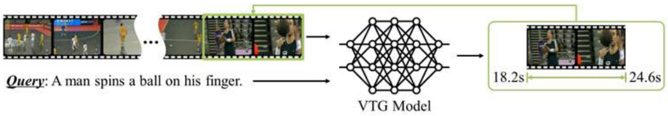
# TAG: A Simple Yet Effective Temporal-Aware Approach for Zero-Shot Video Temporal Grounding

SUNGKYUNKWAN UNIVERSITY(SKKU)
Jin-Seop Lee*, Sungjoon Lee*, Jaehan Ahn, YunSeok Choi, Jee-Hyong Lee†
Sungkyunkwan University, Republic of Korea

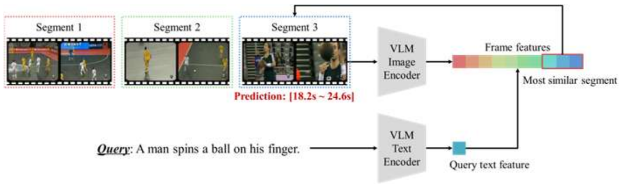
BMVC 2025

## Introduction

- **Video Temporal Grounding** aims to localize the most relevant segment in a video that corresponds to a given natural language query.

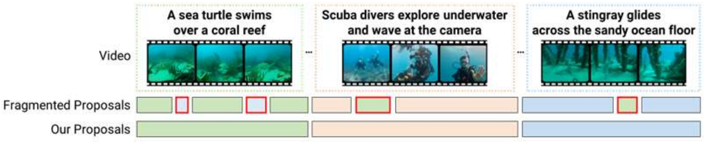



- Traditional VTG methods require annotated training data. Recently, **Zero-Shot approaches leverage large-scale pretrained VLMs** to achieve better generalization.
- Existing Zero-Shot VTG methods first generate candidate segments by **clustering frame features** or **uniformly dividing videos**, and then select the most relevant one based on similarity with the text query.

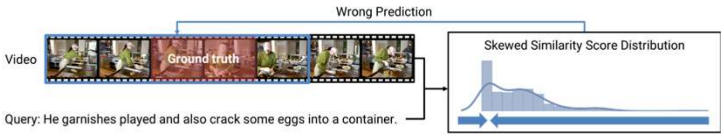



## Motivation

- However, existing methods **fail to consider temporal coherence**, often resulting in **semantic fragmentation** where continuous moments are split into disjoint segments.




- Since the target moment is determined by both **visual and temporal context**, these should be **jointly considered for accurate grounding.**
- In addition, most methods **ignore the distribution of frame-text similarities**, which vary across samples and may cause **biased scoring**, such as overly flat distributions that **hinder reliable segment selection.**




- For To ensure reliable grounding, it is essential to **adjust the similarity distribution**, amplifying meaningful matches and suppressing irrelevant ones

**⇒ We propose a Temporal-Aware Grounding(TAG) that captures temporal context and selects segments via adjusted similarity scores**

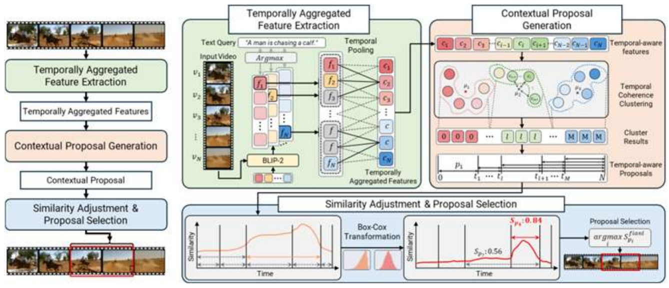
## Proposed Method

### ➢ Temporal-Aware approach for zero-shot video temporal Grounding (TAG)

- Our goal is to reduce segment fragmentation and selection bias by capturing temporal coherence and normalizing similarity distributions.

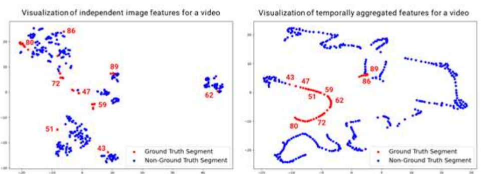
**[Overall Steps]**

**Step 1. Extract frame features and aggregate them using temporal pooling.**
**Step 2. Cluster features to generate proposals based on temporal coherence.**
**Step 3. Adjust the similarity to correct skewed similarity distribution.**
**Step 4. Select the final segment with the highest score as the prediction.**




### (1) Temporally Aggregated Feature Extraction

- Extract frame-wise features using BLIP-2 and apply sliding-window **temporal pooling** to generate temporally-aware representations.




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

### (2) Contextual Proposal Generation

- Cluster temporally-aware features based on **temporal coherence** and generate segment proposals using all combinations of cluster boundaries.
- Temporal coherence clustering objective:

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

- Where $c^{(i)}$ is the feature of the i-th frame, $\mu_j$ is the centroid of the j-th cluster, and $r$ indicates the temporal window size.

### (3) Similarity Adjustment & Proposal Selection

- Normalize frame-text similarity using a Box-Cox transformation, and select the proposal with the highest proposal score.
- The proposal score we use is formulated as follows:

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

- where $S_{p_i}$ is the score for proposal $p_i = (t_i, t_j)$, with $t_i$ and $t_j$ denoting the start and end points of the proposal, respectively, $a_t$ represents the similarity between the text and video frame at time $t$

## Experimental Results

- **Charades-STA and ActivityNet Captions datasets**

| Method | Setting | VLM | LLM | Charades-STA R@7 | Charades-STA mIoU | ActivityNet Captions R@7 | ActivityNet Captions mIoU |
|---|---|---|---|---|---|---|---|
| UniVTG | fully | x | x | 38.6 | 52.3 | - | - |
| Huang et al. | weakly | x | x | 23.9 | 45.2 | - | 41.0 |
| SPL | unsup. | v | v | 19.6 | 40.5 | 15.0 | 35.4 |
| VTimeLLM | fully | v | v | 14.7 | 34.6 | 14.2 | 31.4 |
| VideoChatGPT | zero-shot | v | v | 1.7 | 13.7 | 6.1 | 18.9 |
| UniVTG | zero-shot | v | x | 10.0 | 27.1 | - | - |
| Luo et al. | zero-shot | v | x | 20.1 | 37.9 | 11.6 | 30.5 |
| VTG-GPT | zero-shot | v | v | 25.9 | 39.8 | 12.8 | 30.5 |
| TFTVG | zero-shot | v | v | 24.3 | 44.5 | 13.4 | 34.1 |
| Ours | zero-shot | v | x | **26.7** | **45.7** | **15.1** | **36.6** |

- **Ablation Study**

| TP | TCC | SA | Charades-STA R@5 | Charades-STA R@7 | Charades-STA mIoU | ActivityNet Captions R@5 | ActivityNet Captions R@7 | ActivityNet Captions mIoU |
|---|---|---|---|---|---|---|---|---|
| x | x | x | 40.8 | 19.1 | 39.7 | 26.6 | 14.8 | 35.5 |
| v | x | x | 47.3 | 24.4 | 43.9 | 26.3 | 14.7 | 36.4 |
| v | v | x | 48.0 | 26.3 | 45.4 | 26.3 | 14.6 | 36.4 |
| v | v | v | **48.6** | **26.7** | **45.7** | **28.9** | **15.1** | **36.6** |

- **Analysis of similarity adjustment**

| Setting | Top 25% R@5 | Top 25% mIoU | 25~50% R@5 | 25~50% mIoU | 50~75% R@5 | 50~75% mIoU | 75~100% R@5 | 75~100% mIoU |
|---|---|---|---|---|---|---|---|---|
| w/o SA | 26.4 | 36.9 | 27.0 | 36.6 | 26.1 | 36.2 | 25.7 | 36.2 |
| Ours | 29.7 | 38.1 | 29.6 | 36.8 | 28.4 | 35.8 | 27.9 | 35.5 |
| Diff. | +3.3 | +1.2 | +2.6 | +0.2 | +2.3 | -0.3 | +2.2 | -0.6 |

(Degree of skewness)

- **Analysis of Semantic Fragmentation**

| TP | TCC | Charades-STA # of clusters per GT (↓) | ActivityNet Captions # of clusters per GT (↓) |
|---|---|---|---|
| x | x | 7.90 | 24.32 |
| v | x | 3.43 | 5.13 |
| x | v | 3.54 | 6.74 |
| v | v | 3.57 | 4.65 |

- **Qualitative Result**