Training-free Video Temporal Grounding using Large-scale Pre-trained Models

Minghang Zheng[1], Xinhao Cai[1], Qingchao Chen[2], Yuxin Peng[1], and Yang Liu[1,3⋆]

[1] Wangxuan Institute of Computer Technology, Peking University
[2] National Institute of Health Data Science, Peking University
[3] State Key Laboratory of General Artificial Intelligence, Peking University
{minghang,qingchao.chen,pengyuxin,yangliu}@pku.edu.cn
xinhao.cai@stu.pku.edu.cn

Abstract. Video temporal grounding aims to identify video segments within untrimmed videos that are most relevant to a given natural language query. Existing video temporal localization models rely on specific datasets for training, with high data collection costs, but exhibit poor generalization capability under the across-dataset and out-of-distribution (OOD) settings. In this paper, we propose a **T**raining-**F**ree **V**ideo **T**emporal **G**rounding (TFVTG) approach that leverages the ability of pre-trained large models. A naive baseline is to enumerate proposals in the video and use the pre-trained visual language models (VLMs) to select the best proposal according to the vision-language alignment. However, most existing VLMs are trained on image-text pairs or trimmed video clip-text pairs, making it struggle to (1) grasp the relationship and distinguish the temporal boundaries of multiple events within the same video; (2) comprehend and be sensitive to the dynamic transition of events (the transition from one event to another) in the video. To address these issues, firstly, we propose leveraging large language models (LLMs) to analyze multiple sub-events contained in the query text and analyze the temporal order and relationships between these events. Secondly, we split a sub-event into dynamic transition and static status parts and propose the dynamic and static scoring functions using VLMs to better evaluate the relevance between the event and the description. Finally, for each sub-event description provided by LLMs, we use VLMs to locate the top-k proposals that are most relevant to the description and leverage the order and relationships between sub-events provided by LLMs to filter and integrate these proposals. Our method achieves the best performance on zero-shot video temporal grounding on Charades-STA and ActivityNet Captions datasets without any training and demonstrates better generalization capabilities in cross-dataset and OOD settings. Code is available at https://github.com/minghangz/TFVTG.

Keywords: Video Temporal Grounding, Zero-shot Learning, Large Language Model, Vision Language Model

⋆ Corresponding author

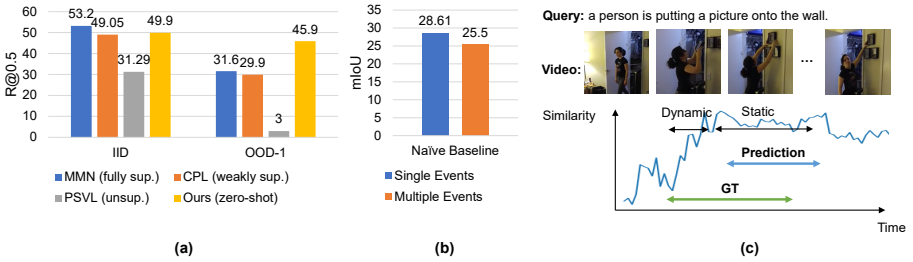

Fig. 1: (a) Evaluation results of existing methods and our method under the IID and OOD setting on the Charades-STA dataset. (b) Evaluation results of the naive baseline on the ActivityNet Datasets when the query describes single or multiple events. (c) The query-frame similarity obtained from the BLIP-2 Q-Former. The naive baseline based on BLIP-2 tends to predict the static parts of the video and ignores the dynamic transitions.

1 Introduction

Video temporal grounding [7] aims to localize the most semantically relevant segment in untrimmed videos according to free-form natural language queries. It has broad application potential in video surveillance [4], video summarization [29], and other fields. Existing video temporal grounding methods [7, 11, 14, 27, 40, 56–58, 62, 63] mainly train models on manually annotated data to understand the alignment between video segments and natural language queries. These methods have achieved remarkable improvements recently on specific datasets, such as ActivityNet Captions [16] and Charades-STA [7]. However, collecting high-quality video temporal grounding datasets is time-consuming and labor-intensive, which hinders the large-scale real-world application of these methods. In addition, they heavily rely on the distribution of the training dataset and the performance degrades significantly in out-of-distribution (OOD) or cross-dataset settings according to previous research [3, 19, 50, 53]. As shown in Figure 1 (a), we report the OOD performance [4] on the Charades-STA dataset of recent fully supervised method MMN [44], weakly supervised method CPL [61], and unsupervised method PSVL [33]. As we can see, their performance all has a significant drop. This is because these models are trained on small-scale video temporal grounding datasets that exhibit certain biases, leading to poor generalization of the models. Therefore, we aim to design a *training-free video temporal grounding* approach that *does not rely on specific video temporal grounding datasets*, so that it can be better generalized to real application scenarios.

Recently, large-scale pre-trained models [1, 21, 26, 34, 36, 42, 43, 55] have shown strong generalization ability in zero-shot retrieval [28, 52], VQA [9, 31], detection [17, 18, 47] et al. This inspires us to transfer the powerful generalization ability of them to video temporal grounding tasks. A naive baseline is to enumerate proposals in the video and use the pre-trained visual language models

[4] We follow [50] and plot the performance under the OOD-1 setting in the figure.

(VLMs) [21, 34, 42, 43] to assess the alignment between these proposals and the query and select the proposal with the highest score. However, this approach has the following drawbacks. *Firstly, it can be challenging for VLMs to understand the temporal boundaries of multiple events in untrimmed video.* Most of the existing image-text or video-text pre-trained VLMs are trained to align single images (e.g. CLIP [34], BLIP [21]) or trimmed video clips (e.g. InterVideo [43]) with texts. However, in the video temporal grounding, the model needs to understand multiple sequential events and their temporal relationships, such as 'She sprays it with a spray bottle and continues brushing her hair'. The skill is seldom emphasized in the image or trimmed video pretraining, and as shown in Figure 1(b), the naive baseline has a poorer performance when the query describes multiple events. *Secondly, we find that directly selecting the most relevant proposal using VLMs often leads to overlooking the dynamic transitions at the beginning of an event.* As shown in Figure 1(c), we show the similarity between video frames and the query text using the BLIP-2 Q-Former [20]. It can be found that in the naive baseline's prediction, the beginning of the event where a person gradually picks up the picture and approaches the wall is ignored. We think this is because these VLMs are trained directly by visual-textual contrastive learning. In such a training paradigm, the model's primary objective is to associate the most discriminative visual cues with their corresponding text descriptions, rather than emphasizing the need to focus on dynamic transitions and ensure the completeness of localized event boundaries.

To address the above problems, we propose to combine the ability of large language models (LLMs) [1, 26, 36] to understand and reason about queries and the ability of visual language models (VLMs) to align vision and text in a training-free manner. Specifically, for the challenge of understanding videos and queries that contain multiple events, we propose to prompt the LLMs to analyze the multiple events that may be contained in the query and give a text description of each single event as sub-queries, the order in which each event occurs, and the relationship between events (sequential or simultaneous). For the sub-query of each single event, we can use the VLMs to locate its possible occurrence in the video. To better use the VLMs to locate the video clip corresponding to the single event query and solve the problem of ignoring the dynamic transitions in the video while enhancing localization completeness, we propose to consider both dynamic transition and static status after transition explicitly when measuring the similarity between the proposal and the text query. For example in Figure 1(c), for the query 'a person put a picture on the wall', the dynamic transition part is where the person gradually raises the picture and approaches the wall, while the static status part is where the person has already hung the picture on the wall and is looking at the camera. A good proposal should begin with a dynamic segment, exhibiting a notable increase in video-text similarity and followed by a static post-status segment characterized by a high average video-text similarity within the static segment, while displaying a low average similarity outside of it. To evaluate each proposal, we compute a matching score by summing the dynamic score, which measures the rate of similarity change

within its dynamic segment, and the static score, which evaluates the comparative similarity within and outside its static post-status segment. Then, we select the top proposals with the highest matching scores as the localization results of the single event description. Finally, the localization results of each single event are combined with the LLM's judgment of the relationship and order of events to filter and integrate the final predictions.

Our contributions are summarized as follows. (1) We propose a training-free pipeline for video temporal grounding (TFVTG) using pre-trained large language models (LLMs) and vison-language models (VLMs). We use the LLMs to split the original query into sub-events and reason the temporal order and relationship between them, use VLMs to localize each sub-event, and filter and integrate the localization results based on the temporal order and relationships. (2) To help VLM better understand the dynamic transitions in the video, we divide the events into dynamic and static parts and model them separately. For the dynamic part, measures the rate of similarity change, and for the static part, we measure the comparative similarity within and outside. (3) Our method achieves the best performance on zero-shot video temporal grounding on both the Charades-STA [7] and ActivityNet Captions [16] datasets and has a greater advantage in cross-dataset and OOD settings.

2 Related Work

2.1 Video Temporal Localization

Fully supervised video temporal grounding methods [5, 7, 11, 14, 25, 27, 32, 44, 50, 57] typically train models using manually annotated queries and start and end timestamps. For example, MMN [44] trains models to distinguish matched and unmatched video-sentence pairs collected from within videos and across videos; VTimeLLM [10] is the first to fine-tune large language models using video temporal grounding data. However, fully supervised methods are often influenced by annotation bias, leading to poor generalization. Weakly supervised learning [13, 60, 61] and unsupervised learning [23, 33, 51, 59] are often used to mitigate the high dependence on human annotation. For instance, PSVL [33] and SPL [59] train models by generating pseudo-labels within videos. Nevertheless, even without using manually annotated data, biases present in the training videos can still affect the generalization of these methods. Therefore, in this work, we focus on training-free video temporal grounding, aiming for stronger generalization and applicability in real-world scenarios. Recently, Luo et al. [28] and VTG-GPT [46] made the first attempt to training-free video temporal grounding. Luo et al. measure the similarity between video proposals and the text query using VLM while VTG-GPT measures the similarity between the video frame captions and the text query using LLM. They then make a prediction based on the similarity. However, they overlooked the order and relationship between the possible multiple events within the query, as well as the issue of VLM's insensitivity to dynamic transitions in the video due to their training scheme. In contrast, we propose to infer multiple sub-events contained in the query and

their order and relationship through LLM, model dynamic changes in the video explicitly to assist VLM in better localizing each sub-event, and filter and integrate the localization results based on the order and relationship between events inferred by LLM.

2.2 Bias in Video Temporal Localization Datasets

The mainstream video temporal grounding datasets [7,16] suffer from certain biases, which affect the generalization ability of models trained on these datasets. Many studies [3, 19, 50, 53] have investigated biases in video temporal grounding datasets. For example, [50, 53] study biases in the location of target segments. When there are significant changes in the distribution of locations, existing methods exhibit noticeable performance degradation. [19] explores biases in query texts and proposes the ActivityNet-CG and Charades-CG datasets to evaluate the generalization ability of existing methods in the novel combinations of phrases and novel words. [3] studies the cross-dataset generalization ability of existing models and finds that models trained on specific datasets perform poorly when testing on another dataset. Some works [3, 19, 49, 50] have focused on improving model generalization to specific biases. However, they are difficult to generalize to address various types of biases in video temporal grounding. Therefore, we attempt to study training-free video temporal grounding to overcome reliance on specific datasets and achieve better generalization across various scenarios.

2.3 Large-scale Pretrained Models in Video Understanding

In recent years, with the development of large-scale corpus [2, 35], model architecture [37, 38], and computational resources, large language models (LLMs) [1, 26, 36] have achieved rapid development, demonstrating powerful capabilities in text generation, chat, problem-solving, reasoning, et al. On the other hand, visual language models (VLMs) [21,34,43,48] have also shown strong generalization abilities in tasks such as multimodal alignment and retrieval. Some existing methods attempt to combine VLMs with LLMs in video tasks. For example, ChatVideo [41] utilizes pre-trained models such as trajectory detection, video captioning, and speech recognition to convert videos into text, which serves as input to LLMs for video understanding. Video-LLaMA [55], and VTimeLLM [10] project video features into the token embedding space of LLMs through fine-tuning to enable LLMs to understand videos. However, these methods perform poorly on video temporal grounding tasks as shown in Table 1. Even VTimeLLM, which specifically fine-tunes LLM using video temporal grounding data, still exhibits a significant performance gap compared to the traditional video temporal grounding method. We think that this may be because fine-tuning VLMs and LLMs on video temporal grounding datasets degrades their generalization, and encoding videos solely as input token sequences for LLMs makes it difficult for LLMs to accurately understand the time boundaries of various events. Therefore, we propose a training-free pipeline to combine the capabilities of LLMs and

VLMs for video temporal grounding tasks. We leverage the strengths of both: using LLMs to reason the sub-events contained in queries, their occurrence order, and relationships, using VLMs to measure the vision-text similarity and localize each sub-event, and integrating the final predictions based on the inferred sub-event order and relationships from LLM.

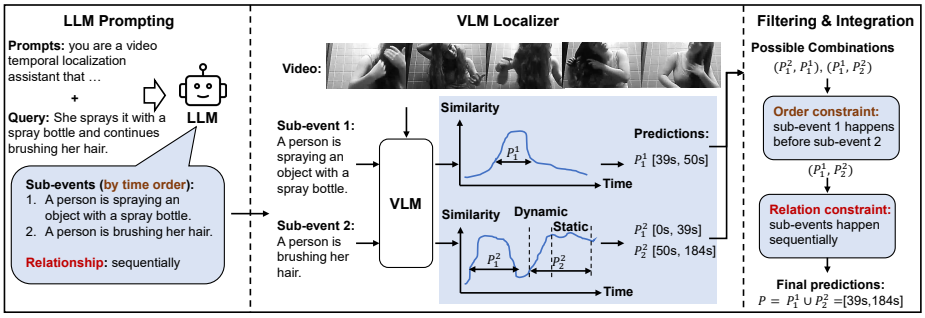

Fig. 2: The pipeline of the proposed method. Firstly, the LLM prompting leverages the large language model (LLM) to analyze sub-events contained in the query and reason the order and the temporal relationship of these sub-events. Then, the VLM localizer uses the vision language models to localize the sub-event in the video. The VLM localizer calculates the similarity between the video frames and the sub-event descriptions, enumerates event proposals in the video, and explicitly considers both dynamic transition and static status post-transition when measuring the similarity between the proposal and the text query, thus selecting proposals as the localization results. Finally, we filter and integrate the results of the VLM localizer based on the order and relationship of sub-events inferred by LLM to make the final prediction.

3 Method

The overview of our model design is illustrated in Figure 2. Our method consists of three parts: Firstly, since the query may describe multiple events that happened sequentially or simultaneously in the video, we propose the LLM prompting to leverage LLM for analyzing sub-events contained in the query and reason the order and the temporal relationship of these sub-events. Then, we propose the VLM localizer that uses the VLM to localize the sub-event in the video. The VLM localizer first calculates the similarity between the video frames and the sub-event descriptions. To solve the problem that VLMs are not sensitive enough to the dynamic transitions in the video, we propose explicitly considering both dynamic transitions and static states following these transitions when evaluating the similarity between the proposal and the text query. A good proposal should commence with a dynamic segment showing a significant rise in video-text similarity, followed by a static post-transition segment characterized by a high average video-text similarity within the static segment, while maintaining

a low average similarity outside of it. To assess each proposal, we enumerate the plausible dynamic and static segments of every event proposal and calculate a matching score by aggregating the dynamic and static scores. The dynamic score measures the rate of similarity change within its dynamic segment, and the static score evaluates the comparative similarity within and outside its static post-status segment. The VLM localizer returns the top-k proposals with the highest sum of dynamic and static scores as the localization results. Finally, we filter and integrate the results of the VLM localizer based on the order and relationship of sub-events inferred by LLM to make the final prediction.

3.1 LLM Prompting

The large language model (LLM) demonstrates powerful capabilities in instruction following, context understanding, and reasoning. Therefore, we propose to prompt LLM to analyze query texts, identifying multiple potential events therein, and analyzing the order and relationship of these events. Specifically, we request the large language model to provide the following information:

- Reasoning: Analyze the user's query and infer the sub-events it may contain.
- Order of sub-events: Provide each sub-events in chronological order.
- Relationships between sub-events: Consider three types of relationships: Single event, simultaneously (i.e. the sub-events occur simultaneously), and sequentially (i.e. the sub-events occur sequentially).
- Textual descriptions: Generate text descriptions for each sub-event.

The complete prompt will be provided in the supplementary materials.

3.2 Grounding with Vison Language Model

Inspired by the powerful multimodal alignment capabilities of VLM, we propose to use VLM as a localizer to locate each sub-event in the video.

Specifically, we choose BLIP-2 Q-Former [20] following [22,55] as the VLM localizer. For a sub-event description c and each video frame $v_1, ..., v_N$, we use VLM to extract their text features $f^c \in \mathbb{R}^D$ and vision features $F^v = [f_1^v, ..., f_N^v] \in \mathbb{R}^{N \times D}$, where N is the number of video frames and D is the feature dimension. As the text and vision space are well aligned, we can directly use the cosine similarity of the text and vision features to measure the relevance of the sub-event description and the video frame:

$$S = \frac{f^c F^{v\intercal}}{\|f^c\| \|F^v\|} \in \mathbb{R}^N \tag{1}$$

We find that directly enumerating proposals within the video and selecting the proposal with the highest average similarity often leads the model to only predict static status in the event while ignoring the dynamic transition at the beginning of an event. In 56.1% of the testing data in the Charades-STA dataset, the naive baseline predicts a start timestamp that is after the ground truth start

timestamp. For example, for the query "a person sits down", the model tends to predict segments where the person is already seated on the chair rather than the process of the person gradually from standing up to sitting down. A good event should commence with a dynamic segment showing a significant rise in video-text similarity, followed by a static post-transition segment characterized by a high average video-text similarity within the static segment, while maintaining a low average similarity outside of it. To assess each proposal, we enumerate the plausible dynamic and static segments of every event proposal and calculate a matching score by aggregating the dynamic and static scores. The dynamic score measures the rate of similarity change within its dynamic segment, and the static score evaluates the comparative similarity within and outside its static post-status segment:

Dynamic Scoring Considering the continuity of the video, in the dynamic part of the target segment, the video content transitions from another event to the target event described by the query, so the relevance between the video frames and the query should quickly increase. Our dynamic scoring aims to provide quantitative scores for segments where the relevance between the video and the query increases rapidly. Firstly, to eliminate the influence of video jitter, we apply a Gaussian filter to the similarity S: $\hat{S} = G(S)$, where $G(\cdot)$ is the Gaussian filter. We expect that the dynamic section contains as many parts as possible where the video-text correlation significantly increases. Therefore, we calculate the difference in similarity $\hat{S}$: $D_i = \hat{S}_i - \hat{S}_{i-1}$. Given a proposal starts from the i-th frame and ends with the k-th frame, we require that all the differential values within the proposal exceed a certain threshold δ. If this condition is met, we sum up the differential values in the proposal to obtain the dynamic score for that proposal:

$$S_{i,k}^{dynamic} = \begin{cases} \sum_{l=i}^{k} D_l, & D_l > \delta, \forall l \in [i,k] \\ 0, & otherwise \end{cases} \tag{2}$$

Static Scoring Inspired by SPL [59], in the static part, we require that the most relevant event for a given query should satisfy the requirement that videos within the event have high relevance to the query and videos outside the event have low relevance to the query. Therefore, given a proposal starts from the k-th frame and ends with the j-th frame, we calculate the average similarity within the proposal and the average similarity outside the proposal, and use the difference between them as the static scores:

$$S_{k,j}^{static} = \frac{1}{j-k} \sum_{l \in [k,j]} S_l - \frac{1}{N-(j-k)} \sum_{l \notin [k,j]} S_l \tag{3}$$

where N is the number of frames.

To localize the target video clip using dynamic and static scoring, we enumerate event proposals in the video. For each proposal (i, j), due to the varying lengths of transition segments in different events, we enumerate all feasible timestamps k where the transition just finished and divide the proposal into two parts: the dynamic part (i, k) and static part (k, j). We then calculate the sum of the dynamic score and static score, and select the maximum value as the score for this proposal:

$$S_{i,j}^{final} = \max_{k=i}^{j}(S_{i,k}^{dynamic} + S_{k,j}^{static}) \tag{4}$$

Finally, we select the top-k proposal with the highest final score S^{final} as the localization results of the VLM localizer.

3.3 Prediction Filtering and Integration

The VLM locator returns the top-k predictions for each sub-event description. To obtain the final prediction, we propose to filter and integrate these predictions from the VLP localizer based on the order of occurrence of sub-events and their relationships derived from LLM. Firstly, for each sub-event, we enumerate one of its top-k predictions, and there are total k^m combinations, where m is the number of sub-events. For a combination $P_1, P_2, ..., P_m$, we filter these combinations by the order constraint: if LLM determines that P_i from s^i to e^i should occur before P_j from s^j to e^j, but the start time of P_i is later than the end time of P_j (i.e. $s^i > e^j$), then this combination will be discarded. After filtering, we can calculate the sum of scores S^{final} returned by the VLM localizer for each combination, selecting the combination with the highest score, and merging these proposals based on the relation constraint: If LLM determines that these sub-events should occur simultaneously, we take the intersection of these proposals as the final prediction; otherwise, we take the union of these proposals as the final prediction:

$$P^{final} = \begin{cases} P_1 \cap P_2 \cap ... \cap P_m, & \text{relation is 'simultaneously'} \\ P_1 \cup P_2 \cup ... \cup P_m, & \text{relation is 'sequentially'} \end{cases} \tag{5}$$

4 Experiments

To comprehensively validate the effectiveness of our method, we conduct experiments on the Charades-STA [7] and ActivityNet Captions [16] datasets. We compare our method with existing methods under the IID, OOD, and cross-dataset settings respectively to demonstrate the generalization capability of our method. We also conduct ablation studies to evaluate the effectiveness of each module.

4.1 Experimental Setup

Dataset: We conduct experiment on two benchmark ActivityNet Captions [16] and Charades-STA [7]. *ActivityNet Captions* contains 20K videos, which is originally collected for video captioning. There are 37,417/17,505/17,031 video-query

pairs in the train /val_1/val_2 split. We follow previous works [28, 44] and report the performance on the val_2 split. *Charades-STA* is built based on the Charades dataset. There are 12,408/3,720 video-query pairs in the train/test split. We report the performance on the test split.

Evaluation Metrics: We follow the evaluation metrics 'R@m' and 'mIoU' in the previous work [28, 44], where m is the predefined temporal Intersection over Union (IoU) threshold. The metric 'R@m' represents the percentage of predictions that have the IoU larger than m. The metric 'mIoU' represents the average IoU of all the predictions.

Implementation Details: We follow [22, 55] to use the BLIP-2 Q-former [20] as the vision-language model. For the large language model, we use the GPT-4 Turbo API. For the VLM localizer, we downsample the input video to 3 FPS and use VLM to calculate the similarity between video frames and text. The localizer returns $k = 3$ predictions for each sub-event. For the hyperparameter, we set δ to 5×10^{-4} across all the datasets.

4.2 Comparison with the SOTAs

Comparison under the IID setting. In Table 1, we compare the performance of our method with existing methods under the IID setting. We use the official splits of Charades-STA and ActivityNet Captions. It can be observed that our method achieves the best performance in the zero-shot setting. For example, on the Charades-STA dataset, our method surpasses the second-ranked VTG-GPT [46] by 6.29% on the R@0.5 metric. Our method also outperforms unsupervised training methods by 9.27% on the R@0.5 metric on the Charades-STA dataset. The methods that utilize both LLM and VLM, such as VideoLLaMA [55] which aligns visual features to the input token space of LLM, have a poor performance on video temporal grounding. Although VTimeLLM [10] and GroundingGPT [24] further finetune LLM using data from ActivityNet-Captions or Charades-STA, the performance remains suboptimal. Our method combines the advantages of LLM in text understanding and reasoning with the advantages of VLM in visual-text alignment, resulting in superior performance.

Comparison under the OOD setting. To study the generalization capability of our method, we conducted experiments under OOD settings. We consider three OOD settings: novel location, novel text, and cross-dataset.

For the novel location, we follow DCM [50] by inserting a segment of random-generated video at the beginning of test videos, as shown in Table 2. In Table 3, we also test the performance on the Charades-CD [53] dataset, which alters the distribution of target location by resplitting the training and test data. As we can see, our method outperforms recent fully supervised methods in this setting without training, proving its superior generalization capability.

For the novel text, we follow VISA [19] and test on the Charades-CG [19] dataset as shown in Table 3. 'Novel-composition' indicates the text contains novel combinations of training vocabulary, while 'novel-word' indicates the text contains novel words. Our method achieves the best performance under the novel-word setting. Under the novel-composition setting, although not as competitive

Method	Setting	VLM	LLM	Charades-STA				ActivityNet Captions			
				R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU
2D-TAN [57]	fully	✗	✗	-	39.81	23.25	-	58.75	44.05	27.38	-
EMB [11]				**72.50**	58.33	39.25	**53.09**	**64.13**	44.81	26.07	**45.59**
MGSL-Net [25]				-	63.98	41.03	-	-	51.87	31.42	-
EaTR [14]				-	**68.47**	**44.92**	-	-	**58.18**	**37.64**	-
CRM [12]	weakly	✗	✗	53.66	34.76	16.37	-	55.26	32.19	-	-
CNM [60]				60.39	35.43	15.45	-	55.68	33.33	-	-
CPL [61]				66.40	49.24	22.39	-	55.73	31.37	-	-
Huang et al. [13]				**69.16**	**52.18**	**23.94**	**45.20**	**58.07**	**36.91**	-	**41.02**
Gao et al. [8]	unsup. [5]	✓	✗	46.69	20.14	8.27	-	46.15	26.38	11.64	-
PSVL [33]				46.47	31.29	14.17	31.24	44.74	30.08	14.74	29.62
PZVMR [39]				46.83	33.21	18.51	32.62	45.73	31.26	**17.84**	30.35
Kim et al. [15]				52.95	37.24	19.33	36.05	47.61	**32.59**	15.42	31.85
SPL [59]				**60.73**	**40.70**	**19.62**	**40.47**	**50.24**	27.24	15.03	**35.44**
GroundingGPT [24]	fully [6]	✓	✓	-	29.6	11.9	-	-	-	-	-
VTimeLLM-13B [10]				**55.3**	**34.3**	**14.7**	**34.6**	**44.8**	**29.5**	**14.2**	**31.4**
VideoChat-7B [22]	zero-shot	✓	✓	9.0	3.3	1.3	6.5	8.8	3.7	1.5	7.2
VideoLLaMA-7B [55]		✓	✓	10.4	3.8	0.9	7.1	6.9	2.1	0.8	6.5
VideoChatGPT-7B [30]		✓	✓	20.0	7.7	1.7	13.7	26.4	13.6	6.1	18.9
Luo et al. [28]		✓	✗	56.77	42.93	20.13	37.92	48.28	27.90	11.57	32.37
VTG-GPT [46]		✓	✓	59.48	43.68	**25.94**	39.81	47.13	28.25	12.84	30.49
Ours w/o LLM		✓	✗	65.46	48.01	22.07	43.37	48.84	26.64	13.10	33.61
Ours		✓	✓	**67.04**	**49.97**	24.32	**44.51**	**49.34**	**27.02**	**13.39**	**34.10**

Table 1: Evaluation Results on the Charades-STA and ActivityNet Captions Datasets.

as methods specifically designed for this scenario, such as DeCo and VISA, our method still outperforms other fully supervised approaches.

For the cross-dataset setting, we follow Debias-TLL [3] where the models are trained on ActivityNet Captions and tested on the Charades-STA, as shown in Table 4. Notably, almost all fully supervised methods experience significant performance degradation when applied across datasets, while our method remains unaffected as it does not rely on training data distribution. These experiments demonstrate that our method has superior generalization capabilities, making it more suitable for practical application requirements.

4.3 Ablation Studies

To validate the effectiveness of each module, we conduct ablation studies on the Charades-STA dataset.

Effectiviness of each component. Table 5 shows the ablation studies on the effectiveness of the proposed modules. When disenabling the VLM localizer, we use the naive baseline. When disenabling Filtering&Integration, we simply take the top-1 predictions of the localizer for each sub-event and consider their union box as the final prediction. (1) From the 2nd row of the table, it can be

[5] Some of them claim to be under the zero-shot setting. However, they still require video data for training. We follow [28] to classify them as unsupervised methods.

[6] They use the data in the ActivityNet Captions or Charades-STA to finetune LLMs.

Method	Setting	Charades-STA						ActivityNet-Captions					
		OOD-1			OOD-2			OOD-1			OOD-2		
		R@0.5	R@0.7	mIoU	R@0.5	R@0.7	mIoU	R@0.5	R@0.7	mIoU	R@0.5	R@0.7	mIoU
LGI [32]	fully	42.1	18.6	41.2	35.8	13.5	37.1	16.3	6.2	22.2	11.0	3.9	17.3
2D-TAN [57]		27.1	13.1	25.7	21.1	8.8	22.5	16.4	6.6	23.2	11.5	3.9	19.4
MMN [44]		31.6	13.4	33.4	27.0	9.3	30.3	20.3	7.1	26.2	14.1	**5.2**	20.6
VDI [27]		25.9	11.9	26.7	20.8	8.7	22.0	**20.9**	7.1	**27.6**	**14.3**	**5.2**	**23.7**
DCM [50]		**44.4**	**19.7**	**42.3**	**38.5**	**15.4**	**39.0**	18.2	**7.9**	24.4	12.9	4.8	20.7
CNM [60]	weakly	9.9	1.7	21.6	6.1	0.5	16.6	**6.1**	**0.4**	21.0	**2.5**	0.1	16.8
CPL [61]		**29.9**	**8.5**	**32.2**	**24.9**	**6.3**	**30.5**	4.7	**0.4**	**21.1**	2.1	**0.2**	**17.7**
PSVL [33]	unsup.	**3.0**	0.7	8.2	**2.2**	0.4	6.8	-	-	-	-	-	-
PZVMR [39]		-	**8.6**	**25.1**	-	**6.5**	**28.5**	-	**4.4**	**28.3**	-	**2.6**	**19.1**
Luo et al. [28]	zero-shot	40.3	18.2	38.2	38.9	17.0	37.8	18.4	6.8	21.1	**18.6**	7.4	20.6
Ours		**45.9**	**20.8**	**43.0**	**43.8**	**20.0**	**42.6**	**20.4**	**11.2**	**31.7**	18.5	**10.0**	**30.3**

Table 2: Results under OOD setting on the Charades and ActivityNet Dataset.

Method	Setting	Charades-CD			Charades-CG					
		test-ood			novel-composition			novel-word		
		R@0.3	R@0.5	R@0.7	R@0.5	R@0.7	mIoU	R@0.5	R@0.7	mIoU
2D-TAN [57]	fully	43.45	30.77	11.75	30.91	12.23	29.75	29.36	13.21	28.47
TSP-PRL [45]		31.93	19.37	6.20	16.30	2.04	13.52	14.83	2.61	14.03
SCDM [54]		**52.38**	**41.60**	**22.22**	27.73	12.25	30.84	-	-	-
VISA [19]		-	-	-	45.41	**22.71**	**42.03**	**42.35**	**20.88**	**40.18**
DeCo [49]		-	-	-	**47.39**	21.06	40.70	-	-	-
WSSL [6]	weakly	**35.86**	**23.67**	**8.27**	3.61	1.21	8.26	2.79	0.73	**7.92**
CPL [61]		-	-	-	**39.11**	**15.60**	**35.53**	**45.90**	**22.88**	-
SPL [59]	unsup.	62.96	38.25	15.53	-	-	-	-	-	-
Luo et al. [28]	zero-shot	-	-	-	40.27	16.27	-	45.04	21.44	-
Ours		**65.07**	**49.24**	**23.05**	**43.84**	**18.68**	**40.19**	**56.26**	**28.49**	**46.90**

Table 3: Results under OOD setting on the Charades-CD and Charades-CG Dataset.

observed that using LLM alone without Filtering&Integration may hurt some metrics. This is because the descriptions of sub-events usually only capture a part of the semantics of the original query, and the localizing results are inaccurate without Filtering&Integration. (2) From the 3rd row of the table, it can be seen that when both LLM prompting and Filtering&Integration are used, the model outperforms the naive baseline by 1.46% in mIoU. (3) From the 4th row of the table, our proposed VLM localizer shows a significant performance improvement, with a 5.69% increase in R@0.5 compared to the naive baseline. (4) In the 6th row, when all three modules are enabled, performance is further improved, demonstrating the effectiveness of our method.

Effectiviness of VLM. Table 6 shows the effectiveness of our two scoring functions in the VLM Localizer. (1) From the second row of the table, it can be observed that the dynamic scoring significantly improves performance, with an increase of 10.07% in mIoU. Since most VLMs are trained on image-text or trimmed video clip-text data, they are not sensitive enough to the dynamic transitions between different events in the same video. Dynamic scoring models the dynamic transitions implicitly, thus demonstrating better performance. (2) From the third row of the table, when using static scoring, the mIoU of the naive baseline improves by 10.2%. Static scoring compared with the naive

Method	R@1 R@0.5	R@1 R@0.7	R@5 R@0.5	R@5 R@0.7
SCDM [54]	15.91	6.19	54.04	30.39
2D-TAN [57]	15.81	6.30	59.06	31.53
Debias-TLL [3]	21.45	10.38	62.34	32.90
Ours	**49.97**	**24.32**	**83.5**	**42.2**

Table 4: Cross-dataset performance when training on ActivityNet captions and evaluate on Charades-STA.

	LLM prompting	VLM localizer	Filtering & Integration	R@0.5	R@0.7	mIoU
1	✗	✗	✗	42.32	18.91	31.61
2	✓			43.17	18.56	32.14
3	✓		✓	44.12	19.21	33.07
4		✓		48.01	22.07	43.37
5	✓	✓		48.41	21.94	42.76
6	✓	✓	✓	**49.97**	**24.32**	**44.51**

Table 5: Ablations on each component.

Dynamic Scoring	Static Scoring	R@0.5	R@0.7	mIoU
✗	✗	42.32	18.91	31.61
✓		47.63	20.13	41.68
	✓	45.48	22.02	41.81
✓	✓	**48.01**	**22.07**	**43.37**

Table 6: Ablations on VLM localizer.

Order Constraint	Relation Constraint	R@0.5	R@0.7	mIoU
✗	✗	42.32	18.91	31.61
✓		43.01	19.03	31.73
	✓	43.97	19.11	32.76
✓	✓	**44.12**	**19.21**	**33.07**

Table 7: Ablations on LLM prompting.

baseline not only requires high visual-textual relevance within the event but also requires low visual-textual relevance outside the event, thereby avoiding the model's focus solely on the most discriminative video segments. Combining both approaches further enhances model performance, demonstrating the effectiveness of our VLM localizer.

In Table 8, we report the performance of different VLMs, including the image-text pre-trained model (CLIP [34] and BLIP-2 [20]) and video-text pre-trained model (InterVideo [43], ViCLIP [42]). It can be observed that BLIP-2 exhibits the best performance, even surpassing models trained on video-text data. We attribute this to the fact that the pretraining data for image-text is much larger than that for video-text (e.g. LAION400M [35] v.s. WebVid10M [2]), thus BLIP-2 demonstrates better generalization capability. Additionally, our designed dynamic scoring helps BLIP-2 better understand the dynamic transition in the videos. Notably, in Table 1, VideoLLaMA [55] and VideoChat [22] utilize the frozen BLIP-2 Q-Former. VTG-GPT employs MiniGPT [64], which is also based on the frozen BLIP-2. Therefore, our comparison with them is fair. Additionally, Luo et al. [28] use InterVideo as the VLM, and as shown in Table 8, our performance using InterVideo still surpasses them.

VLMs	Type	R@0.5	R@0.7	mIoU
CLIP [34]	Image	42.68	18.92	38.89
BLIP-2 [20]	Image	**48.01**	**22.07**	**43.37**
InterVideo [43]	Video	44.60	20.51	40.72
ViCLIP [42]	Video	44.01	20.48	40.25

Table 8: Ablations on the VLMs.

LLMs	R@0.5	R@0.7	mIoU
None	48.01	22.07	43.37
Gemini-1.0-Pro [36]	48.97	22.76	44.12
GPT-3.5 Turbo	49.23	23.11	**44.69**
GPT-4 Turbo	**49.97**	**24.32**	**44.51**

Table 9: Ablations on the LLMs.

Effectiviness of LLM. (1) We utilize the order and relationships of sub-events provided by LLM to filter and integrate the predictions of the VLM localizer. Table 7 verifies the effectiveness of Filtering&Integration. It can be observed that both order and relation constraints improve the performance, with the most significant improvement when both are used simultaneously. (2) In Table 9, we also report the performance of different LLMs, including Gemini-1.0-Pro, GPT-3.5 Turbo, and GPT-4 Turbo. The results indicate that more powerful LLM (e.g. GPT-4) can lead to better performance.

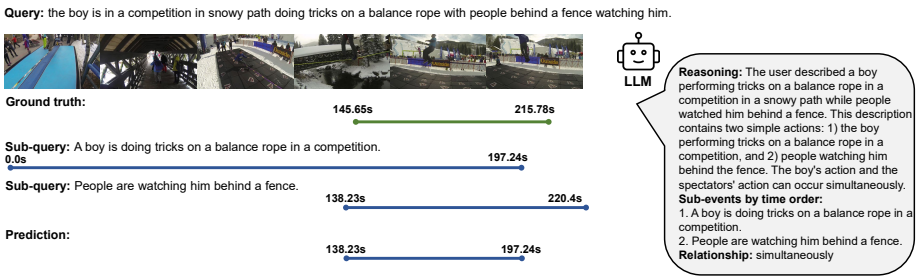

Fig. 3: Qualitative results on the ActivityNet Captions dataset.

4.4 Qualitative Results

Figure 3 presents a visualization. It can be seen that the LLM successfully splits the query into two sub-events and analyzes that both of these sub-events should occur simultaneously in the target query. Our VLM localizer successfully localizes these two sub-events, and their intersection forms the final prediction. More qualitative results can be found in the supplementary materials.

5 Conclusion

In this work, we study the problem of training-free video temporal grounding. We leverage the ability of LLM and VLM, requiring no specific video temporal localization dataset. We propose leveraging LLM to analyze multiple sub-events contained in the query and analyze the temporal order and relationships between these events. Then, we explicitly model the dynamic transition and static status in the video and use the VLM to localize the sub-events and leverage the order and relationships provided by LLMs to integrate the predictions. Our method achieves the best performance on zero-shot video temporal grounding on Charades-STA and ActivityNet Captions datasets without any training and demonstrates better generalization in cross-dataset and OOD settings.

Limitations. LLM are not always reliable in reasoning the order and relationships between sub-events, which can negatively impact the performance. How to validate the reliability of outputs from LLM can be studied in the future.

Acknowledgement

This work was supported by grants from the National Natural Science Foundation of China (62372014, 61925201, 62132001, U22B2048).

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1728–1738 (2021)
3. Bao, P., Mu, Y.: Learning sample importance for cross-scenario video temporal grounding. arXiv preprint arXiv:2201.02848 (2022)
4. Collins, R.T., Lipton, A.J., Kanade, T., Fujiyoshi, H., Duggins, D., Tsin, Y., Tolliver, D., Enomoto, N., Hasegawa, O., Burt, P., et al.: A system for video surveillance and monitoring. VSAM final report **2000**(1-68), 1 (2000)
5. Croitoru, I., Bogolin, S.V., Albanie, S., Liu, Y., Wang, Z., Yoon, S., Dernoncourt, F., Jin, H., Bui, T.: Moment detection in long tutorial videos. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2594–2604 (2023)
6. Duan, X., Huang, W., Gan, C., Wang, J., Zhu, W., Huang, J.: Weakly supervised dense event captioning in videos. Advances in Neural Information Processing Systems **31** (2018)
7. Gao, J., Sun, C., Yang, Z., Nevatia, R.: Tall: Temporal activity localization via language query. In: Proceedings of the IEEE international conference on computer vision. pp. 5267–5275 (2017)
8. Gao, J., Xu, C.: Learning video moment retrieval without a single annotated video. IEEE Transactions on Circuits and Systems for Video Technology **32**(3), 1646–1657 (2021)
9. Guo, J., Li, J., Li, D., Tiong, A.M.H., Li, B., Tao, D., Hoi, S.: From images to textual prompts: Zero-shot visual question answering with frozen large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10867–10877 (2023)
10. Huang, B., Wang, X., Chen, H., Song, Z., Zhu, W.: Vtimellm: Empower llm to grasp video moments. arXiv preprint arXiv:2311.18445 (2023)
11. Huang, J., Jin, H., Gong, S., Liu, Y.: Video activity localisation with uncertainties in temporal boundary. In: European Conference on Computer Vision. pp. 724–740. Springer (2022)
12. Huang, J., Liu, Y., Gong, S., Jin, H.: Cross-sentence temporal and semantic relations in video activity localisation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7199–7208 (2021)
13. Huang, Y., Yang, L., Sato, Y.: Weakly supervised temporal sentence grounding with uncertainty-guided self-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18908–18918 (2023)
14. Jang, J., Park, J., Kim, J., Kwon, H., Sohn, K.: Knowing where to focus: Event-aware transformer for video grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13846–13856 (2023)

15. Kim, D., Park, J., Lee, J., Park, S., Sohn, K.: Language-free training for zero-shot video grounding. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2539–2548 (2023)
16. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Niebles, J.C.: Dense-captioning events in videos. In: 2017 IEEE International Conference on Computer Vision (ICCV) (2017)
17. Lei, T., Yin, S., Liu, Y.: Exploring the potential of large foundation models for open-vocabulary hoi detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16657–16667 (2024)
18. Lei, T., Yin, S., Peng, Y., Liu, Y.: Exploring conditional multi-modal prompts for zero-shot hoi detection. In: Proceedings of the European Conference on Computer Vision (ECCV) (2024)
19. Li, J., Xie, J., Qian, L., Zhu, L., Tang, S., Wu, F., Yang, Y., Zhuang, Y., Wang, X.E.: Compositional temporal grounding with structured variational cross-graph correspondence learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3032–3041 (2022)
20. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
21. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. arXiv preprint arXiv:2201.12086 (2022)
22. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355 (2023)
23. Li, Z., Wang, P., Wang, Z., Zhan, D.c.: Flowgananomaly: Flow-based anomaly network intrusion detection with adversarial learning. Chinese Journal of Electronics **33**(1), 58–71 (2024)
24. Li, Z., Xu, Q., Zhang, D., Song, H., Cai, Y., Qi, Q., Zhou, R., Pan, J., Li, Z., Vu, V.T., et al.: Lego: Language enhanced multi-modal grounding model. arXiv preprint arXiv:2401.06071 (2024)
25. Liu, D., Qu, X., Di, X., Cheng, Y., Xu, Z., Zhou, P.: Memory-guided semantic learning network for temporal sentence grounding. arXiv preprint arXiv:2201.00454 (2022)
26. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36** (2024)
27. Luo, D., Huang, J., Gong, S., Jin, H., Liu, Y.: Towards generalisable video moment retrieval: Visual-dynamic injection to image-text pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23045–23055 (2023)
28. Luo, D., Huang, J., Gong, S., Jin, H., Liu, Y.: Zero-shot video moment retrieval from frozen vision-language models. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5464–5473 (2024)
29. Ma, Y.F., Lu, L., Zhang, H.J., Li, M.: A user attention model for video summarization. In: Proceedings of the Tenth ACM International Conference on Multimedia. p. 533–542. MULTIMEDIA '02, Association for Computing Machinery, New York, NY, USA (2002). https://doi.org/10.1145/641007.641116, https://doi.org/10.1145/641007.641116
30. Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models. arXiv preprint arXiv:2306.05424 (2023)

31. Mo, W., Liu, Y.: Bridging the gap between 2d and 3d visual question answering: A fusion approach for 3d vqa. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4261–4268 (2024)
32. Mun, J., Cho, M., Han, B.: Local-global video-text interactions for temporal grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10810–10819 (2020)
33. Nam, J., Ahn, D., Kang, D., Ha, S.J., Choi, J.: Zero-shot natural language video localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1470–1479 (2021)
34. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
35. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 (2021)
36. Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
37. Tian, Y., Fu, Y., Zhang, J.: Transformer-based under-sampled single-pixel imaging. Chinese Journal of Electronics **32**(5), 1151–1159 (2023)
38. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
39. Wang, G., Wu, X., Liu, Z., Yan, J.: Prompt-based zero-shot video moment retrieval. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 413–421 (2022)
40. Wang, H., Zha, Z.J., Li, L., Liu, D., Luo, J.: Structured multi-level interaction network for video moment localization via language query. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7026–7035 (June 2021)
41. Wang, J., Chen, D., Luo, C., Dai, X., Yuan, L., Wu, Z., Jiang, Y.G.: Chatvideo: A tracklet-centric multimodal and versatile video understanding system. arXiv preprint arXiv:2304.14407 (2023)
42. Wang, Y., He, Y., Li, Y., Li, K., Yu, J., Ma, X., Li, X., Chen, G., Chen, X., Wang, Y., et al.: Internvid: A large-scale video-text dataset for multimodal understanding and generation. arXiv preprint arXiv:2307.06942 (2023)
43. Wang, Y., Li, K., Li, Y., He, Y., Huang, B., Zhao, Z., Zhang, H., Xu, J., Liu, Y., Wang, Z., et al.: Internvideo: General video foundation models via generative and discriminative learning. arXiv preprint arXiv:2212.03191 (2022)
44. Wang, Z., Wang, L., Wu, T., Li, T., Wu, G.: Negative sample matters: A renaissance of metric learning for temporal grounding. In: AAAI. pp. 2613–2623. AAAI Press (2022)
45. Wu, J., Li, G., Liu, S., Lin, L.: Tree-structured policy based progressive reinforcement learning for temporally language grounding in video. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12386–12393 (2020)
46. Xu, Y., Sun, Y., Xie, Z., Zhai, B., Du, S.: Vtg-gpt: Tuning-free zero-shot video temporal grounding with gpt. Applied Sciences **14**(5), 1894 (2024)
47. Yang, D., Liu, Y.: Active object detection with knowledge aggregation and distillation from large models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16624–16633 (2024)

48. Yang, D., Xu, Z., Mo, W., Chen, Q., Huang, S., Liu, Y.: 3d vision and language pretraining with large-scale synthetic data. IJCAI (2024)
49. Yang, L., Kong, Q., Yang, H.K., Kehl, W., Sato, Y., Kobori, N.: Deco: Decomposition and reconstruction for compositional temporal grounding via coarse-to-fine contrastive ranking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23130–23140 (2023)
50. Yang, X., Feng, F., Ji, W., Wang, M., Chua, T.S.: Deconfounded video moment retrieval with causal intervention. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1–10 (2021)
51. Ye, Z., He, X., Peng, Y.: Unsupervised cross-media hashing learning via knowledge graph. Chinese Journal of Electronics **31**(6), 1081–1091 (2022)
52. Yelamarthi, S.K., Reddy, S.K., Mishra, A., Mittal, A.: A zero-shot framework for sketch based image retrieval. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 300–317 (2018)
53. Yuan, Y., Lan, X., Wang, X., Chen, L., Wang, Z., Zhu, W.: A closer look at temporal sentence grounding in videos: Dataset and metric. In: Proceedings of the 2nd international workshop on human-centric multimedia analysis. pp. 13–21 (2021)
54. Yuan, Y., Ma, L., Wang, J., Liu, W., Zhu, W.: Semantic conditioned dynamic modulation for temporal sentence grounding in videos. Advances in Neural Information Processing Systems **32** (2019)
55. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858 (2023)
56. Zhang, M., Yang, Y., Chen, X., Ji, Y., Xu, X., Li, J., Shen, H.T.: Multi-stage aggregated transformer network for temporal language localization in videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12669–12678 (2021)
57. Zhang, S., Peng, H., Fu, J., Luo, J.: Learning 2d temporal adjacent networks for moment localization with natural language. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12870–12877 (2020)
58. Zhao, Y., Zhao, Z., Zhang, Z., Lin, Z.: Cascaded prediction network via segment tree for temporal video grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4197–4206 (June 2021)
59. Zheng, M., Gong, S., Jin, H., Peng, Y., Liu, Y.: Generating structured pseudo labels for noise-resistant zero-shot video sentence localization. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 14197–14209 (2023)
60. Zheng, M., Huang, Y., Chen, Q., Liu, Y.: Weakly supervised video moment localization with contrastive negative sample mining. In: Proceedings of the AAAI Conference on Artificial Intelligence (2022)
61. Zheng, M., Huang, Y., Chen, Q., Peng, Y., Liu, Y.: Weakly supervised temporal sentence grounding with gaussian-based contrastive proposal learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
62. Zheng, M., Li, S., Chen, Q., Peng, Y., Liu, Y.: Phrase-level temporal relationship mining for temporal sentence localization. In: Proceedings of the AAAI Conference on Artificial Intelligence (2023)
63. Zhou, H., Zhang, C., Luo, Y., Chen, Y., Hu, C.: Embracing uncertainty: Decoupling and de-bias for robust temporal grounding. In: 2021 IEEE/CVF Conference

on Computer Vision and Pattern Recognition (CVPR). pp. 8441–8450. IEEE Computer Society, Los Alamitos, CA, USA (jun 2021)
64. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)