TAG: A Simple Yet Effective Temporal-Aware Approach for Zero-Shot Video Temporal Grounding

Jin-Seop Lee*,[1]
wlstjq0602@skku.edu
SungJoon Lee*,[1]
sjoon8379@skku.edu
Jaehan Ahn[2]
ajh508@skku.edu
YunSeok Choi[3]
ys.choi@skku.edu
Jee-Hyong Lee†,[1,4]
john@skku.edu

[1] Department of Artificial Intelligence, Sungkyunkwan University
[2] Department of AI system Engineering, Sungkyunkwan University
[3] Department of Immersive Media Engineering, Sungkyunkwan University
[4] Department of Computer Science and Engineering, Sungkyunkwan University

Abstract

Video Temporal Grounding (VTG) aims to extract relevant video segments based on a given natural language query. Recently, zero-shot VTG methods have gained attention by leveraging pretrained vision-language models (VLMs) to localize target moments without additional training. However, existing approaches suffer from semantic fragmentation, where temporally continuous frames sharing the same semantics are split across multiple segments. When segments are fragmented, it becomes difficult to predict an accurate target moment that aligns with the text query. Also, they rely on skewed similarity distributions for localization, making it difficult to select the optimal segment. Furthermore, they heavily depend on the use of LLMs which require expensive inferences. To address these limitations, we propose a *TAG*, a simple yet effective Temporal-Aware approach for zero-shot video temporal Grounding, which incorporates temporal pooling, temporal coherence clustering, and similarity adjustment. Our proposed method effectively captures the temporal context of videos and addresses distorted similarity distributions without training. Our approach achieves state-of-the-art results on Charades-STA and ActivityNet Captions benchmark datasets without rely on LLMs. Our code is available at github.com/Nuetee/TAG.

1 Introduction

Platforms like YouTube have become essential digital resources, offering vast video content for users to explore. However, manually locating specific segments within lengthy videos remains a time-consuming and labor-intensive task. Video Temporal Grounding (VTG) aims to automatically extract relevant segments from videos given a natural language query.

* Equal Contribution. † Corresponding author.

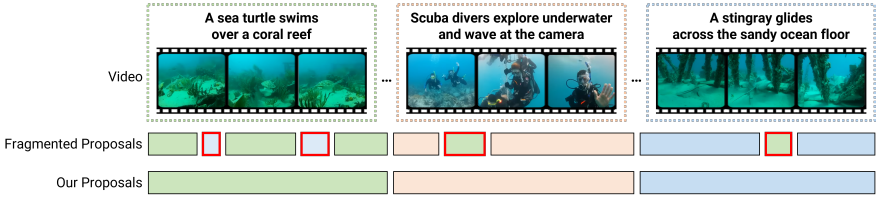


Figure 1: Motivation of our proposed method. Existing approaches tend to generate fragmented proposals, where semantically continuous video segments are divided into multiple disjoint proposals (red boxes). This is due to a lack of consideration for temporal context and temporal coherence during proposal generation.

Previous VTG approaches have been proposed to train models using pairs of natural language queries and their corresponding video moments [8, 9, 10, 11, 17, 20, 38, 39, 40]. However, creating well-annotated paired datasets is expensive to produce, and the models trained with such datasets demonstrate poor generalization performance as the datasets have a limited set of videos and texts. To address these issues, there has been growing interest in the field of zero-shot video temporal grounding, which leverages pretrained vision-language models (VLMs) to localize the target moment.

Since VLMs are pretrained on large datasets of image and text pairs, zero-shot video temporal grounding (ZSVTG) methods based on VLMs demonstrate strong generalization performance without additional fine-tuning [22, 32, 42]. To localize the target moment, they first generate candidate segments either by clustering video frame features [22] or by heuristic approaches (e.g., uniformly dividing the video into pre-defined intervals) [32, 42]. Then, the most relevant segment is selected based on the similarity between the segments and the given text query.

While these approaches [22, 32, 42] can localize segments that roughly correspond to the query, they often suffer from *semantic fragmentation*, where temporally continuous frames sharing the same semantics are split across multiple segments. Although adjacent video frame features which belong to the same action or scene, transient noise (e.g., variations in camera angle or lighting) can cause inconsistencies in their visual representations, as shown in fig. 3(a). As a result, they are assigned to different clusters, and their proposals are split regardless of the context, as illustrated in fig. 1. To mitigate semantic fragmentation, it is necessary to consider temporal context and enhance temporal coherence to accurately predict the target moment.

Another limitation is that most methods simply use alignment similarities without considering the similarity distribution when selecting the candidate segment most relevant to the query. When either a majority or a minority of frames in the video are close to the query, this can lead to skewed similarity distributions. Such skewed distributions may hinder the selection of the optimal segment, resulting in a decline in overall performance. To address this, adaptive similarity adjustment based on the similarity distribution is required.

We present TAG, a simple yet effective Temporal-Aware approach for zero-shot video temporal Grounding. To effectively capture the temporal context of images within videos, we propose temporal pooling and temporal coherence clustering. In temporal pooling, we incorporate temporal information into the extracted image features by aggregating features from adjacent images. Based on these temporally aggregated features, we generate contextual proposals by temporal coherence clustering. These methods help the model effectively capture the video context and generate boundary-aligned candidate proposals. Also, To pre-

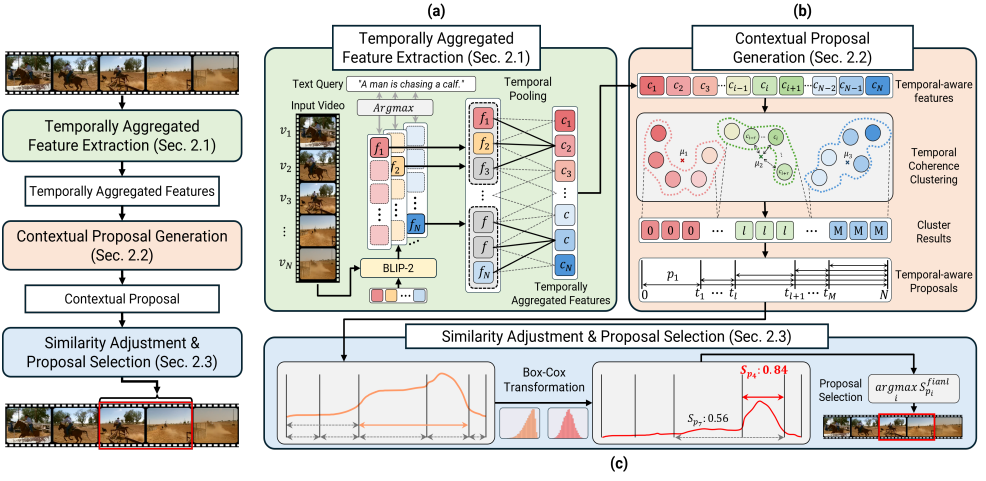


Figure 2: **Overview of our proposed method**. We first incorporate temporal dependencies of consecutive features using a temporal pooling (section 2.1). Based on this, we extract temporally aggregated features. Then, we generate contextual proposals. To effectively generate segment proposals, we propose temporal coherence clustering (section 2.2). Lastly, we adjustment similarities, and then, select the most relevant proposal (section 2.3).

vent performance degradation by the distorted similarity distribution when selecting the most suitable proposal, we propose similarity adjustment. This transformation normalized similarities, where higher values are amplified, and lower values are dampened.

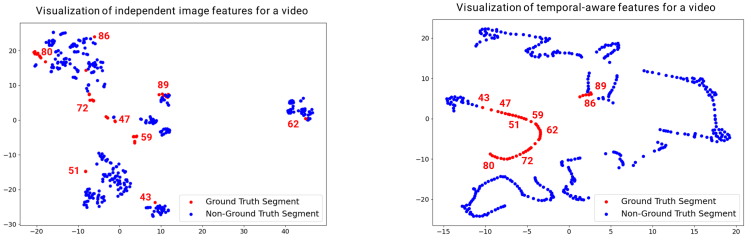
Notably, TAG outperforms existing approaches that heavily rely on large language models (LLMs), despite not using LLMs. This demonstrates that our method is not only simple and effective, but also cost-efficient. We validate our approach through experiments on the Charades-STA [5] and ActivityNet Captions [14] datasets and various scenarios. We achieve state-of-the-art performance across all datasets and scenarios, with up to 2.65% mIoU improvement on Charades-STA and 7.18% on ActivityNet Captions in general settings.

2 Proposed Method

We aim to the model effectively capture the video's temporal context information and enhance temporal coherence without any training. Also, we address the distorted similarity distributions and the reliance on LLMs. To achieve this goal, we introduce a TAG, simple yet effective Temporal-Aware approach for zero-shot video temporal Grounding. In this section, we introduce temporal pooling (section 2.1), temporal coherence clustering (section 2.2), and similarity adjustment (section 2.3).

2.1 Temporally Aggregated Feature Extraction

For a given video and text query, our first objective is to extract temporally aggregated features that incorporate temporal context information. We first extract single-frame image features using VLMs. As shown in fig. 2(a), we use pretrained BLIP-2 image encoder, text encoder, Q-Former, and learned queries [16]. Each image is extracted into multiple representations based on the number of learned queries, and among these, the image representa-



(a) Image representations from VLM (b) Temporal-aware representations

Figure 3: T-SNE visualizations for a single video. Red and blue points indicate representations of the ground truth segment and other segments, respectively. The numbers represent frame indices. (a) Features extracted from VLM are scattered without regard to time order. (b) Temporal-aware features are arranged in temporal sequence.

tion most similar to the text representation is selected. Then, we could extract single-frame image features $F = [f_1, f_2, \ldots, f_N]$ for N frames. However, they still do not contain temporal context information. To inject the temporal-aware information into the features, we propose a temporal-aware pooling, which uses sliding window-based average pooling.

The temporal pooling is a non-parametric convolutional layer, with a large window size w and a stride of 1. For frame i, the aggregated feature c_i is computed by averaging features across consecutive frames, and it is as follows:

$$c_i = \frac{1}{w} \sum_{j=i-(w-1)/2}^{i+(w-1)/2} f_j \tag{1}$$

where w is the kernel window size, and f_j represents the feature of the j-th frame. With a large window size, it enables capturing a wider range of context information. To maintain the original sequence length N, we apply feature padding at the edges.

Through the temporal pooling, which aggregates temporal context information across consecutive image features, we can extract temporally aggregated features $C = [c_1, \ldots, c_N]$. It is very simple yet effective at incorporating temporal context information into the model without any training. As shown in fig. 3(b), our features are not only located in the feature space according to the time order, but are also clearly separable based on temporal context.

2.2 Contextual Proposal Generation

With the extracted temporally aggregated features C, we aim to generate contextual proposals that align with context boundaries. To achieve this, we introduce temporal coherence clustering which enhances temporal coherence. Unlike conventional clustering methods [7] that consider only individual frames, our method assigns clusters by also taking into account temporally neighboring frames. Our temporal coherence clustering objective is as follows:

$$\arg\min_{\mathcal{S}} \sum_{j=1}^{k} \sum_{c^{(i)} \in S_j} \sum_{\Delta=-(r-1)/2}^{(r-1)/2} \left\| c^{(i+\Delta)} - \mu_j \right\|^2, \tag{2}$$

where $c^{(i)}$ is the feature of the i-th frame, μ_j is the centroid of the j-th cluster, and r indicates the temporal window size. When $r = 0$, the formulation is same as naive k-means clustering.

This objective encourages temporally adjacent frames to be assigned to the same cluster when they exhibit similar features, thereby enhancing temporal coherence and identifying the points where the video's temporal context changes.

As shown in fig. 2(b), based on temporal coherence clustering, we obtain clustering results $L = [l_1, \ldots, l_N]$ for all aggregated features. Since we use temporal pooling to extract aggregated features, adjacent aggregated features are very similar. Also, temporal coherence clustering encourages adjacent frames to be assigned to the same cluster. So, if video content does not change, images will belong to the same cluster. For all frames, we identify the points where the cluster labels change. These change points T are formulated as follows:

$$T = \{0, t_1, \cdots, t_M, N\}, \;\; t_m = \min\{i \mid \{l_i \neq l_{i+1}\} \cap \{i > t_{m-1}\}\}, \;\; m \in \{1, \cdots, M\} \quad (3)$$

where l_i is the cluster label of frame i, M is the total number of label changes. Here, M is not a hyperparameter, which is determined from the clustering result. Each t_m represents the m-th point where a cluster label change occurs, marking the boundary for segment proposals, and we add the start and end points of the frames to T.

Based on these change points T, we generate contextual proposals with diverse ranges. The generated contextual proposals are formulated as follows:

$$P = \left\{ p_k \mid p_k = (t_i, t_j),\, t_i, t_j \in T,\, i < j \right\}, \;\; k \in \left\{ 1, \cdots, \frac{(M+1)(M+2)}{2} \right\} \quad (4)$$

where P is the set of contextual proposals, and t_i and t_j denote the start and end points of each segment, respectively. This results in $(M+1)(M+2)/2$ possible segments, providing diverse temporal ranges within the video.

Our approach allows for generating more proposals for videos with frequent context changes and fewer proposals for videos with infrequent context changes. Additionally, since our candidate proposals are formed from all possible combinations of context boundaries, they have a diverse range of lengths.

2.3 Similarity Adjustment & Proposal Selection

To localize the target moment, we need to choose the most relevant segment among the temporal-aware proposals. To evaluate the semantic relevance for segment proposals, we calculate the alignment similarities between the text query q and single-frame image features F. At each i-th frame, the basic similarity is represented as $f_i \cdot q$.

However, the basic similarity-based proposal selection method does not consider the similarity distribution, which leads to skewed similarity distributions and hinders the selection of the optimal segment. To address this issue, we propose similarity adjustment. To make the distribution closer to a normal distribution, we apply a Box-Cox transformation [2]. The adjusted similarities are formulated as follows:

$$\mathrm{a}_i = \frac{(f_i \cdot q)^\lambda - 1}{\lambda}, \;\; i \in \{1, 2, \cdots, N\} \quad (5)$$

where λ is the parameter for the Box-Cox transformation, and $f_m \cdot q$ represents the similarity between the feature f_m of frame m and the query q. When similarities are normalized, it mitigates the skewness in the similarity distributions, as shown in fig. 2(c). This approach enables adaptive normalization based on the similarity distribution, which helps in selecting the optimal segment.

With these normalized similarities, we calculate the scores for all contextual proposals P. The proposal score we use is formulated as follows:

$$S = \{S_{p_i} \mid p_i = (t_i, t_j),\ p_i \in P\},\ \ S_{p_i} = \frac{1}{t_j - t_i} \sum_{m=t_i}^{t_j} a_m - \frac{1}{N - (t_j - t_i)} \sum_{n \notin [t_i, t_j)} a_n \tag{6}$$

where S is the set of proposal scores S_{p_i} for each proposal $p_i = (t_i, t_j) \in P$, with t_i and t_j representing the start and end points of the proposal. These proposal scores represent the difference in relevance between the segment and the non-segment regions.

Finally, based on the proposal scores, we rank all temporal-aware proposals, and select the proposal with the highest score as the final output. With our proposed method, the model effectively captures the temporal context of images within videos, and selects the most suitable proposal based on a non-distorted similarity distribution.

3 Experiments

3.1 Datasets

General Settings. To verify the effectiveness of our method, we conduct experiments on Charades-STA [8] and ActivityNet Captions [14] benchmark datasets. The Charades-STA dataset is an extension of the original Charades dataset, designed for video-query tasks. It contains 12,408 and 3,720 video-query pairs for the training and test splits, respectively, and we evaluate performance using the test split. The ActivityNet Captions dataset, initially created for video captioning tasks, consists of 20,000 videos. It includes 37,417, 17,505, and 17,031 video-query pairs in the train, valid-1, and valid-2 splits, respectively. Following previous works [22, 30, 42], we evaluate performance using the valid-2 split.

OOD Settings. To demonstrate that our training-free approach effectively enhances robustness and generalization, we conduct experiments across three OOD scenarios: altered target moment distributions, inserted noise moments, and unseen natural language queries.

In the first scenario, we examine the impact of altered target moment distributions. The Charades-STA dataset's training, validation, and test sets are restructured so that the target moment distribution in the test set differs from that in the training set.

The second scenario examines the effect of inserting noise into the video. We insert randomly generated video segments at the beginning of the test videos to create two types of noise-augmented data, as following DCM [35]. The temporal length of each video is extended to $\tau + \rho$, and the target moment's timestamps are adjusted to $(\tau_s + \rho, \tau_e + \rho)$ accordingly. For evaluation, ρ is set to $\{10, 15\}$ for Charades-STA and $\{30, 60\}$ for ActivityNet Captions. This allows us to evaluate the model's ability to locate target segments based on video context rather than relying on positional patterns.

For the third scenario, we generate unseen words in natural language queries during training with two types of textual noise: Insertion Composition Words, where queries combine words seen during training in novel ways, and Insertion Unseen Words, where queries include new words not encountered during training. This setup evaluates the model's ability to generalize to diverse textual queries without over-relying on the training vocabulary.

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 [58]	fully	✗	✗	-	39.81	23.25	-	58.75	44.05	27.38	-
EMB [48]		✓	✗	72.50	58.33	39.25	53.09	64.13	44.81	26.07	45.59
MGSL-Net [48]		✓	✗	-	63.98	41.03	-	51.87	31.42	-	
EaTR [28]		✓	✗	-	68.47	44.92	-	58.18	37.64	-	
UniVTG [58]		✓	✗	72.63	60.19	38.55	52.17			-	
CRM [6]	weakly	✗	✗	53.66	34.76	16.37	-	55.26	32.19	-	-
CNM [36]		✗	✗	60.39	35.43	15.45	-	55.68	33.31	-	-
CPL [69]		✗	✗	66.40	49.24	22.39	-	55.73	31.37	-	-
Huang et al. [28]		✗	✗	69.16	52.18	23.94	45.20	58.07	36.91	-	41.02
Gao et al. [6]	unsup.	✓	✗	46.69	20.14	8.27	-	46.15	26.38	11.64	-
PSVL [28]		✓	✗	46.47	31.29	14.17	31.24	44.74	30.06	14.74	29.62
PZVMR [36]		✓	✓	46.83	33.21	19.14	36.15	45.63	32.14	18.71	30.35
Kim et al. [28]		✓	✓	52.95	37.24	19.33	36.05	47.61	32.59	15.42	32.45
SPL [48]		✓	✓	60.73	40.70	19.62	40.47	50.24	27.24	15.03	35.44
GroundingGPT [28]	fully	✓	✓	-	29.6	11.9	-	-	-	-	-
TimeChat-7B [48]		✓	✓	40.6	23.8	9.7	26.2	25.0	13.2	6.1	18.5
VTimeLLM-13B [8]		✓	✓	55.3	34.3	14.7	34.6	44.8	29.5	14.2	31.4
VideoChat-7B [48]	zero-shot	✓	✓	9.0	3.3	1.3	6.5	8.8	3.7	1.5	7.2
VideoLLaMA-7B [28]		✓	✓	10.4	3.8	0.9	7.1	6.9	2.1	1.1	6.3
VideoChatGPT-7B [48]		✓	✓	20.0	7.7	1.7	13.7	26.4	13.6	6.1	18.9
UniVTG [58]		✓	✗	44.09	25.22	10.03	27.12	-	-	-	-
Luo et al.[48]		✓	✗	56.77	42.93	20.13	37.92	48.28	27.90	11.57	30.45
VTG-GPT [58]		✓	✓	59.48	43.68	25.94	39.81	47.13	28.25	12.84	30.49
TFVTG [48]		✓	✓	67.04	**49.97**	24.32	44.51	49.34	27.02	13.39	34.10
Ours	zero-shot	✓	✗	**67.82**	48.58	**26.67**	**45.69**	**51.88**	**28.91**	**15.07**	**36.55**

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

Method	Setting	Charades-CD		
		R@0.3	R@0.5	R@0.7
2D-TAN [58]	fully	43.45	30.77	11.75
TSP-PRL [48]	fully	31.93	19.37	6.20
SCDM [58]	fully	52.38	41.60	22.22
WSSL [6]	weakly	35.86	23.67	8.27
SPL [48]	unsup.	62.96	38.25	15.53
TFVTG [48]	zero-shot	65.07	49.24	23.05
Ours	zero-shot	**68.09**	**49.45**	**26.58**

Table 2: Results under OOD setting with altered target moment distributions on the Charades-CD dataset.

3.2 Results on General Settings

Our baselines consist of general video understanding models including VideoChat, VideoLLaMa, VideoChatGPT, and UniVTG [17, 19, 23, 37], along with zero-shot VTG models including Luo et al., VTG-GPT, and TGVTG [22, 32, 42]. Table 1 shows the comparison between our proposed method and recent state-of-the-art VTG methods on Charades-STA and ActivityNet Captions datasets, respectively. Our proposed method, which does not rely on LLMs, achieves the best performance across all evaluation metrics except for R@0.5 on Charades-STA. Notably, our method significantly surpasses others with mIoU performance improvements of 2.65% and 7.18%. These results clearly indicate that our method is highly effective in the zero-shot video temporal grounding task.

3.3 Results on OOD settings

Our proposed method demonstrates superior performance across all metrics in generalization, as shown in table 2. Since it effectively captures the temporal context of the given video, it can generate segment proposals that adapt to the video's temporal dynamics. Consequently, our method can generate accurate segments regardless of the statistical priors of the actual video segment data.

Method	Setting	Charades-STA						ActivityNet-Captions					
		ρ = 30			ρ = 60			ρ = 10			ρ = 15		
		R@.5	R@.7	mIoU	R@.5	R@.7	mIoU	R@.5	R@.7	mIoU	R@.5	R@.7	mIoU
LGI [24]	fully	42.1	18.6	41.2	35.8	15.4	37.1	16.3	6.2	22.2	11.0	3.9	17.3
2D-TAN [36]		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 [31]		31.6	13.4	33.4	27.0	9.3	30.3	23.0	7.1	26.2	14.3	5.2	20.6
VDI [25]		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 [35]		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 [38]	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 [34]		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 [18]	unsup.	3.0	0.7	8.2	2.2	0.4	6.8	-	-	-	-	-	-
PZVMR [23]		-	8.6	25.1	-	6.5	28.5	-	4.4	28.3	-	2.6	19.1
Luo et al. [20]	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
TFVTG [28]		**45.9**	20.8	43.0	43.8	20.0	42.6	20.4	11.2	31.7	18.5	10.0	30.3
Ours	zero-shot	45.3	**23.2**	**44.7**	**44.1**	**22.0**	**44.6**	**28.5**	**14.7**	**36.2**	**28.3**	**14.5**	**36.1**

Table 3: Results under OOD settings with inserted noise moments on benchmark datasets.

Method	Setting	Charades-CG					
		Insertion Composition Words			Insertion Unseen Words		
		R@0.5	R@0.7	mIoU	R@0.5	R@0.7	mIoU
2D-TAN [36]	fully	30.91	12.23	29.75	29.36	13.21	28.47
TSP-PRL [29]		16.30	2.04	13.52	14.83	2.61	14.03
SCDM [33]		27.73	12.25	30.84	-	-	-
VISA [37]		45.41	22.71	42.03	42.35	20.88	40.18
DeCo [32]		47.39	21.06	40.70	-	-	-
WSSL [1]	weakly	3.61	1.21	8.26	2.79	0.73	7.92
CPL [34]		39.11	15.60	35.53	45.90	22.88	-
Luo et al. [20]	zero-shot	40.27	16.27	-	45.04	21.44	-
TFVTG [28]		**43.84**	18.68	40.19	**56.26**	28.49	46.90
Ours	zero-shot	43.55	**21.30**	**41.95**	52.37	**32.66**	**47.86**

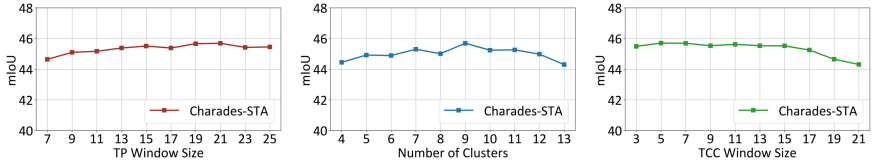
Table 4: Results under OOD settings with unseen text queries on the Charades-CG dataset.

TP	TCC	SA	Charades-STA				ActivityNet-Captions			
			R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU
✗	✗	✗	60.99	40.83	19.06	39.71	49.78	26.63	14.83	35.49
✓	✗	✗	65.16	47.34	24.38	43.91	50.90	26.32	14.72	36.41
✓	✓	✗	67.74	48.01	26.34	45.44	50.81	26.31	14.60	36.44
✓	✓	✓	**67.82**	**48.58**	**26.67**	**45.69**	**51.88**	**28.91**	**15.07**	**36.55**

Table 5: Performance with our proposed modules. TP, TCC, and SA represent the temporal pooling, temporal coherent clustering, and similarity adjustment, respectively.

Table 3 presents the experimental results for the noise insertion video OOD setting. Overall, zero-shot approaches demonstrate stronger robustness compared to training-based methods under noise-inserted scenarios. For Charades-STA dataset, our proposed method demonstrates the best performance across all evaluation metrics except for R@0.5 with $\rho = 30$. In contrast, for the ActivityNet-Captions dataset, our method significantly outperforms all metrics. The mIoU improvements are 14.20% and 19.14%, respectively. This demonstrates that our method achieves significant robustness against noise-inserted scenarios. It shows that our proposed method effectively aggregates consecutive features, which helps to mitigate the influence of noise-inserted segments.

In table 4, we observed that our proposed method shows better mIoU performance under two types of textual noise scenarios. For the insertion composition words scenario, our method consistently demonstrates good performance, and for the insertion unseen words scenario, it performs well in most cases. This result highlights that simply employing LLMs for query diversification is insufficient; instead, incorporating robust feature extraction and temporal context is crucial for handling noisy queries effectively.



(a) (b) (c)

Figure 4: Analysis on hyperparameters. mIoU with respect to (a) temporal pooling window size w, (b) number of clusters k, and (c) window size r for temporal coherence clustering.

Feature Extraction	R@0.3	R@0.5	R@0.7	mIoU
Gaussian ($\sigma = 1$)	65.27	47.72	24.27	43.96
Gaussian ($\sigma = 4$)	67.63	48.44	26.24	45.36
TP (Ours)	**67.82**	**48.58**	**26.67**	**45.39**

Table 6: Analysis of temporal pooling (TP)

Clustering	R@0.3	R@0.5	R@0.7	mIoU
K-Means	65.40	47.74	24.92	44.22
TW-FINCH	66.83	47.34	24.78	44.67
TCC (Ours)	**67.82**	**48.58**	**26.67**	**45.69**

Table 7: Analysis of temporal coherence clustering (TCC)

Setting	Top 25% Skewness			25-50% Skewness			50-75% Skewness			75-100% Skewness		
	R@.5	R@.7	mIoU	R@.5	R@.7	mIoU	R@.5	R@.7	mIoU	R@.5	R@.7	mIoU
Ours w/o SA	26.40	16.44	36.89	27.01	14.91	36.57	26.14	13.88	36.17	25.69	13.18	36.15
Ours	29.74	17.27	38.09	29.61	15.64	36.76	28.42	13.74	35.82	27.88	13.60	35.54
Difference	+3.34	+0.82	+1.21	+2.61	+0.73	+0.20	+2.28	-0.14	-0.34	+2.18	+0.42	-0.61

Table 8: Analysis of the effect of similarity adjustment on partitions ordered by similarity score skewness, from highest to lowest.

3.4 Ablation Study

Table 5 presents the performance of our method on both datasets with different combinations of our proposed modules. This demonstrates that our proposed method, which incorporates temporal pooling, temporal coherence clustering, and similarity adjustment, effectively generates high-quality proposals and localizes proposals more accurately.

3.5 Further Analysis

Hyperparameters. To validate the sensitivity of the hyperparameters, we conducted experiments on Charades-STA by varying the window size w for temporal pooling, the number of clusters k for temporal coherence clustering, and the window size r for temporal coherence clustering. The results are shown in fig. 4. Overall, these results demonstrate that our method is not sensitive for all hyperparameters. For the Charades dataset, the optimal values were set to $w = 21$, $k = 9$, and $r = 7$. Importantly, we did not tune hyperparameters separately for each dataset; instead, we used the same values for all experiments across both Charades-STA and ActivityNet.

Temporal pooling. While we uniformly pooled frame features over adjacent frames, other methods, such as gaussian pooling, can also be applied. Table 6 presents the results with a gaussian approach, showing that as the sigma value increases, the results become more similar to our method.

Temporal coherence clustering. As shown in table 7, our proposed temporal coherence clustering achieves higher performance compared to other clustering methods [7, 28]. This demonstrates that enforcing temporal consistency during clustering leads to more coherent segment boundaries, thereby improving proposal quality and localization accuracy.

Similarity adjustment. Since the ActivityNet dataset consists of diverse scenes, it exhibits higher similarity distribution skewness. Therefore, we analyze the effect of similarity adjustment across different levels of skewness on ActivityNet, as shown in table 8. Our method shows strong improvements under highly skewed conditions. This confirms that adjusting similarity distributions is crucial for robust proposal scoring and optimal segment selection.

4 Conclusion

In this paper, we presented TAG, a simple yet effective Temporal-Aware approach for zero-shot video temporal grounding. By incorporating temporal pooling, temporal coherence clustering, and similarity adjustment, our proposed method effectively captured the temporal context of videos and enhanced temporal coherence without additional training. Our approach achieved state-of-the-art results on the Charades-STA and ActivityNet benchmark datasets, without reliance on LLMs.

Acknowledgments

This work partly supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2019-II190421, AI Graduate School Support Program(Sungkyunkwan University), 20%), (RS-2025-02653113, High-Performance Research AI Computing Infrastructure Support at the 2 PFLOPS Scale, 20%), MSIT(Ministry of Science and ICT), Korea, under the Graduate School of Metaverse Convergence support program(IITP-2025-RS-2023-00254129, 20%), Artificial intelligence industrial convergence cluster development project funded by the Ministry of Science and ICT(MSIT, Korea) & Gwangju Metropolitan City (20%) and National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST) (RS-2024-00352717, 20%).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.

[2] George EP Box and David R Cox. An analysis of transformations. Journal of the Royal Statistical Society Series B: Statistical Methodology, 26(2):211–243, 1964.

[3] Ruizhe Chen, Zhiting Fan, Tianze Luo, Heqing Zou, Zhaopeng Feng, Guiyang Xie, Hansheng Zhang, Zhuochen Wang, Zuozhu Liu, and Huaijian Zhang. Datasets and recipes for video temporal grounding via reinforcement learning, 2025. URL https://arxiv.org/abs/2507.18100.

[4] Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. Weakly supervised dense event captioning in videos. Advances in Neural Information Processing Systems, 31, 2018.

[5] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In Proceedings of the IEEE international conference on computer vision, pages 5267–5275, 2017.

[6] Junyu Gao and Changsheng Xu. Learning video moment retrieval without a single annotated video. IEEE Transactions on Circuits and Systems for Video Technology, 32 (3):1646–1657, 2021.

[7] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. Journal of the royal statistical society. series c (applied statistics), 28(1): 100–108, 1979.

[8] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14271–14280, 2024.

[9] Jiabo Huang, Yang Liu, Shaogang Gong, and Hailin Jin. Cross-sentence temporal and semantic relations in video activity localisation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 7199–7208, 2021.

[10] Jiabo Huang, Hailin Jin, Shaogang Gong, and Yang Liu. Video activity localisation with uncertainties in temporal boundary. In European Conference on Computer Vision, pages 724–740. Springer, 2022.

[11] Yifei Huang, Lijin Yang, and Yoichi Sato. Weakly supervised temporal sentence grounding with uncertainty-guided self-training. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 18908–18918, 2023.

[12] Jinhyun Jang, Jungin Park, Jin Kim, Hyeongjun Kwon, and Kwanghoon Sohn. Knowing where to focus: Event-aware transformer for video grounding. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 13846–13856, 2023.

[13] Dahye Kim, Jungin Park, Jiyoung Lee, Seongheon Park, and Kwanghoon Sohn. Language-free training for zero-shot video grounding. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 2539–2548, 2023.

[14] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In Proceedings of the IEEE international conference on computer vision, pages 706–715, 2017.

[15] Juncheng Li, Junlin Xie, Long Qian, Linchao Zhu, Siliang Tang, Fei Wu, Yi Yang, Yueting Zhuang, and Xin Eric Wang. Compositional temporal grounding with structured variational cross-graph correspondence learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3032–3041, 2022.

[16] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In International conference on machine learning, pages 19730–19742. PMLR, 2023.

[17] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355, 2023.

[18] Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Vu Tu, et al. Groundinggpt: Language enhanced multi-modal grounding model. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6657–6678, 2024.

[19] Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. Univtg: Towards unified video-language temporal grounding. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 2794–2804, 2023.

[20] Daizong Liu, Xiaoye Qu, Xing Di, Yu Cheng, Zichuan Xu, and Pan Zhou. Memory-guided semantic learning network for temporal sentence grounding. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 1665–1673, 2022.

[21] Dezhao Luo, Jiabo Huang, Shaogang Gong, Hailin Jin, and Yang Liu. 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, pages 23045–23055, 2023.

[22] Dezhao Luo, Jiabo Huang, Shaogang Gong, Hailin Jin, and Yang Liu. Zero-shot video moment retrieval from frozen vision-language models. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 5464–5473, 2024.

[23] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. arXiv preprint arXiv:2306.05424, 2023.

[24] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10810–10819, 2020.

[25] Jinwoo Nam, Daechul Ahn, Dongyeop Kang, Seong Jong Ha, and Jonghyun Choi. Zero-shot natural language video localization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 1470–1479, 2021.

[26] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. the Journal of machine Learning research, 12:2825–2830, 2011.

[27] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14313–14323, 2024.

[28] M. Saquib Sarfraz, Naila Murray, Vivek Sharma, Ali Diba, Luc Van Gool, and Rainer Stiefelhagen. Temporally-weighted hierarchical clustering for unsupervised action segmentation, 2021. URL https://arxiv.org/abs/2103.11264.

[29] Guolong Wang, Xun Wu, Zhaoyuan Liu, and Junchi Yan. Prompt-based zero-shot video moment retrieval. In Proceedings of the 30th ACM International Conference on Multimedia, pages 413–421, 2022.

[30] Zhenzhi Wang, Limin Wang, Tao Wu, Tianhao Li, and Gangshan Wu. Negative sample matters: A renaissance of metric learning for temporal grounding. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 2613–2623, 2022.

[31] Jie Wu, Guanbin Li, Si Liu, and Liang Lin. Tree-structured policy based progressive reinforcement learning for temporally language grounding in video. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 12386–12393, 2020.

[32] Yifang Xu, Yunzhuo Sun, Zien Xie, Benxiang Zhai, and Sidan Du. Vtg-gpt: Tuning-free zero-shot video temporal grounding with gpt. Applied Sciences, 14(5):1894, 2024.

[33] Yifang Xu, Yunzhuo Sun, Zien Xie, Benxiang Zhai, and Sidan Du. Vtg-gpt: Tuning-free zero-shot video temporal grounding with gpt. Applied Sciences, 14(5):1894, 2024.

[34] Lijin Yang, Quan Kong, Hsuan-Kung Yang, Wadim Kehl, Yoichi Sato, and Norimasa Kobori. 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, pages 23130–23140, 2023.

[35] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. Deconfounded video moment retrieval with causal intervention. In Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval, pages 1–10, 2021.

[36] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. Advances in Neural Information Processing Systems, 32, 2019.

[37] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858, 2023.

[38] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 12870–12877, 2020.

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

[40] Minghang Zheng, Yanjie Huang, Qingchao Chen, Yuxin Peng, and Yang Liu. Weakly supervised temporal sentence grounding with gaussian-based contrastive proposal learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15555–15564, 2022.

[41] Minghang Zheng, Shaogang Gong, Hailin Jin, Yuxin Peng, and Yang Liu. 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), pages 14197–14209, 2023.

[42] Minghang Zheng, Xinhao Cai, Qingchao Chen, Yuxin Peng, and Yang Liu. Training-free video temporal grounding using large-scale pre-trained models. In European Conference on Computer Vision, pages 20–37. Springer, 2025.