands as a crucial task within the realm of computer vision, entailing the identification of temporal moments in untrimmed videos that bear semantic relevance to the supplied natural language queries. This work delves into a relatively unexplored facet of the task: the transferability of video moment localization models. This concern is addressed by evaluating moment localization models within a cross-domain transfer setting. In this setup, we curate multiple datasets distinguished by substantial domain gaps. The model undergoes training on one of these datasets, while validation and testing are executed using the remaining datasets. To confront the challenges inherent in this scenario, we draw inspiration from the recently introduced large-scale pre-trained vision-language models. Our focus is on exploring how the strategic utilization of these resources can bolster the capabilities of a model designed for video moment localization. Nevertheless, the distribution of language queries in video moment localization usually diverges from the text used by pre-trained models, exhibiting distinctions in aspects such as length, content, expression, and more. To mitigate the gap, this work proposes a Moment-Guided Query Prompting (MGQP) method for video moment localization. Our key idea is to generate multiple distinct and complementary prompt primitives through stratification of the original queries. Our approach is comprised of a prompt primitive constructor, a multimodal prompt refiner, and a holistic prompt incorporator. We carry out extensive experiments on Charades-STA, TACoS, DiDeMo, and YouCookII datasets, and investigate the efficacy of the proposed method using various pre-trained models, such as CLIP, ActionCLIP, CLIP4Clip, and VideoCLIP. The experimental results demonstrate the effectiveness of our proposed method.

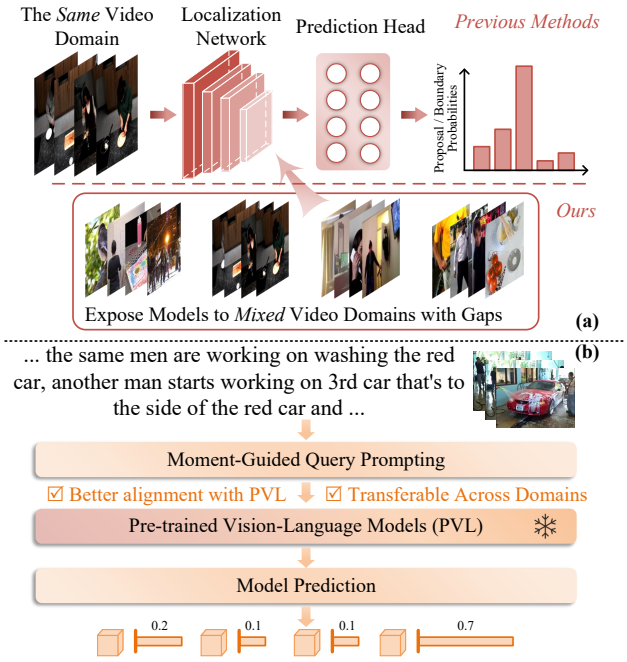

Figure 1: Schematic diagram of our main idea. (a): Previous studies typically train and test models within a single domain, making it difficult to understand how well models can transfer to new domains. (b): The distinction between text inputs used in pre-trained vision-language models and the sentence queries required for moment localization creates a gap. Our proposed method effectively diminishes this gap.

1 Introduction

In recent years, the realm of computer vision has undergone rapid progress in techniques for comprehending and analyzing video content. A particularly notable task, video moment localization (Anne Hendricks et al. 2017; Gao et al. 2017; Hu et al. 2021; Ma, Zhu, and Yang 2022; Shao et al. 2018; Yang et al. 2022; Zhang et al. 2020b), which centers on identifying temporal moments in videos based on natural language queries, has garnered escalating interest from both the academic and industrial sectors. Video moment localization holds a broad spectrum of applications, including video highlight detection (Liu et al. 2022), video summarization (Jiang and Mu 2022), etc.

Existing methods in this field can be broadly categorized into fully-supervised (Gao and Xu 2021; Wang et al. 2021; Zhang et al. 2021b; Zhao et al. 2021; Zhou et al. 2021), weakly-supervised (Huang et al. 2021; Mithun, Paul, and Roy-Chowdhury 2019; Yang et al. 2021; Zheng et al. 2022), or unsupervised (Nam et al. 2021) approaches, based on

*Corresponding Author.

the manner in which temporal boundaries of moments are annotated. Over recent years, substantial efforts have been dedicated to modeling visual features (Wang, Huang, and Wang 2019; Zhang et al. 2021b), textual features (Ding et al. 2022; Gao and Xu 2021), and cross-modal feature interactions (Chen and Jiang 2020; Mun, Cho, and Han 2020; Zeng et al. 2020). However, despite the notable strides made by these studies, a crucial question remains unexplored: Can video moment localization models achieve satisfactory transferability? Most prior research evaluates moment localization models solely within their original domain, which diverges from practical application scenarios and restricts an analysis of model transferability.

To tackle these concerns, we suggest investigating video moment localization models in cross-domain transfer scenarios. In this framework, we curate multiple datasets with relatively substantial domain gaps, specifically Charades (Daily Indoor Videos), TACoS (Cooking), DiDeMo (Open Domains), and YouCookII (Instructional Videos). We proceed by training the model on one of these datasets and subsequently carry out validation and testing on the remaining datasets, as depicted in Figure 1(a). In a previously related work (Li et al. 2022a), the focus was on exploring the word compositional generalization ability of moment localization models. Differing from this approach, in our cross-domain transfer setting, the domain gap doesn't solely emerge from the text domain, but also encompasses the visual domain. This aligns more closely with the practical contexts of moment localization applications.

To tackle the challenges posed by this setting, drawing inspiration from the recently introduced large-scale pre-trained vision-language models, we delve into how these pre-trained resources can enhance a video moment localization model. Intuitively, models like CLIP (Radford et al. 2021) encapsulate tremendous knowledge about the correlation between vision and language, thereby aiding in generalizing to new scenes encountered during testing. Nonetheless, this solution encounters certain challenges. As depicted in Figure 1(b), in video moment localization tasks, queries commonly take the form of natural language sentences describing temporal events. This gives rise to disparities in length, content, and expression style when compared to the text employed for optimizing pre-trained vision language models. Neglecting this distinction could potentially impede the performance of moment localization models.

To address this problem, we propose a moment-guided query prompting (MGQP) method, as showcased in Figure 2. It consists of three key modules: *prompt primitive constructor*, *multimodal prompt refiner*, and *holistic prompt incorporator*. Specifically, the prompt primitive constructor sifts through the querying sentences by a set of predefined primitives. In this manner, we achieve not only the decomposition of the original sentences into elementary prompts, but also harness these fundamental prompt primitives to enhance invariance and transferability across diverse domains. Due to the predominant emphasis on textual queries in prompt primitives, with insufficient attention to visual content, we introduce a multimodal prompt refiner. The objective of this module is to integrate visual information into the prompt primitives, thereby creating prompts that are attuned to visual cues. Lastly, recognizing that the pre-designed prompt primitives might not comprehensively encompass the diverse semantics of the queries, a holistic prompt incorporator is presented. This module encodes the content not covered by the aforementioned prompts via employing some learnable global operations.

The main contributions are summarized as follows:

- This work explores the cross-domain transfer setting of video moment localization models, mirroring the domain gap observed in real-world scenarios. It assesses the model's ability to transfer to unfamiliar contexts and scenes.
- We propose a moment-guided query prompting method, which efficiently harnesses the support of pre-trained vision-language models to amplify the efficacy of moment localization methods.
- We extensively experiment with Charades-STA, TACoS, DiDeMo, and YouCookII datasets. Additionally, we examine the effectiveness of the proposed method across diverse pre-trained models, including CLIP, ActionCLIP, CLIP4Clip, and VideoCLIP. The results demonstrate the superiority of the proposed method over baseline methods. We release the code of this work on this website[1].

2 Related Work

Video moment localization. The related methods can be categorized into two groups: proposal-free methods (Ghosh et al. 2019; Zeng et al. 2020) and proposal-based methods (Xu et al. 2019; Yuan et al. 2019; Zhang et al. 2019; Ning et al. 2021; Kim et al. 2023; Wang et al. 2022a; Xiao et al. 2021; Liu et al. 2023). For example, Zhang *et al.* (Zhang et al. 2023) co-train vision encoder and language encoder in a 2D proposal-free manner with prompts. Nonetheless, the majority of current methods concentrate on formulating expressive feature learning and interaction modules, often overlooking the investigation into the transferability of models. In contrast, our approach investigates the challenges of cross-domain transfer settings, dedicating efforts to amplify the model's transfer capabilities.

Prompt learning. It is originally proposed in the field of natural language processing, which bridges the pre-trained language models and a plethora of downstream tasks in a lightweight and efficient way (Zhang et al. 2022; Guo, Yang, and Abbasi 2022; Hu et al. 2022). Recently, profitting from the development of large-scale pre-trained vision-language models, prompt learning methods have been increasingly applied to the field of computer vision (Li et al. 2022b; Rao et al. 2022; Lu et al. 2022; Wang et al. 2022c; Du et al. 2022). In this paper, we embark on an exploration of video moment-oriented prompt construction, with a consequential focus on fortifying the capability of video moment localization.

Domain generalization. For domain generalization, models are trained on one or more given domains and tested

[1]https://code-website.wixsite.com/prompt-code.

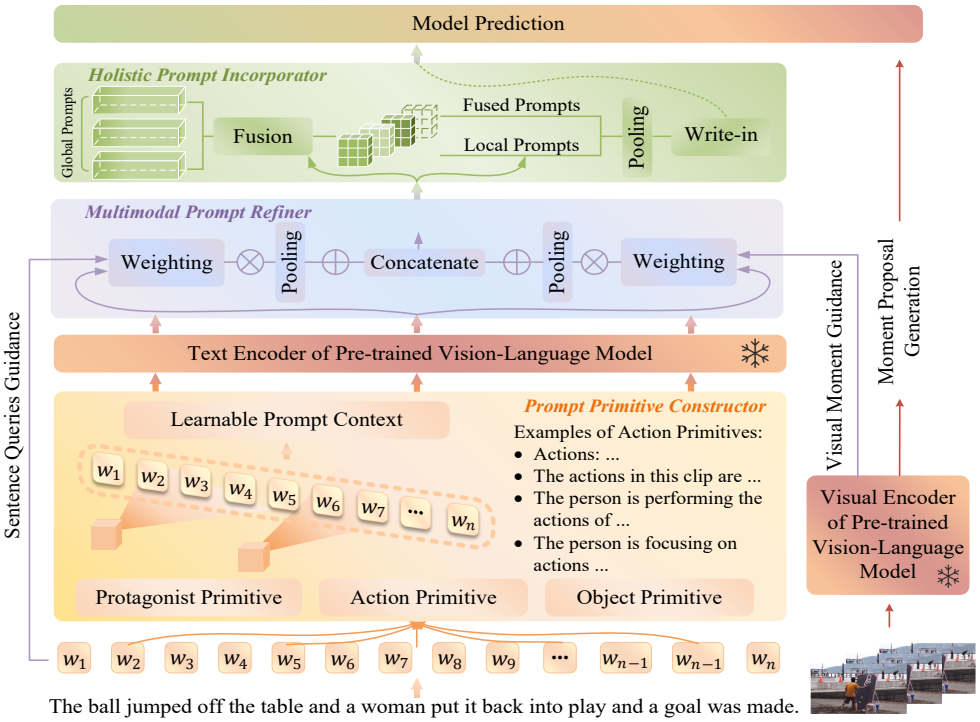

Figure 2: Pipeline of the proposed moment-guided query prompting method for video moment localization. It consists of prompt primitive constructor, multimodal prompt refiner, and holistic prompt incorporator.

on unseen domains. Prior work can be divided into three categories: data manipulation (Yue et al. 2019; Robey, Pappas, and Hassani 2021), representation learning (Mahajan, Tople, and Sharma 2021; Choi et al. 2021), and learning strategies design (Tian et al. 2022; Kim et al. 2021). In this paper, we study the cross-domain transfer problem in video moment localization, which is of great importance in practical applications and is rarely explored in previous work.

3 The Proposed Method

3.1 Overview

In this section, we present the architecture of the proposed moment-guided query prompting method for video moment localization. As depicted in Figure 2, the proposed method comprises three modules. These modules facilitate query sentence understanding by acquiring multiple interpretable primitives and generating visually-aware prompts through multimodal prompt refinement. The holistic prompt incorporator further enhances prompt modeling by incorporating a global perspective. Subsequent sections provide a detailed description of these modules.

3.2 Prompt Primitive Constructor

As previously mentioned, our approach involves harnessing the knowledge from large-scale pre-trained vision-language models to enhance the effectiveness of video moment localization models. Nevertheless, an inevitable challenge arises from the disparities, such as variations in length, content, and expression, between the text used during the pre-training of the vision-language model and the sentence queries encountered in the moment localization task. Employing sentence queries directly as prompts may not represent the most effective approach. To address this problem, this section introduces a prompt primitive constructor, which learns a series of accessible prompt primitives aimed at encapsulating distinct facets of sentence queries and mitigates the divergence in text modalities between domains. Our observaiton centers around that, despite variations in text queries across different domains, they share common query primitives, as elaborated below. By unifying the representation of texts into a common primitive framework, we further enhance the model's transferring capacity across diverse distributions.

Formally, for the k-th moment-query pair $< \boldsymbol{v}_k, \boldsymbol{q}_k >$ sampled from the video-query sets $\mathcal{V}, \mathcal{Q}$, we denote the natural language query as $\boldsymbol{q}_k = \{\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_n\}$. Here, $\boldsymbol{w}_i (1 \leq i \leq n)$ signifies a word in the sentence query $\boldsymbol{q}_k$, and n stands for the length of the sentence. To derive prompts from sentence queries, we establish a collection of prompt primitives denoted as $\mathcal{P} = \{\boldsymbol{p}_{pro}, \boldsymbol{p}_{act}, \boldsymbol{p}_{obj}\}$, in which $\boldsymbol{p}_{pro}, \boldsymbol{p}_{act}, \boldsymbol{p}_{obj}$ correspond to the protagonist prompt, action prompt, and object prompt, respectively. Protagonist prompt takes the form of "The protagonist in this clip is {_}"; action prompt is structured as "The actions in this clip are {_}"; object prompt follows the format of "The objects in this clip are {_}". In cases where the query lacks a verb or noun, the related action or object primitive becomes a rudimentary prompt. This implies that the blank

remains unfilled, and solely the primitive's external features surrounding the blank are employed as prompts.

We populate the prompt primitives $\mathcal{P}$ using a simple approach. Recognizing that noun chunks (verb chunks) within query sentences can convey more comprehensive semantic information compared to isolated nouns (verbs), such as in the case of "a boy wearing a blue T-shirt" containing more context than the noun "boy", we initially establish a series of sentence-chunking parsers. The sentence chunking parser serves the purpose of extracting relevant phrases from sentences. We empirically incorporate standard rules for noun phrases and verb phrases into the sentence chunking parser. For $\boldsymbol{p}_{act}$ and $\boldsymbol{p}_{obj}$, we initiate the process by identifying verb chunks and noun chunks within $\boldsymbol{q}_k$, denoting them as $\boldsymbol{q}_k^{vphr} = \{\boldsymbol{w}_{n_1}^{vphr}, \boldsymbol{w}_{n_2}^{vphr}, ..., \boldsymbol{w}_{n_{vphr}}^{vphr}\}$ and $\boldsymbol{q}_k^{nphr} = \{\boldsymbol{w}_{n_1}^{nphr}, \boldsymbol{w}_{n_2}^{nphr}, ..., \boldsymbol{w}_{n_{nphr}}^{nphr}\}$, respectively. Subsequently, we employ $\boldsymbol{q}_k^{vphr}$ and $\boldsymbol{q}_k^{nphr}$ to populate the placeholders within $\boldsymbol{p}_{act}$ and $\boldsymbol{p}_{obj}$, resulting in action prompts and object prompts. Regarding protagonist prompts, we adopt a simple approach of identifying sentence chunks associated with pre-defined common protagonists (*e.g.*, the man, the child), and then substituting them into $\boldsymbol{p}_{pro}$. In cases where the sentence commences with a personal pronoun, we substitute the pronoun with the relevant chunk (*e.g.*, replace "he" with "the man", "they" with "the people").

Prior studies (Zhou et al. 2022a,b) have suggested that incorporating learnable prefixes or suffixes can enhance prompt flexibility and generalization. Drawing inspiration from this, we introduce low-dimensional learnable vectors as contextual cues for $\boldsymbol{p}_{pro}$, $\boldsymbol{p}_{act}$, and $\boldsymbol{p}_{obj}$, enriching the information associated with prompt keywords. The resulting prompts are denoted as $\overline{\mathcal{P}} = \{\overline{\boldsymbol{p}}_{pro}, \overline{\boldsymbol{p}}_{act}, \overline{\boldsymbol{p}}_{obj}\}$. We introduce learnable context for $\boldsymbol{p}_{pro}$, $\boldsymbol{p}_{act}$, and $\boldsymbol{p}_{obj}$ rather than for $\boldsymbol{q}_k^{vphr}$ and $\boldsymbol{q}_k^{nphr}$ due to the fact that $\boldsymbol{p}_{pro}$, $\boldsymbol{p}_{act}$, and $\boldsymbol{p}_{obj}$ encompass essential paraphrases of prompts (such as "protagonist," "actions," "objects").

3.3 Multimodal Prompt Refiner

Considering the acquired prompt $\mathcal{P}$, it is essential to note that the existing information within $\mathcal{P}$ originates exclusively from text queries, neglecting the inclusion of visual information. Visual information assumes a critical role in enhancing prompt perception of video moments and facilitating the capture of multimodal feature fusion. Hence, the integration of visual cues into $\mathcal{P}$ becomes imperative, as it serves to offer complementary modality information and amplify overall efficacy. Taking this into consideration, we introduce a multimodal prompt refiner that enhances prompts by integrating guidance from visual moments and sentence queries. It processes prompt primitive features $\{\overline{\boldsymbol{p}}_{pro}, \overline{\boldsymbol{p}}_{act}, \overline{\boldsymbol{p}}_{obj}\}$ as its input and produces refined multimodal prompt representations $\{\widetilde{\boldsymbol{p}}_{pro}, \widetilde{\boldsymbol{p}}_{act}, \widetilde{\boldsymbol{p}}_{obj}\}$.

Formally, for each $\overline{\boldsymbol{p}}_j \in \{\overline{\boldsymbol{p}}_{pro}, \overline{\boldsymbol{p}}_{act}, \overline{\boldsymbol{p}}_{obj}\}$, we commence by calculating weighted scores between $\overline{\boldsymbol{p}}_j$ and either the visual moment $\boldsymbol{v}_k$ or the sentence query $\boldsymbol{q}_k$. We leverage a pre-trained text encoder $\mathbf{T}$ (frozen) to represent $\boldsymbol{q}_k$, and a frozen visual encoder $\mathbf{I}$ is employed to extract features for

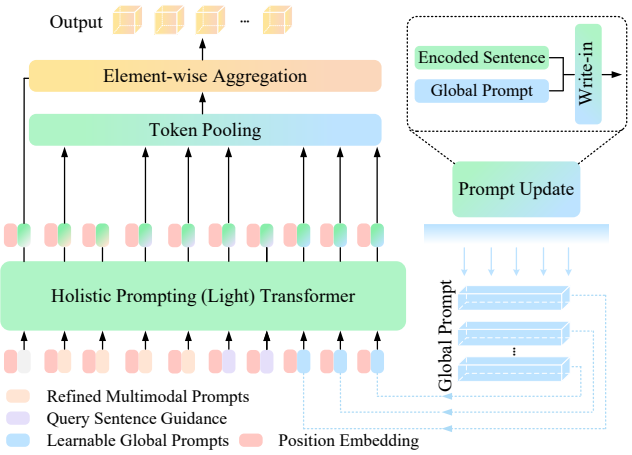

Figure 3: Pipeline of holistic prompt incorporator.

$\boldsymbol{v}_k$. The computation of weighted scores unfolds in the subsequent manner:

$$\begin{cases} \boldsymbol{w}_{j\leftarrow q} = \overline{\boldsymbol{p}}_j \cdot \big(\mathbf{T}(\boldsymbol{q}_k)\big)^\top, \\ \boldsymbol{w}_{j\leftarrow v} = \overline{\boldsymbol{p}}_j \cdot \big(\mathbf{I}(\boldsymbol{v}_k)\big)^\top, \end{cases} \tag{1}$$

where $\boldsymbol{w}_{j\leftarrow q}$ signifies the weighted score between prompt $\overline{\boldsymbol{p}}_j$ and text query $\boldsymbol{q}_k$, whereas $\boldsymbol{w}_{j\leftarrow v}$ represents the weighted score between prompt $\overline{\boldsymbol{p}}_j$ and the visual moment $\boldsymbol{v}_k$. Subsequently, a weighted summation operation is executed to consolidate information from various prompts:

$$\begin{cases} \boldsymbol{p}_q^{(w)} = \sum_{j=1}^{|\mathcal{P}|} \boldsymbol{w}_{j\leftarrow q} \cdot \overline{\boldsymbol{p}}_j, \\ \boldsymbol{p}_v^{(w)} = \sum_{j=1}^{|\mathcal{P}|} \boldsymbol{w}_{j\leftarrow v} \cdot \overline{\boldsymbol{p}}_j, \end{cases} \tag{2}$$

where $\boldsymbol{p}_q^{(w)}$ and $\boldsymbol{p}_v^{(w)}$ symbolize the prompt features guided by sentence queries and moments, respectively. Afterwards, residual connections are applied to $\boldsymbol{p}_q^{(w)}$ and $\boldsymbol{p}_v^{(w)}$, resulting in the attainment of enhanced prompt representations $\widetilde{\boldsymbol{p}}_j$:

$$\begin{cases} \boldsymbol{p}_{j\leftarrow q}^{(r)} = \overline{\boldsymbol{p}}_j \oplus \boldsymbol{p}_q^{(w)}, \boldsymbol{p}_{j\leftarrow v}^{(r)} = \overline{\boldsymbol{p}}_j \oplus \boldsymbol{p}_v^{(w)}, \\ \widetilde{\boldsymbol{p}}_j = \boldsymbol{p}_{j\leftarrow q}^{(r)} \| \boldsymbol{p}_{j\leftarrow v}^{(r)}, \end{cases} \tag{3}$$

where $\|$ signifies the concatenation operation, and $\oplus$ stands for element-wise addition.

3.4 Holistic Prompt Incorporator

The prompts established in $\mathcal{P}$ predominantly focus on three key aspects of sentence queries: protagonist, action, and object. Nevertheless, it's worth highlighting that several other facets of sentences remain unaddressed, including the background in which actions unfold, the temporal correlation between actions, the causal linkages within a sentence, and more. To facilitate a more comprehensive comprehension of prompts, a holistic prompt incorporator is curated. This module offers a broader perspective and conveys the content that is not covered by the constructed prompts. Figure 3 depicts the presented module, and the detailed architecture

of the holistic prompting transformer is elaborated upon in the implementation details section. Our main idea revolves around optimizing a set of adaptable global prompts that gather insights from local prompts and adjust them accordingly. The incorporation of trainable global prompts offers several benefits. It broadens the prompt's purview, enhancing its comprehensiveness, while also facilitating the interaction of features across diverse prompts.

Formally, we represent a set of learnable global prompts as $\mathcal{G} = \{\boldsymbol{g}_1, \boldsymbol{g}_2, ..., \boldsymbol{g}_z\}$, where z denotes the number of global prompts. $\widetilde{\boldsymbol{p}}_j (j \in \{pro, act, obj\})$, $\mathbf{T}(\boldsymbol{q}_k)$, $\mathcal{G}$ are fed into the transformer encoder, fostering interactions among global prompt features, local prompt features, and original sentence features:

$$\widetilde{\boldsymbol{p}}^{(o)}, \boldsymbol{q}^{(o)}, \boldsymbol{g}^{(o)} = \text{HolisticTransformer}(\widetilde{\boldsymbol{p}}, \mathbf{T}(\boldsymbol{q}_k), \mathcal{G}), \quad (4)$$

where $\widetilde{\boldsymbol{p}}$ stands for the collective set of $\widetilde{\boldsymbol{p}}_j$, while $\widetilde{\boldsymbol{p}}^{(o)}$, $\boldsymbol{q}^{(o)}$, and $\boldsymbol{g}^{(o)}$ denote the resultant vectors from the local prompt, sentence, and global prompt, respectively. This approach empowers the module to strengthen the learnable global prompt representation $\mathcal{G}$ by assimilating information from the local prompt $\widetilde{\boldsymbol{p}}_j$, ensuring that the local prompt information is captured within the global prompts. Furthermore, the global prompt undergoes iterative updates through interactions with sentence features $\mathbf{T}(\boldsymbol{q}_k)$, facilitating the acquisition of supplementary content that might not be explicitly encompassed within the local prompts. Subsequently, we concatenate the [CNT] token with prompt features produced by the transformer, resulting in the ultimate outcome:

$$\boldsymbol{o} = [\text{CNT Token}] \big\| \text{Pooling}\big[\widetilde{\boldsymbol{p}}^{(o)} \big\| \boldsymbol{q}^{(o)}\big]. \quad (5)$$

For optimization, we adhere to proposal-based methods such as (Wang et al. 2022b) and (Zhang et al. 2020a). The objective remains consistent with (Wang et al. 2022b).

4 Experiments

4.1 Datasets and Evaluation Metrics

TACoS contains 127 videos of kitchen scenes (Regneri et al. 2013) and $18,818$ video-language pairs. We follow the segmentation by (Gao et al. 2017), with 10,146, 4,589, 4,083 video-query pairs in training, validation, and test sets.

Charades-STA contains 9,848 videos of daily indoor activities (Sigurdsson et al. 2016). Training and test sets contain 12,408 and 3,720 video-query pairs, respectively.

DiDeMo contains 10,464 Flickr videos and 40,543 annotated queries (Anne Hendricks et al. 2017). Each annotated segment is 5 seconds long.

YouCookII is an instructional video dataset collected by (Zhou, Xu, and Corso 2018), comprising 2,000 long untrimmed videos from YouTube. Each video encompasses procedural steps annotated with temporal boundaries and elucidated through imperative sentences.

Method	Rank@1, IoU=			Rank@5, IoU=		
	0.3	0.5	0.7	0.3	0.5	0.7
2D-TAN-Pool	30.16	18.09	5.46	75.40	58.17	28.33
2D-TAN-Conv	31.40	14.95	6.29	76.96	61.37	29.25
VSLNet-RNN	30.81	14.16	5.21	–	–	–
VSLNet-Transf	32.11	19.22	7.06	–	–	–
EAMAT	22.82	10.40	3.31	–	–	–
MMN	36.21	20.16	8.15	80.73	64.17	32.58
MMN-Feature	36.26	21.26	9.25	84.09	68.31	34.44
CoOp	36.53	22.46	9.58	88.04	68.28	35.05
CoCoOp	30.75	17.70	7.29	89.03	70.17	35.25
MGQP (Ours)	**39.36**	**23.96**	**10.39**	**89.82**	**71.30**	**38.14**

Table 1: Comparison of the performance of our proposed method and baseline models on TACoS → Charades.

Method	Rank@1		Rank@5	
	IoU=0.3	IoU=0.5	IoU=0.3	IoU=0.5
2D-TAN-Pool	7.19	1.69	26.51	8.63
2D-TAN-Conv	5.35	1.34	27.32	10.66
VSLNet-RNN	3.81	1.00	–	–
VSLNet-Transf	3.38	0.88	–	–
EAMAT	5.88	2.13	–	–
MMN	7.81	2.78	28.35	11.91
MMN-Feature	7.88	2.60	28.96	13.02
CoOp	8.13	3.24	32.22	14.64
CoCoOp	7.44	2.67	30.43	13.70
MGQP (Ours)	**9.20**	**3.56**	**33.45**	**15.04**

Table 2: Comparison of the performance of our proposed method and baseline models on TACoS → YouCookII.

Evaluation Protocals. We experiment on TACoS → Charades, TACoS → YouCookII, Charades → TACoS, and DiDeMo → YouCookII. We use the former dataset for training and the latter dataset for validation and testing. We evaluate $Rank@n, IoU = m$ as metrics. It represents the percentage of queries for which at least one proposal in top n predictions satisfies "IoU with ground truth greater than m".

4.2 Implementation Details

The pretrained vision-language models used in our experiments encompass CLIP (Radford et al. 2021), ActionCLIP (Wang, Xing, and Liu 2021), VideoCLIP (Xu et al. 2021), and CLIP4Clip (Luo et al. 2022). Their parameters are frozen in the experiments. The number of global prompts z is set to 4. The number of sampled video clips used in the model is 32, and the size of the 2D moment map is 16, which is consistent with the baseline model in the experiment. The hidden dimension of the model is 512, and the number of layers in the transformer block is 6. The number of heads in the multi-head attention layer is 8. Other parameter settings (*e.g.*, non-maximum suppression threshold, scaling thresholds) are consistent with baseline methods (Wang et al. 2022b; Zhang et al. 2020a). AdamW optimizer (Loshchilov and Hutter 2018) is adopted in the experiment.

Method	Rank@1, IoU=			Rank@5, IoU=		
	0.1	0.3	0.5	0.1	0.3	0.5
2D-TAN-Pool	25.92	6.62	1.60	64.31	29.17	11.47
2D-TAN-Conv	26.17	6.02	1.07	68.61	27.69	11.67
VSLNet-RNN	26.47	6.90	1.32	–	–	–
VSLNet-Transf	22.19	6.35	1.67	–	–	–
EAMAT	24.24	5.80	1.70	–	–	–
MMN	27.62	7.97	1.45	69.73	30.87	11.75
MMN-Feature	27.07	7.30	1.62	71.13	31.12	11.92
CoOp	28.65	7.90	1.88	73.04	33.79	13.43
CoCoOp	28.90	6.96	1.14	**73.80**	32.37	12.42
MGQP (Ours)	**29.56**	**10.24**	**3.54**	71.71	**33.94**	**14.97**

Table 3: Comparison of the performance of our proposed method and baseline models on Charades → TACoS.

Method	Rank@1		Rank@5	
	IoU=0.3	IoU=0.5	IoU=0.3	IoU=0.5
2D-TAN-Pool	5.31	1.56	21.01	6.88
2D-TAN-Conv	6.35	1.97	28.01	9.72
VSLNet-RNN	5.53	1.75	–	–
VSLNet-Transf	6.10	1.81	–	–
EAMAT	6.03	1.69	–	–
MMN	6.66	1.97	30.42	10.41
MMN-Feature	7.85	2.38	32.95	11.57
CoOp	7.75	2.47	33.95	12.88
CoCoOp	7.94	2.22	33.89	12.57
MGQP (Ours)	**8.51**	**3.01**	**34.74**	**13.69**

Table 4: Comparison of the performance of our proposed method and baseline models on DiDeMo → YouCookII.

4.3 Experimental Results

The experimental results on TACoS → Charades, TACoS → YouCookII, Charades → TACoS, and DiDeMo → YouCookII are shown in Tables 1∼4. Several baseline methods (2D-TAN-Pool, 2D-TAN-Conv (Zhang et al. 2020a), VSLNet-RNN, VSLNet-Transf (Zhang et al. 2021a), EAMAT (Yang and Wu 2022), MMN (Wang et al. 2022b)) leverage VGG, C3D, and I3D features for DiDeMo, TACoS, and Charades, respectively. On the other hand, MMN-Feature (Wang et al. 2022b), CoOp (Zhou et al. 2022b), and CoCoOp (Zhou et al. 2022a) employ pre-trained vision-language models (*i.e.*, ActionCLIP) to extract visual/textual features. The experimental results show that the baseline models, including 2D-TAN, VSLNet, and EAMAT, does not achieve satisfactory results, possibly indicating their limited exploration of model design in the context of transfer scenarios, resulting in suboptimal performance. For MMN-Feature baseline, we solely substituted the input features with features extracted by the pre-trained models. However, this replacement does not yield substantial performance enhancement. This outcome demonstrates that simply relying on features extracted from pre-trained models does not yield satisfactory results. In addition, we also conduct comparisons against previous pre-trained model prompting methods, such as CoOp and CoCoOp. The experiments demonstrate that MGQP outperforms these alternatives, potentially due to the fact that the baselines rely mostly on straightforward prompting strategies, which exhibit suboptimal performance in experimental settings.

Model Variants	Rank@5		
	IoU=0.3	IoU=0.5	IoU=0.7
MGQP-P	87.37	69.82	37.40
MGQP-L	87.68	69.41	35.95
MGQP-G	89.38	71.08	37.09
MGQP (Ours)	**89.82**	**71.30**	**38.14**

Table 5: Ablation experiment results. We compare the performance of three different model variants, MGQP-P, MGQP-L and MGQP-G.

Pre-trained Vision-Language Model: VideoCLIP (Xu et al. 2021)						
TACoS→Charades	Rank@1, IoU=			Rank@5, IoU=		
	0.3	0.5	0.7	0.3	0.5	0.7
Base Method	36.12	18.05	7.04	86.69	70.89	32.80
MGQP	**39.04**	**23.72**	**10.75**	**93.01**	**73.85**	**38.07**
Pre-trained Vision-Language Model: CLIP (Radford et al. 2021)						
TACoS→Charades	Rank@1, IoU=			Rank@5, IoU=		
	0.3	0.5	0.7	0.3	0.5	0.7
Base Method	32.45	18.25	7.39	86.29	66.91	34.68
MGQP	**36.10**	**19.89**	**7.72**	**90.26**	**70.05**	**37.94**

Table 6: Results comparison of the proposed method for VideoCLIP (Xu et al. 2021) and CLIP (Radford et al. 2021).

4.4 Ablation Study

To assess the effectiveness of each proposed module, we conduct ablation experiments, as presented in Table 5. In MGQP-P, we omit the proposed prompt primitive. For this variant, the model does not differentiate between the protagonist, action, and object in a sentence, blending them as prompt input to the model. The experimental results demonstrate a degradation in the performance of MGQP-P, underscoring the necessity of constructing prompt primitives. In MGQP-L, we exclude the multimodal prompt refiner, resulting in a model variant that does not integrate multimodal feature information from sentence queries and visual moments. The experimental results indicate that MGQP-L achieves lower performance than MGQP, demonstrating the effectiveness of the multimodal prompt refiner. In the MGQP-G variant, we remove the holistic prompt incorporator. This results in a model lacking learnable global prompts to complement the information provided by the constructed prompt primitives. As observed in the experimental results, this variant achieves suboptimal performance, indicating the effectiveness of our model design.

4.5 Study on General Vision-Language Models

To substantiate the efficacy of the proposed MGQP mechanism, we extend our evaluation beyond ActionCLIP (Wang, Xing, and Liu 2021) to encompass a broader array of

MSVD (Chen and Dolan 2011)	Rank@1, IoU= 0.3	0.5	0.7	Rank@5, IoU= 0.3	0.5	0.7
Base Method	33.39	22.66	8.76	82.98	66.77	31.83
MGQP	**34.77**	**23.68**	**10.90**	**86.91**	**71.30**	**35.66**
MSR-VTT (Xu et al. 2016)	Rank@1, IoU= 0.3	0.5	0.7	Rank@5, IoU= 0.3	0.5	0.7
Base Method	35.28	21.16	8.44	81.29	67.82	33.44
MGQP	**37.40**	**25.52**	**11.54**	**85.85**	**71.32**	**37.96**
WebVid (Bain et al. 2021)	Rank@1, IoU= 0.3	0.5	0.7	Rank@5, IoU= 0.3	0.5	0.7
Base Method	33.76	23.57	11.09	85.75	70.13	35.43
MGQP	**35.68**	**24.50**	**12.40**	**89.17**	**74.82**	**38.04**

Table 7: Comparison of experimental results of the proposed method for CLIP4Clip (Luo et al. 2022). CLIP4Clip is pre-trained on MSVD, MSR-VTT, and WebVid dataset respectively.

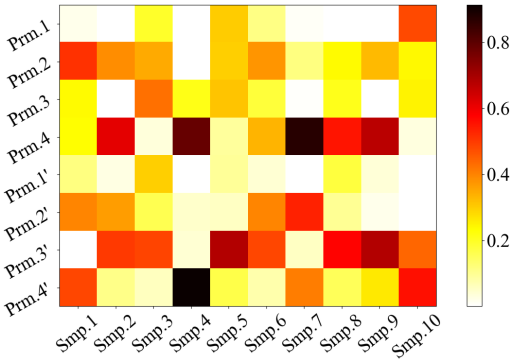

Figure 4: Visualization of visual moment guidance in multi-modal prompt refiner.

pre-trained vision-language models, including CLIP (Radford et al. 2021), VideoCLIP (Xu et al. 2021), and CLIP4Clip (Luo et al. 2022). In Table 6, we present the results of the proposed method on VideoCLIP and CLIP, with experiments conducted on the TACoS→Charades dataset. Here, we opt for MMN-Feature as the base method. Table 7 showcases the performance of MGQP when utilizing CLIP4Clip as the backbone network. We explore various pre-training datasets, including MSVD (Chen and Dolan 2011), MSR-VTT (Xu et al. 2016), and WebVid (Bain et al. 2021). The experimental results indicate the effectiveness of our proposed method across different pre-trained models.

4.6 Visualization of Visual Moment Guidance

Figure 4 illustrates the outcomes of visual moment attentions in the multimodal prompt refiner, depicting the weight matrix assigned to distinct prompts in alignment with visual guidance. Along the horizontal axis, 10 samples are randomly chosen, while the vertical axis corresponds to the protagonist, action, object, and sentence prompts. Evidently, varying prompts assume varying degrees of significance contingent on the specific visual moment, where the

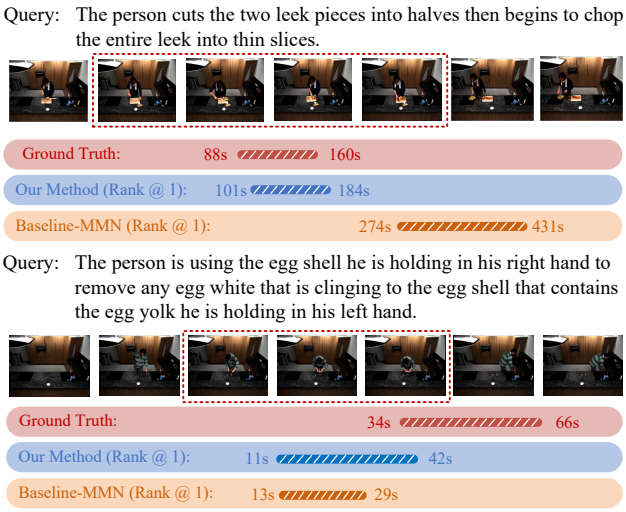

Figure 5: Illustration of qualitative experimental results. Videos s28-d39 (upper) and s27-d50 (lower) are selected from TACoS and we visualize Rank@1 model predictions.

object and sentence prompts tend to hold more substantial roles.

4.7 Qualitative Experiment and Failure Case

Visualizations of the predictions generated by MGQP and key baseline (Wang et al. 2022b) are provided in Figure 5. As evident from the results in the upper sub-figure, our model exhibits a greater capacity to produce accurate predictions compared to the baseline. In the lower sub-figure, we present a failure case of the model's prediction. In this example, the query sentence features intricate relationships, including phrases like "egg white that is clinging to the egg shell that contains the egg yolk he is holding in his left hand". Given the relatively simplistic nature of the constructed prompts, the model's prediction accuracy might be constrained when confronted with queries that repeatedly feature complex relationships between objects. A potential solution is to incorporate additional prompts that can capture nuanced relationships within sentences, including temporal relationships, causal connections, etc.

5 Conclusion

We investigate query sentence prompting methods to transfer rich knowledge from large-scale pre-trained vision-language models to the domain of video moment localization. We propose MGQP, which facilitates sentence content understanding by acquiring a set of prompts, and realizes cross-prompt complementary modeling through local- and global-level interaction mechanisms. Quantitative and qualitative experiments verify the effectiveness of the proposed method.

Acknowledgements

The research is supported by National Key R&D Program of China (2022ZD0160305).

References

Anne Hendricks, L.; Wang, O.; Shechtman, E.; Sivic, J.; Darrell, T.; and Russell, B. 2017. Localizing moments in video with natural language. In *ICCV*, 5803–5812.

Bain, M.; Nagrani, A.; Varol, G.; and Zisserman, A. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 1728–1738.

Chen, D.; and Dolan, W. B. 2011. Collecting highly parallel data for paraphrase evaluation. In *ACL*, 190–200.

Chen, S.; and Jiang, Y.-G. 2020. Hierarchical visual-textual graph for temporal activity localization via language. In *ECCV*, 601–618.

Choi, S.; Jung, S.; Yun, H.; Kim, J. T.; Kim, S.; and Choo, J. 2021. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *CVPR*, 11580–11590.

Ding, X.; Wang, N.; Zhang, S.; Huang, Z.; Li, X.; Tang, M.; Liu, T.; and Gao, X. 2022. Exploring language hierarchy for video grounding. *TIP*, 4693–4706.

Du, Y.; Wei, F.; Zhang, Z.; Shi, M.; Gao, Y.; and Li, G. 2022. Learning to Prompt for Open-Vocabulary Object Detection with Vision-Language Model. In *CVPR*, 14084–14093.

Gao, J.; Sun, C.; Yang, Z.; and Nevatia, R. 2017. Tall: Temporal activity localization via language query. In *ICCV*, 5267–5275.

Gao, J.; and Xu, C. 2021. Fast video moment retrieval. In *ICCV*, 1523–1532.

Ghosh, S.; Agarwal, A.; Parekh, Z.; and Hauptmann, A. G. 2019. ExCL: Extractive Clip Localization Using Natural Language Descriptions. In *ACL*, 1984–1990.

Guo, Y.; Yang, Y.; and Abbasi, A. 2022. Auto-Debias: Debiasing Masked Language Models with Automated Biased Prompts. In *ACL*, 1012–1023.

Hu, S.; Ding, N.; Wang, H.; Liu, Z.; Wang, J.; Li, J.; Wu, W.; and Sun, M. 2022. Knowledgeable Prompt-tuning: Incorporating Knowledge into Prompt Verbalizer for Text Classification. In *ACL*, 2225–2240.

Hu, Y.; Liu, M.; Su, X.; Gao, Z.; and Nie, L. 2021. Video moment localization via deep cross-modal hashing. *TIP*, 4667–4677.

Huang, J.; Liu, Y.; Gong, S.; and Jin, H. 2021. Cross-sentence temporal and semantic relations in video activity localisation. In *ICCV*, 7199–7208.

Jiang, H.; and Mu, Y. 2022. Joint Video Summarization and Moment Localization by Cross-Task Sample Transfer. In *CVPR*, 16388–16398.

Kim, D.; Park, J.; Lee, J.; Park, S.; and Sohn, K. 2023. Language-free Training for Zero-shot Video Grounding. In *WACV*, 2539–2548.

Kim, D.; Yoo, Y.; Park, S.; Kim, J.; and Lee, J. 2021. Selfreg: Self-supervised contrastive regularization for domain generalization. In *ICCV*, 9619–9628.

Li, J.; Xie, J.; Qian, L.; Zhu, L.; Tang, S.; Wu, F.; Yang, Y.; Zhuang, Y.; and Wang, X. E. 2022a. Compositional temporal grounding with structured variational cross-graph correspondence learning. In *CVPR*, 3032–3041.

Li, M.; Chen, L.; Duan, Y.; Hu, Z.; Feng, J.; Zhou, J.; and Lu, J. 2022b. Bridge-Prompt: Towards Ordinal Action Understanding in Instructional Videos. In *CVPR*, 19880–19889.

Liu, D.; Fang, X.; Zhou, P.; Di, X.; Lu, W.; and Cheng, Y. 2023. Hypotheses tree building for one-shot temporal sentence localization. In *AAAI*.

Liu, Y.; Li, S.; Wu, Y.; Chen, C.-W.; Shan, Y.; and Qie, X. 2022. UMT: Unified Multi-modal Transformers for Joint Video Moment Retrieval and Highlight Detection. In *CVPR*, 3042–3051.

Loshchilov, I.; and Hutter, F. 2018. Decoupled Weight Decay Regularization. In *ICLR*.

Lu, Y.; Liu, J.; Zhang, Y.; Liu, Y.; and Tian, X. 2022. Prompt Distribution Learning. In *CVPR*, 5206–5215.

Luo, H.; Ji, L.; Zhong, M.; Chen, Y.; Lei, W.; Duan, N.; and Li, T. 2022. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 293–304.

Ma, F.; Zhu, L.; and Yang, Y. 2022. Weakly Supervised Moment Localization with Decoupled Consistent Concept Prediction. *IJCV*, 1244–1258.

Mahajan, D.; Tople, S.; and Sharma, A. 2021. Domain generalization using causal matching. In *ICML*, 7313–7324.

Mithun, N. C.; Paul, S.; and Roy-Chowdhury, A. K. 2019. Weakly supervised video moment retrieval from text queries. In *CVPR*, 11592–11601.

Mun, J.; Cho, M.; and Han, B. 2020. Local-global video-text interactions for temporal grounding. In *CVPR*, 10810–10819.

Nam, J.; Ahn, D.; Kang, D.; Ha, S. J.; and Choi, J. 2021. Zero-shot natural language video localization. In *ICCV*, 1470–1479.

Ning, K.; Xie, L.; Liu, J.; Wu, F.; and Tian, Q. 2021. Interaction-integrated network for natural language moment localization. *TIP*, 2538–2548.

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

Rao, Y.; Zhao, W.; Chen, G.; Tang, Y.; Zhu, Z.; Huang, G.; Zhou, J.; and Lu, J. 2022. Denseclip: Language-guided dense prediction with context-aware prompting. In *CVPR*, 18082–18091.

Regneri, M.; Rohrbach, M.; Wetzel, D.; Thater, S.; Schiele, B.; and Pinkal, M. 2013. Grounding action descriptions in videos. *TACL*, 25–36.

Robey, A.; Pappas, G. J.; and Hassani, H. 2021. Model-based domain generalization. *NeurIPS*, 20210–20229.

Shao, D.; Xiong, Y.; Zhao, Y.; Huang, Q.; Qiao, Y.; and Lin, D. 2018. Find and focus: Retrieve and localize video events with natural language queries. In *ECCV*, 200–216.

Sigurdsson, G. A.; Varol, G.; Wang, X.; Farhadi, A.; Laptev, I.; and Gupta, A. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 510–526.

Tian, C. X.; Li, H.; Xie, X.; Liu, Y.; and Wang, S. 2022. Neuron coverage-guided domain generalization. *TPAMI*, 1302–1311.

Wang, G.; Wu, X.; Liu, Z.; and Yan, J. 2022a. Prompt-based Zero-shot Video Moment Retrieval. In *MM*, 413–421.

Wang, H.; Zha, Z.-J.; Li, L.; Liu, D.; and Luo, J. 2021. Structured multi-level interaction network for video moment localization via language query. In *CVPR*, 7026–7035.

Wang, M.; Xing, J.; and Liu, Y. 2021. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*.

Wang, W.; Huang, Y.; and Wang, L. 2019. Language-driven temporal activity localization: A semantic matching reinforcement learning model. In *CVPR*, 334–343.

Wang, Z.; Wang, L.; Wu, T.; Li, T.; and Wu, G. 2022b. Negative sample matters: A renaissance of metric learning for temporal grounding. In *AAAI*, 2613–2623.

Wang, Z.; Zhang, Z.; Lee, C.-Y.; Zhang, H.; Sun, R.; Ren, X.; Su, G.; Perot, V.; Dy, J.; and Pfister, T. 2022c. Learning to prompt for continual learning. In *CVPR*, 139–149.

Xiao, S.; Chen, L.; Zhang, S.; Ji, W.; Shao, J.; Ye, L.; and Xiao, J. 2021. Boundary proposal network for two-stage natural language video localization. In *AAAI*, 2986–2994.

Xu, H.; Ghosh, G.; Huang, P.-Y.; Okhonko, D.; Aghajanyan, A.; Metze, F.; Zettlemoyer, L.; and Feichtenhofer, C. 2021. VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding. In *EMNLP*, 6787–6800.

Xu, H.; He, K.; Plummer, B. A.; Sigal, L.; Sclaroff, S.; and Saenko, K. 2019. Multilevel language and vision integration for text-to-clip retrieval. In *AAAI*, 9062–9069.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 5288–5296.

Yang, S.; and Wu, X. 2022. Entity-aware and Motion-aware Transformers for Language-driven Action Localization. In *IJCAI*, 1552–1558.

Yang, W.; Zhang, T.; Zhang, Y.; and Wu, F. 2021. Local correspondence network for weakly supervised temporal sentence grounding. *TIP*, 3252–3262.

Yang, X.; Wang, S.; Dong, J.; Dong, J.; Wang, M.; and Chua, T.-S. 2022. Video moment retrieval with cross-modal neural architecture search. *TIP*, 1204–1216.

Yuan, Y.; Ma, L.; Wang, J.; Liu, W.; and Zhu, W. 2019. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. In *NeurIPS*.

Yue, X.; Zhang, Y.; Zhao, S.; Sangiovanni-Vincentelli, A.; Keutzer, K.; and Gong, B. 2019. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *ICCV*, 2100–2110.

Zeng, R.; Xu, H.; Huang, W.; Chen, P.; Tan, M.; and Gan, C. 2020. Dense regression network for video grounding. In *CVPR*, 10287–10296.

Zhang, D.; Dai, X.; Wang, X.; Wang, Y.-F.; and Davis, L. S. 2019. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *CVPR*, 1247–1257.

Zhang, H.; Sun, A.; Jing, W.; Zhen, L.; Zhou, J. T.; and Goh, R. S. M. 2021a. Natural language video localization: A revisit in span-based question answering framework. *TPAMI*.

Zhang, M.; Yang, Y.; Chen, X.; Ji, Y.; Xu, X.; Li, J.; and Shen, H. T. 2021b. Multi-stage aggregated transformer network for temporal language localization in videos. In *CVPR*, 12669–12678.

Zhang, R.; Yu, Y.; Shetty, P.; Song, L.; and Zhang, C. 2022. Prompt-Based Rule Discovery and Boosting for Interactive Weakly-Supervised Learning. In *ACL*, 745–758.

Zhang, S.; Peng, H.; Fu, J.; and Luo, J. 2020a. Learning 2d temporal adjacent networks for moment localization with natural language. In *AAAI*, 12870–12877.

Zhang, Y.; Chen, X.; Jia, J.; Liu, S.; and Ding, K. 2023. Text-visual prompting for efficient 2d temporal video grounding. In *CVPR*, 14794–14804.

Zhang, Z.; Zhao, Z.; Lin, Z.; He, X.; et al. 2020b. Counterfactual contrastive learning for weakly-supervised vision-language grounding. In *NeurIPS*, 18123–18134.

Zhao, Y.; Zhao, Z.; Zhang, Z.; and Lin, Z. 2021. Cascaded prediction network via segment tree for temporal video grounding. In *CVPR*, 4197–4206.

Zheng, M.; Huang, Y.; Chen, Q.; Peng, Y.; and Liu, Y. 2022. Weakly Supervised Temporal Sentence Grounding With Gaussian-Based Contrastive Proposal Learning. In *CVPR*, 15555–15564.

Zhou, H.; Zhang, C.; Luo, Y.; Chen, Y.; and Hu, C. 2021. Embracing uncertainty: Decoupling and de-bias for robust temporal grounding. In *CVPR*, 8445–8454.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *CVPR*, 16816–16825.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *IJCV*, 2337–2348.

Zhou, L.; Xu, C.; and Corso, J. 2018. Towards automatic learning of procedures from web instructional videos. In *AAAI*.