STEER: STATE-TRANSITION EVENT EVIDENCE REASONING FOR TRAINING-FREE VIDEO TEMPORAL GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Video Temporal Grounding (VTG) aims to localize the temporal boundaries of a target event in an untrimmed video according to a natural language query. Existing training-free methods typically identify target moments based on video-query semantic relevance. However, semantic relevance does not necessarily indicate that the target event is actually occurring. Moments before and after the event may contain similar persons, objects, and scenes, leading to ambiguous temporal boundaries. To address this issue, we propose STEER (**S**tate-**T**ransition **E**vent **E**vidence **R**easoning), a training-free framework that explicitly reasons event progression through complementary semantic, motion, and query-independent caption evidence. STEER first parses the query into a structured event attributes, which guides the construction of complementary evidence from video-text semantics, action-aware human motion, and query-independent video captions. These cues are jointly used for event-state reasoning, and the resulting state transitions are used to predict the temporal boundaries. Experiments on Charades-STA and ActivityNet Captions demonstrate competitive performance, while ablation studies further validate the effectiveness of the proposed evidence construction and state-transition reasoning. Code and related resources will be publicly available upon publication.

1 INTRODUCTION

Video Temporal Grounding (VTG) (Gao et al., 2017; Hendricks et al., 2017) aims to locate the start and end timestamps of a video segment that matches a given natural language query in an untrimmed video. By establishing fine-grained temporal alignment between language queries and video content, VTG serves as a foundation for a wide range of downstream video understanding tasks, including video retrieval (Lin et al., 2023), video question answering (Ren et al., 2024), and long-video understanding (Grauman et al., 2022). In recent years, supervised deep learning methods (Zhang et al., 2020b; Lei et al., 2021a; Moon et al., 2023; Lin et al., 2023) have achieved strong performance on VTG . However, these methods usually rely on large-scale video-text annotations with precise temporal boundaries for training, resulting in high annotation costs and potentially limiting generalization to unseen datasets and real-world scenarios.

With the rapid development of Vision-Language Models (VLMs) (Bai et al., 2023; Wang et al., 2024), recent studies have increasingly explored training-free VTG (Luo et al., 2024; Xu et al., 2024; 2025; Zhang et al., 2025; Xu et al., 2026; Zhang et al., 2026), leveraging pre-trained vision-language knowledge to improve generalization without task-specific training or extensive temporal annotations. Some methods (Zheng et al., 2024; Zhang et al., 2026) estimate the video-query similarities to select candidate regions, while others (Ren et al., 2024; Huang et al., 2024; Xu et al., 2025) reason over sampled video content and directly infer temporal segments. Both approaches primarily predict temporal boundaries based on cross-modal semantic relevance over sampled clips. However, semantically similar content may persist across different stages of an event, making it difficult to distinguish different event states and resulting in ambiguous temporal localization.

This raises a fundamental question: rather than measuring semantic relevance, *can we determine which state of the target event a moment represents*? The difficulty lies in distinguishing these event

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

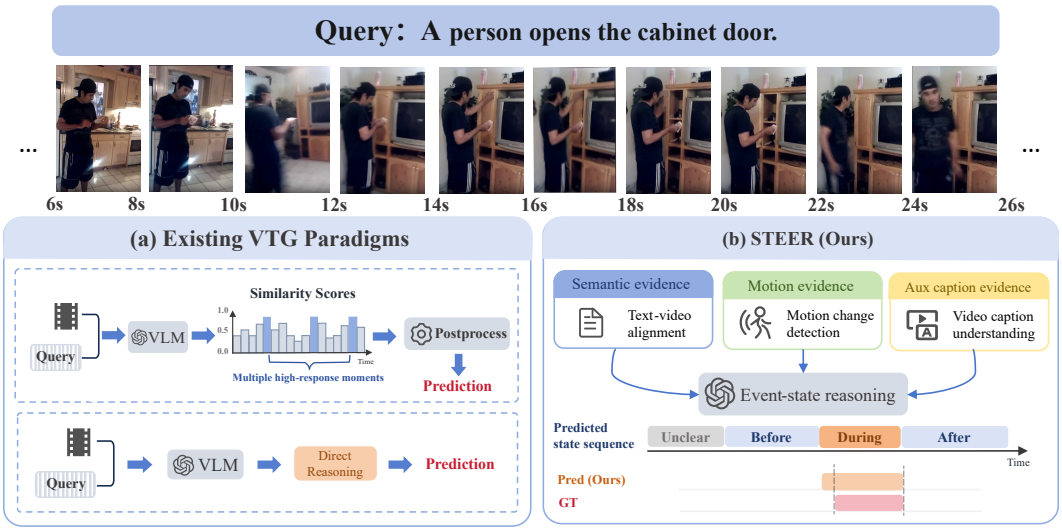


Figure 1: Motivation of STEER for training-free video temporal grounding. (a) Existing methods may produce similar semantic responses across different event stages, leading to ambiguous temporal localization. (b) STEER combines semantic, motion, and query-independent caption evidence to model BEFORE-DURING-AFTER transitions for temporal localization.

states, as moments before, during, and after an event may contain similar visual elements,such as people, objects, and scenes, and thus exhibit similarly high query relevance. Consider the query "A person opens the cabinet door." Before the action, the person may reach toward the door; during the action, the door is being opened; after the action, the door is already open. Although these stages share similar visual content, they correspond to different event states: closed → opening → open. As illustrated in Fig. 1(a), high semantic relevance does not necessarily indicate that the target event is occurring. Accurate temporal localization therefore requires modeling how the event progresses from Before → During → After. However, inferring such event transitions from semantic relevance alone remains challenging, as semantic correspondence does not explicitly capture the temporal evolution of the event.

Motivated by these observations, we propose STEER (**S**tate-**T**ransition **E**vent **E**vidence **R**easoning), an event-state reasoning framework for training-free VTG, as illustrated in Fig. 1(b). STEER first decomposes the query into a structured event attributes using a frozen VLM, including the event subject, core action, relevant object, and their relations. This representation guides the construction of complementary event evidence from three perspectives. Video-text semantic evidence measures query relevance, human pose and motion evidence captures action-related dynamics, and query-independent video captioning provides contextual descriptions beyond semantic matching. Using these cues, STEER jointly reasons over the current video content and the constructed evidence to classify each sampled moment as Before, During, or After. The resulting state transitions are then used to predict the start and end boundaries of the target event. We evaluate STEER on two standard VTG benchmarks, Charades-STA and ActivityNet Captions. Experimental results show that STEER achieves competitive temporal grounding performance without task-specific training. Further ablation studies verify the effectiveness of multi-source event evidence and explicit event-state reasoning.

The main contributions of this work are summarized as follows:

- We revisit training-free VTG and show that semantic relevance does not necessarily indicate event occurrence. We highlight the importance of modeling the Before → During → After progression for accurate temporal localization.
- We propose STEER, which integrates complementary evidence from video-text semantics, human motion, and query-independent video captions to perform explicit event-state reasoning.
- We evaluate STEER through extensive experiments on Charades-STA and ActivityNet Captions, demonstrating its effectiveness across various methods and datasets.

2 RELATED WORKS

2.1 VIDEO TEMPORAL GROUNDING

Video Temporal Grounding (VTG) aims to localize the temporal segment in an untrimmed video that corresponds to a natural language query. Existing methods can be broadly divided into proposal-based and proposal-free approaches. Proposal-based methods first generate temporal proposals of different lengths and then select the most relevant segment through cross-modal video-text matching. For example, 2D-TAN (Zhang et al., 2020b) models candidate moments on a two-dimensional temporal map to capture their temporal relations . In contrast, proposal-free methods directly predict the start and end boundaries of the target event. VSLNet (Zhang et al., 2020a) formulates VTG as a span prediction problem , while Moment-DETR (Lei et al., 2021b) adopts a Transformer-based framework for end-to-end temporal localization . Although these methods have achieved strong performance, most of them rely on temporally annotated data and task-specific training. This limitation has motivated increasing research on zero-shot and training-free VTG.

2.2 TRAINING-FREE VIDEO TEMPORAL GROUNDING

To reduce reliance on temporal annotations and task-specific training, recent studies have explored training-free video temporal grounding. TFVTG (Zheng et al., 2024) decomposes queries with an LLM and measures their relevance to video content using vision-language models, while VTime-CoT (Zhang et al., 2025) introduces visual temporal chain-of-thought reasoning. HiTeA (Xu et al., 2026) and DART (Zhang et al., 2026) further improve training-free localization through hierarchical temporal alignment and difficulty-adaptive reasoning, respectively. Despite these advances, existing methods provide limited explicit modeling of event progression over time. As a result, visually similar segments corresponding to different stages of the same event can be difficult to distinguish. This motivates us to explicitly model event-state transitions for temporal localization.

2.3 HUMAN MOTION AND ACTION UNDERSTANDING

Human pose and motion provide complementary cues to visual appearance for action understanding. Early methods (Cheron et al., 2015; Zolfaghari et al., 2017) combine appearance, motion, and pose information to improve action recognition. Later studies (Das et al., 2020; Duan et al., 2022) exploit pose-aware attention and skeleton representations to capture fine-grained motion patterns. PI-ViT (Reilly & Das, 2024) further incorporates 2D and 3D pose information into video Transformers to better distinguish visually similar actions. Prior work also shows that different human joints contribute unequally to action recognition, motivating joint-aware modeling (Shah et al., 2022). These findings suggest that action-relevant joint motion can provide useful cues for characterizing event progression. Inspired by this, we select joints according to the queried action and use their motion changes as auxiliary evidence for event-state reasoning.

3 METHOD

3.1 OVERVIEW

We propose STEER, an event-state reasoning framework for training-free video temporal grounding, as illustrated in Fig. 2. Given a text query and an untrimmed video, STEER first uses a frozen VLM to parse the query into a structured representation of the target event, including the subject, core action, related objects and their relations. Guided by this representation, STEER employs an **Event Evidence Extraction** module to construct complementary event evidence from video-text semantic relevance, human motion cues, and query-independent video captions. At each sampled moment, STEER incorporates **Event-State Reasoning** into the CoT process, where the current video content and constructed evidence are jointly analyzed to infer one of four states: BEFORE, DURING, AFTER, or UNCLEAR, forming an event-state sequence over time. Finally, STEER localizes the temporal boundaries by exploiting the state transitions between BEFORE, DURING, and AFTER.

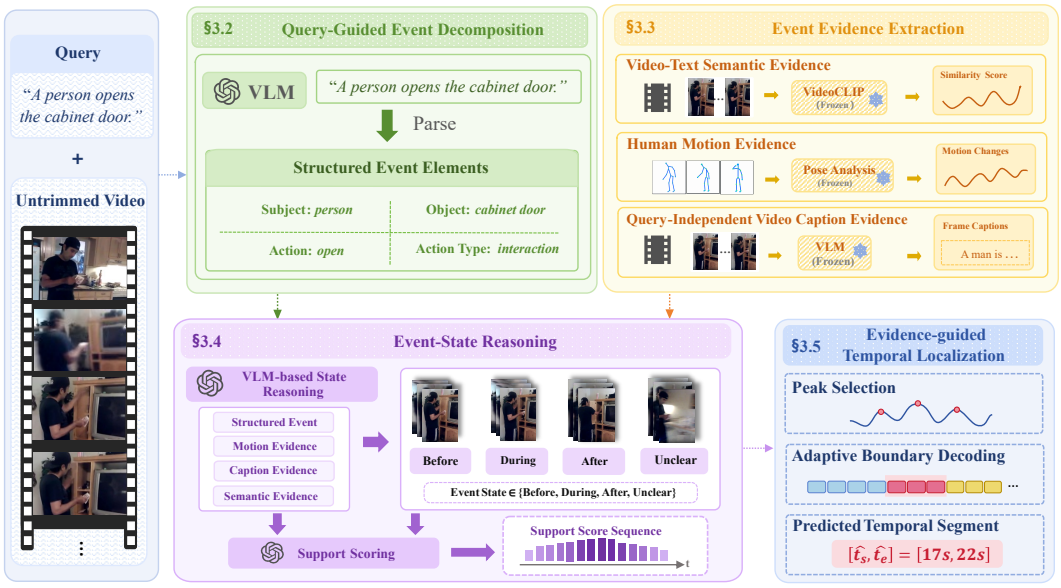


Figure 2: **Overview of STEER.** Given an untrimmed video and a text query, STEER decomposes the query into structured event attributes and constructs complementary semantic, motion, and query-independent caption evidence. These cues are used for event-state reasoning and event-support scoring, followed by evidence-guided peak selection and adaptive boundary decoding for temporal localization.

3.2 Query-guided Event Decomposition

Given a text query Q, the first step is to identify the target event from the query. A natural language query typically contains a subject, an action, related objects, and their interactions. Directly matching the full query with video content may be influenced by less relevant details, obscuring the core event semantics. We therefore use a frozen VLM to parse Q into structured event attributes. The resulting attributes define the target event and guides subsequent evidence construction and event-state reasoning. At each sampled moment, these event attributes are used to interpret the video content and infer its state. Further details on the query representation and action-specific visual focus are provided in Appendix A.1. The details of prompt used for event decompostion are provided in Appendix A.3.1.

Specifically, a frozen VLM is employed to extract the subject, core action, relevant objects, and their relations:

$$\text{Attr}^{\text{query}} = \{\text{Sub}, \text{Act}, \text{Obj}, \text{Rel}\}, \tag{1}$$

where Sub, Act, Obj, and Rel denote the event subject, core action, relevant objects, and interaction relations, respectively. The resulting representation defines the target event and guides subsequent evidence construction and event-state reasoning. At each sampled moment, these event elements are used to interpret the video content and infer its state.

3.3 Event Evidence Extraction

Video-Text Semantic Evidence. To measure video-text semantic relevance, we use a frozen VideoCLIP (Xu et al., 2021) to encode video and the text query into a shared feature space. Given an untrimmed videos $V = \{v_1, \cdots, v_T\}$ composing T frames and corresponding text query Q, we can extract fearure representations as follows:

$$F_t^{\text{video}}, f^{\text{query}} = \text{Encoder}(V, Q), \tag{2}$$

where $\text{Encoder}(\cdot)$ denotes the encoder of VideoCLIP, f^{query} denote extracted query features, and $F_t^{\text{video}} = \{f_1^{\text{video}}, \cdots, f_T^{\text{video}}\}$ defines extracted video features. Then, we can obtain the semantic evidence by calculating the semantic similarity scores:

$$\mathcal{E}^{\text{sem}} = \{s_t^{\text{sem}}\}_{t=1}^{t}, \quad s_t^{\text{sem}} = \frac{f_t^{\text{video}} \cdot f^{\text{query}}}{\|f_t^{\text{video}}\| \|f^{\text{query}}\|}. \tag{3}$$

Since semantic relevance alone cannot determine whether the target event is occurring, $\mathcal{E}^{\text{sem}}$ is used as complementary evidence for subsequent event-state reasoning.

Human Motion Evidence. High video-text similarity does not necessarily indicate that the target action is occurring. To capture action-related state changes, we further introduce human pose and motion evidence based on body keypoints. Specifically, we employ a pre-trained pose estimator (Jocher & Qiu, 2024) to extract key points for each frame in the video:

$$P_t = \Phi_{\text{pose}}(v_t) = \{p_t^j = (x_t^j, y_t^j)\}_{j=1}^{J}, \tag{4}$$

where p_t^j denotes the position of the j-th keypoint in frame v_t, J denotes the number of extracted keypoints, $\Phi_{\text{pose}}(\cdot)$ denotes the pre-trained pose estimator, and v_t is the t-th frame in video V.

Since different action types involve distinct body parts and motion patterns, treating all keypoints equally may introduce irrelevant motion cues. We therefore select action-relevant joints according to the action category. Specifically, the selected joint set is defined as

$$\mathcal{J}_{\text{Act}} = g(c_{\text{Act}}), \tag{5}$$

where c_{Act} denotes the action category and $g(\cdot)$ is a predefined action-to-joint mapping. For example, interaction actions mainly focus on the wrists, elbows, and shoulders, while locomotion actions emphasize the hips, knees, and ankles.

Based on the selected joints, we compute query-guided motion evidence $\mathcal{E}^{\text{mot}}$ by measuring the temporal displacement of keypoints between consecutive frames:

$$\mathcal{E}^{\text{mot}} = \{s_t^{\text{mot}}\}_{t=1}^{T}, \quad s_t^{\text{mot}} = \frac{1}{|\mathcal{J}_{\text{Act}}|} \sum_{j \in \mathcal{J}_{\text{Act}}} \left\| p_t^j - p_{t-1}^j \right\|_2, \tag{6}$$

where p_t^j and p_{t-1}^j denote the j-th keypoint of frame v_t and v_{t-1}, respectively. The resulting motion evidence captures action-related body changes while suppressing irrelevant movements.

Query-Independent Video Captioning. When a VLM is conditioned on both the query and video frames, it may overemphasize query-related visual cues and overlook subtle changes in actions or object states. This query-induced bias can make visually similar observations appear to correspond to the same event stage, which may affect subsequent event-state reasoning.

To reduce this bias, we introduce query-independent video captioning. The text query Q is not provided to the model. Instead, only the current video content is fed into the frozen VLM to generate an auxiliary description:

$$\mathcal{E}^{\text{cap}} = \{C_t\}_{t=1}^{T}, \quad C_t = \Phi_{\text{VLM}}(\text{Prompt}_{\text{cap}}, v_t), \tag{7}$$

where $\text{Prompt}_{\text{cap}}$ instructs the VLM to describe the current video content without access to the query, and $\mathcal{E}^{\text{cap}}$ denotes the resulting query-independent caption evidence sequence. These captions provide independent scene and object-state information rather than directly predicting temporal boundaries. They are combined with video-text semantic and human motion evidence for subsequent event-state reasoning. The details of the captioning prompt are provided in Appendix A.3.2.

3.4 Event-State Reasoning

VLM-based State Reasoning. The previous modules provide complementary evidence from video-text semantics, human motion, and query-independent video captions. Since each source alone may be ambiguous, we use a VLM to jointly reason over the current video content and these cues to infer the event state for each frame. Specifically, given a sampled frame v_t, the VLM infers the corresponding event state as:

$$z_t = \Phi_{\text{VLM}}\left(\text{Prompt}_{\text{state}}, v_t, \text{Attr}^{\text{query}}, \mathcal{E}_t\right), \tag{8}$$

where $\text{Prompt}_{\text{state}}$ denotes the event-state reasoning prompt, $\text{Attr}^{\text{query}}$ represents the structured event attributes extracted from the query, $\mathcal{E}_t = \{s_t^{\text{sem}}, s_t^{\text{mot}}, C_t\}$ denotes the event evidence collected for v_t, and z_t denotes the inferred event state belonging to four categories four categories: BEFORE,

DURING, AFTER, or UNCLEAR. BEFORE indicates that the target event has not started, DURING that it is occurring, and AFTER that it has ended. UNCLEAR is assigned when the available evidence is insufficient to determine the state reliably.

Event Support Scoring. The discrete state z_t describes the temporal phase of the target event, but does not indicate how strongly frame v_t supports the event. We therefore combine the predicted state with the semantic, motion, and caption evidence to obtain a continuous event-support score r_t for each sampled frame. Applying this scoring process along the video timeline yields the event-support score sequence $R = \{r_t\}_{t=1}^{T}$. The state sequence describes how the target event progresses over time, while the score sequence reflects the strength of event support at different frames. The support scores are subsequently used for evidence-guided temporal localization.The prompt templates used for event-state reasoning are provided in Appendix A.3.3.

3.5 Evidence-guided Temporal Localization

Peak Selection. Given the continuous event-support sequence R, we first identify the temporal region that most likely contains the target event. We first detect local peaks in the event-support sequence and rank them by their response scores. Temporally separated peaks are regarded as different candidate regions. When multiple candidate regions remain, we further compare their evidence and select the one that best matches the target event.

Adaptive Boundary Decoding. Starting from the selected peak, we scan the support sequence in both temporal directions. Strongly supported observations are included in the predicted interval, while short weak-support gaps are tolerated to handle local uncertainty. The scan stops when sustained low support or negative evidence is encountered. The outermost supported observations determine the start and end boundaries.

Temporal Output. The resulting start and end boundaries are denoted by $\hat{t}_s$ and $\hat{t}_e$, respectively. The final temporal prediction is

$$\hat{\tau} = [\hat{t}_s, \hat{t}_e]. \tag{9}$$

Unlike fixed-window localization, this strategy adapts the predicted interval to the temporal extent of each event.

4 Experiment

4.1 Experimental Setups

Datasets. We evaluate our method on two standard VTG benchmarks, Charades-STA (Gao et al., 2017) and ActivityNet Captions (Krishna et al., 2017). Charades-STA is built on the Charades dataset and contains 12,408 training pairs and 3,720 test pairs. Following the standard evaluation protocol, we report results on its official test split. ActivityNet Captions is a large-scale benchmark containing approximately 20K videos with temporally localized descriptions. Its train, val_1, and val_2 splits contain 37,417, 17,505, and 17,031 video-text pairs, respectively. Following previous works (Zheng et al., 2022b; Yan et al., 2023), we report results on val_2. We adopt a fully training-free setting for both benchmarks. All pre-trained models are used only for inference, without any parameter updates or task-specific fine-tuning.

Evaluation Metrics. We use R1@$\{0.3, 0.5, 0.7\}$ and mIoU as evaluation metrics. R1@0.3, R1@0.5, and R1@0.7 denote the percentages of samples whose predicted temporal segments achieve an IoU of at least 0.3, 0.5, and 0.7 with the ground-truth segments, respectively. mIoU measures the average IoU between the predicted and ground-truth segments over all samples.

Implementation Details. We use Qwen2.5-VL-7B-Instruct (Bai et al., 2025) for vision-language reasoning and keep its parameters frozen throughout all experiments. Video-text semantic matching is performed by a pre-trained VideoCLIP. When the similarity of the original query is low, we also compute the similarity of the simplified query and use the higher score as semantic evidence. Human pose is extracted by YOLO11m-Pose (Jocher & Qiu, 2024). We select action-related body keypoints based on the query and record their position changes over time. Query-independent video captions are also generated offline by the frozen Qwen2.5-VL-7B-Instruct and used for subsequent event-state reasoning. For videos of different lengths, we select observation positions according

Table 1: Performance comparison on Charades-STA and ActivityNet Captions. Setting: FS (Fully-Supervised), WS (Weakly-Supervised), US (Unsupervised), and ZS (Zero-Shot). The best and second-best results under the ZS setting are highlighted in bold and underline, respectively.

Method	Setting	Charades-STA				ActivityNet Captions			
		R1@0.3	R1@0.5	R1@0.7	mIoU	R1@0.3	R1@0.5	R1@0.7	mIoU
2D-TAN (Zhang et al., 2020b)	FS	–	39.70	23.31	–	59.45	44.51	26.54	–
Moment-DETR (Lei et al., 2021b)	FS	62.10	48.20	25.30	42.30	52.60	32.50	15.30	37.80
VTimeLLM (Huang et al., 2024)	FS	55.30	34.30	14.70	34.60	44.80	29.50	14.20	31.40
SCANet (Yoon et al., 2023)	WS	68.04	50.85	24.07	–	83.62	56.07	31.52	–
Huang et al. (2023)	WS	59.20	44.20	22.10	39.40	54.80	32.90	–	36.40
CNM (Zheng et al., 2022a)	WS	60.04	35.15	14.95	–	78.13	55.68	33.33	–
DSCNet (Liu et al., 2022)	US	44.15	28.73	14.67	–	–	50.12	31.07	–
CMKT (Fang et al., 2023)	US	47.80	30.96	18.87	30.42	50.69	31.28	16.42	38.79
CORONET (Holla & Lourentzou, 2024)	US	49.21	34.60	17.93	32.73	46.05	28.19	12.84	31.11
Luo et al. (2024)	ZS	57.77	42.93	20.13	37.92	**48.28**	27.90	11.57	32.37
VTG-GPT (Xu et al., 2024)	ZS	59.48	43.68	**25.94**	39.81	<u>47.13</u>	<u>28.25</u>	12.84	30.49
VTimeLLM (Huang et al., 2024)	ZS	51.00	27.50	11.40	31.20	44.00	27.80	14.30	30.40
TimeChat (Ren et al., 2024)	ZS	51.50	32.20	13.40	–	–	–	–	–
VTimeCoT (Zhang et al., 2025)	ZS	66.96	38.79	20.83	43.41	–	–	–	–
NumPro (Wu et al., 2025)	ZS	60.70	36.80	15.90	38.50	44.20	26.40	14.40	31.30
T2SGrid (Guo et al., 2026)	ZS	**71.40**	<u>47.00</u>	20.70	<u>44.90</u>	44.80	26.80	**16.10**	<u>32.50</u>
STEER (Ours)	ZS	<u>70.80</u>	**47.82**	<u>24.57</u>	**45.68**	46.82	**28.82**	<u>15.17</u>	**33.44**

to the video duration. Further details on video sampling and temporal alignment are provided in Appendix A.5.

4.2 MAIN RESULTS

Table 1 compares STEER with representative VTG methods under fully supervised, weakly supervised, unsupervised, and zero-shot settings. Since STEER requires no task-specific training or fine-tuning, we mainly focus on comparisons with zero-shot methods.

Performance on Charades-STA. STEER achieves 70.80% R1@0.3, 47.82% R1@0.5, 24.57% R1@0.7, and 45.68% mIoU on Charades-STA. Compared with T2SGrid, the strongest zero-shot baseline in terms of mIoU, STEER improves R1@0.3 from 69.30% to 70.80%, R1@0.5 from 47.00% to 47.82%, and mIoU from 44.90% to 45.68%. These improvements show that STEER can achieve more accurate temporal localization without task-specific supervision. At the stricter IoU threshold of 0.7, STEER remains competitive, although VTG-GPT obtains a slightly higher result. This suggests that STEER provides strong overall temporal overlap, while highly precise boundary alignment remains challenging.

Performance on ActivityNet Captions. On ActivityNet Captions, STEER achieves 46.82% R1@0.3, 28.82% R1@0.5, 15.17% R1@0.7, and 33.44% mIoU. It obtains the best mIoU and R1@0.5 among the compared zero-shot methods. This result is notable because ActivityNet Captions contains more diverse activities and temporal structures than Charades-STA. The consistent performance indicates that STEER is not limited to a specific video domain. Instead, event-state reasoning can generalize to videos with different content and temporal characteristics.

Across both benchmarks, STEER consistently achieves strong mIoU under the zero-shot setting. This suggests that its advantage does not come from simply enlarging the predicted temporal interval to increase coarse overlap. Instead, STEER explicitly models the temporal evolution of the target event and identifies boundaries according to state transitions. In addition, video-text semantics, human motion, and query-independent captions provide complementary cues for event-state reasoning. Their combination helps reduce the ambiguity caused by similar semantic responses at multiple temporal positions and leads to more reliable temporal localization.

4.3 ABLATION STUDIES

We mainly conduct ablation studies on the Charades-STA dataset to evaluate the contribution of each component.

Table 2: Ablation study of different evidence sources on Charades-STA. A checkmark indicates the enabled components. VLM denotes VLM-based event-state reasoning, VC denotes video-text semantic evidence, Aux denotes query-independent video captions, and Pose denotes human motion evidence.

VLM	VC	Aux	Pose	R1@0.5	R1@0.7	mIoU
✓				34.85	16.18	36.39
✓	✓			38.76	17.59	40.17
✓		✓		38.98	18.84	39.47
✓			✓	38.87	19.94	39.37
✓		✓	✓	40.62	21.64	42.18
✓	✓		✓	40.87	20.26	42.35
✓	✓	✓		41.86	21.53	42.23
✓	✓	✓	✓	**47.82**	**24.57**	**45.68**

Table 3: Ablation study of different pose strategies on Charades-STA. AA Pose denotes Action-aware Pose.

Strategy	R1@0.3	R1@0.5	R1@0.7	mIoU
w/o Pose	65.79	41.86	21.53	42.23
All-joint Pose	63.06	41.37	20.23	41.56
AA Pose	**70.80**	**47.82**	**24.57**	**45.68**

Table 4: Ablation study of different boundary strategies on Charades-STA. AE Boundary denotes Adaptive Evidence Boundary.

Strategy	R1@0.3	R1@0.5	R1@0.7	mIoU
Peak $\pm$ 4s	60.37	**48.11**	22.51	41.02
Peak $\pm$ 6s	67.43	47.32	20.17	42.45
Peak $\pm$ 8s	68.76	38.42	14.71	41.07
AE Boundary	**70.80**	47.82	**24.57**	**45.68**

Effect of Complementary Evidence. We first analyze the contribution of different evidence sources. Using only the VLM for event-state reasoning gives an mIoU of 36.39%. Adding VC, Aux, and Pose individually improves the mIoU to 40.17%, 39.47%, and 39.37%, respectively. Among them, VC provides the largest individual improvement in mIoU, indicating that video-text similarity is an important cue for identifying query-relevant temporal regions. In contrast, Pose provides the largest improvement at R1@0.7, increasing the score from 16.18% to 19.94%. This suggests that human motion is particularly useful for identifying more precise temporal changes around the target event.

However, no single evidence source is sufficient to achieve the performance of the full model. When VC, Aux, and Pose are combined, STEER reaches 47.82% R1@0.5, 24.57% R1@0.7, and 45.68% mIoU. Removing any individual source consistently degrades the results. In particular, removing VC reduces R1@0.5 by 7.20%, while removing Aux causes a 4.31% drop at R1@0.7. Removing Pose also decreases mIoU from 45.68% to 42.23%.

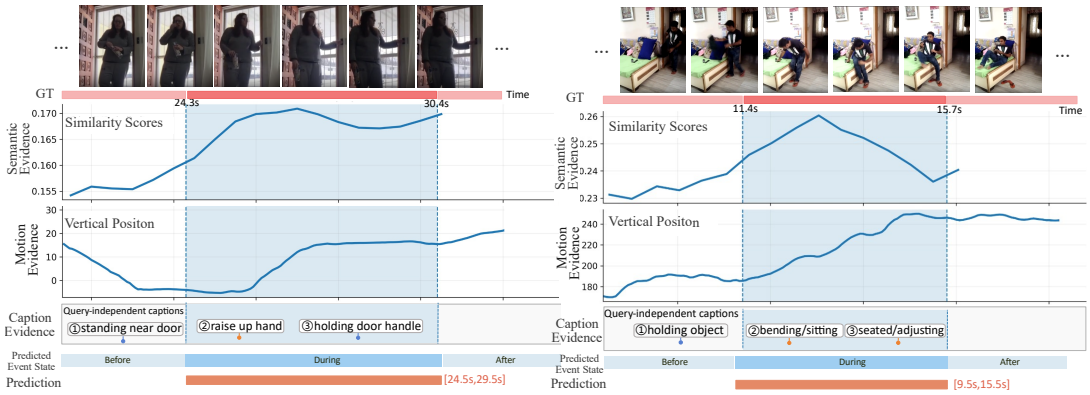
These results reveal that the three evidence sources play different but complementary roles. VC mainly measures semantic relevance between the query and video content. Pose captures local body changes associated with the target action, while Aux provides query-independent contextual information about human behavior, objects, and visual states. Their combination helps resolve cases where one type of evidence is ambiguous, leading to more reliable event-state reasoning and temporal localization.

Effect of Action-Aware Motion Strategies. We compare three pose modeling strategies: without Pose, using all body joints, and action-aware joint selection. The results are shown in Table 3. Simply introducing human pose does not necessarily improve temporal grounding. Using all joints reduces mIoU from 42.23% to 41.56% and also decreases all R1 metrics compared with the setting without Pose. This indicates that indiscriminately aggregating body motion may introduce irrelevant movement and obscure the motion patterns associated with the target event.

This issue is especially important because different actions are characterized by different body regions. For interaction actions, changes around the wrists, elbows, and shoulders are usually more informative, whereas locomotion actions are more closely related to the hips, knees, and ankles. Motion from unrelated joints may reflect natural body movement rather than the target action itself.

In contrast, Action-aware Pose selects joints according to the action type and consistently improves all metrics. Compared with w/o Pose, it improves R1@0.3, R1@0.5, R1@0.7, and mIoU by 5.01%, 5.96%, 3.04%, and 3.45%, respectively. Compared with All-joint Pose, the mIoU improvement reaches 4.12%. These results show that the benefit of human motion does not come from using more pose information, but from selecting motion cues that are relevant to the target action.·

Effect of Temporal Localization Strategies. We compare our adaptive boundary strategy with fixed windows of different lengths. The results are shown in Table 4. The fixed-window methods are sensitive to the window size. As the window increases from ±4s to ±8s, R1@0.3 rises from



(a) Turning on a light. (b) Sitting on a bed.

Figure 3: **Qualitative visualization of complementary event evidence.** We show two examples of (a) turning on a light and (b) sitting on a bed. For each example, we visualize VideoCLIP similarity, the vertical-position trajectory of the action-related body joint, query-independent video captions, and the predicted event-state sequence along the video timeline. The blue shaded region denotes the ground-truth interval, while the orange bar indicates the final prediction. These complementary cues support event-state reasoning and temporal localization.

60.37% to 68.76%. However, R1@0.5 and R1@0.7 drop to 38.42% and 14.71%, respectively. This shows that a fixed window cannot adapt well to different event durations and boundary changes.

In contrast, Adaptive Evidence Boundary achieves more balanced performance. It reaches 70.80% R1@0.3, 24.57% R1@0.7, and 45.68% mIoU. Compared with Peak ±6s, which gives the best mIoU among the fixed-window settings, our method improves mIoU by 3.23%. Although Peak ±4s is slightly higher on R1@0.5 by 0.29%, its overall performance is lower. These results show that adapting the boundaries to event-state changes provides better overall localization than using a fixed temporal window.

4.4 QUALITATIVE ANALYSIS

To further examine how complementary evidence contributes to event-state reasoning and temporal localization, we visualize two representative examples in Fig. 3. VideoCLIP similarity captures semantic relevance, human motion reflects action-related body changes, and query-independent captions describe local visual states. Their combination supports event-state prediction and final temporal localization.

In the light-turning example, semantic similarity remains high over several nearby frames, making the boundary ambiguous when used alone. Motion cues and captions provide additional evidence for the progression from BEFORE to DURING and then AFTER, leading to a prediction that closely overlaps the ground truth. In the sitting example, the vertical-position trajectory clearly captures the body transition, while semantic similarity and captions provide complementary context. Together, these cues produce a consistent BEFORE–DURING–AFTER progression and a temporal prediction that largely overlaps the ground truth.

5 CONCLUSION

We propose a training-free framework for video temporal grounding. It combines complementary evidence from video-text semantics, human motion, and query-independent video captions to model the temporal progression of the target event. A VLM reasons over these cues to infer event states, while continuous support scores measure how strongly each sampled frame supports the target event. Based on the resulting support sequence, we identify the most relevant temporal region and adaptively decode its start and end boundaries without additional training or task-specific supervision. Extensive experiments validate the effectiveness of the proposed method on multiple benchmarks. The results further show that combining event-state reasoning with complementary evidence provides more accurate and stable temporal localization than relying only on semantic relevance.

6 Reproducibility Statement

We provide detailed descriptions of STEER in the main paper and supplementary material. Additional implementation details on query and action representation, action-aware human motion, and video sampling and temporal alignment are provided in Appendices A.1, A.2, and A.5, respectively. The prompt templates for query decomposition, query-independent video captioning, and event-state reasoning are provided in Appendix A.3. We also report the model configurations, sampling settings, dataset splits, and evaluation protocol used in our experiments. All models remain frozen during inference, with no task-specific training or fine-tuning. The code will be released upon publication.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Guilhem Cheron, Ivan Laptev, and Cordelia Schmid. P-cnn: Pose-based cnn features for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3218–3226, 2015.

Srijan Das, Saurav Sharma, Rui Dai, Francois Bremond, and Monique Thonnat. Vpn: Learning video-pose embedding for activities of daily living. In *European Conference on Computer Vision*, pp. 72–90. Springer, 2020.

Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2969–2978, 2022.

Xiang Fang, Daizong Liu, Wanlong Fang, Pan Zhou, Yu Cheng, Keke Tang, and Kai Zou. Annotations are not all you need: A cross-modal knowledge transfer network for unsupervised temporal sentence grounding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 8721–8733. Association for Computational Linguistics, 2023.

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*, 2017.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18995–19012, 2022.

Chaohong Guo, Yihan He, Yongwei Nie, Fei Ma, Xuemiao Xu, and Chengjiang Long. T2sgrid: Temporal-to-spatial gridification for video temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3443–3454, 2026.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

Meghana Holla and Ismini Lourentzou. Commonsense for zero-shot natural language video localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 2166–2174, 2024.

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*, 2024.

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*, pp. 18908–18918, 2023.

Glenn Jocher and Jing Qiu. Ultralytics yolo11, 2024. URL `https://github.com/ultralytics/ultralytics`.

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*, pp. 706–715, 2017.

Jie Lei, Tamara L. Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. In *Advances in Neural Information Processing Systems*, 2021a.

Jie Lei, Tamara L. Berg, and Mohit Bansal. Qvhighlights: Detecting moments and highlights in videos via natural language queries. In *Advances in Neural Information Processing Systems*, volume 34, 2021b.

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*, 2023.

Daizong Liu, Xiaoye Qu, Yinzhen Wang, Xing Di, Kai Zou, Yu Cheng, Zichuan Xu, and Pan Zhou. Unsupervised temporal video grounding with deep semantic clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 1683–1691, 2022.

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*, pp. 5464–5473, 2024.

WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Dominick Reilly and Srijan Das. Just add π! pose induced video transformers for understanding activities of daily living. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18340–18350, 2024.

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*, 2024.

Anshul Shah, Shlok Mishra, Ankan Bansal, Jun-Cheng Chen, Rama Chellappa, and Abhinav Shrivastava. Pose and joint-aware action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3850–3860, 2022.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pp. 396–416. Springer, 2024.

Yongliang Wu, Xinting Hu, Yuyang Sun, Yizhou Zhou, Wenbo Zhu, Fengyun Rao, Bernt Schiele, and Xu Yang. Number it: Temporal grounding videos like flipping manga. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13754–13765, 2025.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6787–6800, 2021.

Xinyi Xu, Hongsong Wang, Guo-Sen Xie, Caifeng Shan, and Fang Zhao. Hitea: Hierarchical temporal alignment for training-free long-video temporal grounding. In *International Conference on Learning Representations*, 2026.

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.

Yifang Xu, Yunzhuo Sun, Benxiang Zhai, Ming Li, Wenxin Liang, Yang Li, and Sidan Du. Zero-shot video moment retrieval via off-the-shelf multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8978–8986, 2025. doi: 10.1609/aaai.v39i9.32971.

Shen Yan, Xuehan Xiong, Arsha Nagrani, Anurag Arnab, Zhonghao Wang, Weina Ge, David Ross, and Cordelia Schmid. Unloc: A unified framework for video localization tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Sunjae Yoon, Gwanhyeong Koo, Dahyun Kim, and Chang D. Yoo. Scanet: Scene complexity aware network for weakly-supervised video moment retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6543–6554. Association for Computational Linguistics, 2020a. doi: 10.18653/v1/2020.acl-main.585.

Jinglei Zhang, Yuanfan Guo, Rolandos Alexandros Potamias, Jiankang Deng, Hang Xu, and Chao Ma. Vtimecot: Thinking by drawing for video temporal grounding and reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24203–24213, 2025.

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, pp. 12870–12877, 2020b. doi: 10.1609/aaai.v34i07.6984.

Zhengbo Zhang, Mark He Huang, Zhigang Tu, and Ming-Hsuan Yang. Dart: Difficulty-adaptive routing for zero-shot video temporal grounding. In *European Conference on Computer Vision*, 2026.

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, pp. 3517–3525, 2022a.

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*, pp. 15555–15564, 2022b.

Minghang Zheng, Xinhao Cai, Qingchao Chen, Yuxin Peng, and Yang Liu. Training-free video temporal grounding using large-scale pre-trained models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Mohammadreza Zolfaghari, Gabriel L. Oliveira, Nima Sedaghat, and Thomas Brox. Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2923–2932, 2017.

Table 5: Action categories and corresponding human motion evidence.

Action Type	Typical Visual Focus	Role of Motion Evidence	Example
Interaction	Hands / upper body	Hand-arm motion and human-object interaction	open a cabinet
Locomotion	Lower body	Body displacement and leg motion	walk across a room
Stative	Task-dependent	Weak or subtle motion; used as auxiliary evidence	remain standing

A Appendix

A.1 Query and Action Representation

Based on the query Q and its structured event attributes $\mathrm{Attr}^{\mathrm{query}}$ defined in the main paper, we further derive the action category and primary visual focus for subsequent evidence construction, while retaining the relevant objects from the query attributes. The action category characterizes the dominant motion pattern of the target action, while the visual focus indicates the body region, object, or scene most relevant to the event. These attributes serve only as visual priors for evidence construction, whereas the original query Q and $\mathrm{Attr}^{\mathrm{query}}$ preserve the complete semantics of the target event.

Action category. Based on the dominant motion pattern of the core action Act, we categorize the target action into three types: interaction, locomotion, and stative. Interaction mainly involves human-object manipulation, locomotion describes clear body displacement, while stative refers to activities with limited or subtle body motion. The action category is used to guide the selection of human motion features and does not directly determine the temporal boundaries.

Target objects and visual focus. For queries involving specific objects, we retain the objects that are directly related to the target event. These objects provide auxiliary cues for understanding human-object interactions and object states. We further determine the primary visual focus according to the visual characteristics of the target event. The focus may correspond to the hands, upper body, lower body, body center, full body, object, or scene.

For example, interaction actions involving object manipulation mainly focus on the hands or upper body, while locomotion actions emphasize the lower body. For stative actions, the visual focus is determined by the specific event, and human motion is treated only as auxiliary evidence when body movement is weak. For object- or scene-centered queries, the focus is placed on the relevant object or scene, and human pose is not treated as the primary evidence.

Preserving complete query semantics. The auxiliary action and visual-focus attributes do not replace the original query. They only guide the evidence modules toward the action type, relevant objects, and body regions that are most informative for the target event. During subsequent event-state reasoning, the complete event semantics are preserved by the original query Q together with its structured attributes $\mathrm{Attr}^{\mathrm{query}}$.

In particular, the model does not treat the presence of a query-related person or object as sufficient evidence that the target event is occurring. It further verifies whether the target subject, action, object, and their relations are jointly consistent with the complete event. For human-object interaction events, both the target action and the corresponding interaction relation must be satisfied. For object-state events, the model instead determines whether the target object is in the state described by the query. The action-dependent visual focus and the role of human motion evidence are summarized in Table 5. For object- or scene-centered queries, the visual focus is set to the relevant object or scene, and human pose is not treated as the primary motion evidence.

A.2 Action-aware Human Motion Evidence

Human motion provides additional dynamic cues for distinguishing temporally different event states with similar semantic content. Rather than uniformly using all body keypoints, we select the body regions most relevant to the target action, thereby reducing noise from unrelated joint movements.

For each video, we first use a frozen human pose estimator to extract 17 body keypoints in the COCO format. The pose at the t-th observation is represented as

$$P_t = \{p_t^j = (x_t^j, y_t^j, c_t^j)\}_{j=1}^{17}, \tag{10}$$

where p_t^j denotes the j-th keypoint at observation t, (x_t^j, y_t^j) denotes its spatial coordinates, and c_t^j is the corresponding detection confidence. Low-confidence keypoints are ignored in subsequent motion computation. When multiple persons are detected in the same frame, we select the pose with the highest overall keypoint confidence as the motion representation. **Action-aware joint selection** Different actions are typically characterized by different body regions. For example, object-manipulation actions mainly involve wrist and arm movements, whereas locomotion actions rely more on changes in the hips, knees, and ankles. We therefore select action-relevant joints according to the action representation obtained in Appendix A.1

$$\mathcal{J}_{\mathrm{Act}} = g(c_Q, f_Q), \tag{11}$$

where c_Q and f_Q denote the action category and primary visual focus derived from the core action Act, respectively.

Specifically, interaction actions mainly focus on the wrists, elbows, and shoulders, while locomotion actions focus on the hips, knees, and ankles. For stative actions, human motion is generally less informative because the target event may involve only subtle or limited body movement. We therefore reduce the contribution of motion cues and rely more on semantic and visual-state evidence.

For implementation, the motion score s_t^{mot} follows the definition in the main paper. Only action-relevant keypoints with sufficient confidence at both adjacent observation positions are used for motion computation. This confidence-aware filtering reduces the influence of unreliable pose estimates, while action-aware joint selection suppresses motion from irrelevant body regions.

For actions that cannot be sufficiently characterized by local keypoint displacement alone, we additionally incorporate motion cues that reflect their characteristic body changes. For locomotion actions, lower-body motion is complemented by body-center displacement. For actions involving posture changes, body-center movement and vertical position changes provide additional cues. For stative actions, motion evidence is treated only as auxiliary information because salient body movement is usually limited.

Since the temporal resolution of the pose features may differ from the observation timeline defined in Appendix A.5, the motion scores are aligned to the unified observation grid using the same nearest-timestamp matching strategy. The aligned motion score at the i-th observation position is denoted by $\widetilde{s}_i^{\mathrm{mot}}$.

Table 6: Action-specific body regions, representative keypoints, and motion cues used for human motion evidence.

Action Type	**Visual Focus**	**Representative Keypoints**	**Main Motion Cue**
Interaction	Hands / upper body	Wrists, elbows, shoulders	Hand–arm displacement and interaction-related motion
Locomotion	Lower body	Hips, knees, ankles	Body displacement and leg motion
Stative	Task-dependent	–	Weak auxiliary motion cue

A.3 Prompt Template

STEER uses a frozen VLM for query analysis, query-independent video captions, and event-state reasoning. To improve reproducibility, we provide the core prompt templates used in these modules below. Variables enclosed in braces are replaced by the corresponding query- or video-specific information during inference. The output format is explicitly constrained to reduce response ambiguity.

A.3.1 QUERY-GUIDED EVENT DECOMPOSITION PROMPT

We use the following prompt to decompose the input query into a structured event representation. Given a natural-language query, the VLM extracts the subject, core action, relevant objects, relation, and action type. The resulting structured representation is used for subsequent evidence construction and event-state reasoning.

```
You are analyzing a text query for video temporal grounding.

Query:
"{QUERY}"

Identify the target event described by the query.

Return the following structured information:
- Subject: the actor or entity involved in the target event.
- Core action: the main action or state that should be temporally
    localized.
- Relevant objects: objects directly involved in the target event.
- Relation: the interaction or relation between the subject, action,
    and objects.
- Action type: the general motion category of the target event.

Return ONLY a valid JSON object:

{
  "subject": "...",
  "core_action": "...",
  "objects": ["..."],
  "relation": "...",
  "action_type": "..."
}
```

A.3.2 QUERY-INDEPENDENT CAPTION PROMPT

The following prompt is used to generate query-independent descriptions for each video observation. The VLM receives only the current visual content and is explicitly instructed not to use the text query or infer temporal context beyond the observation. This provides complementary visual evidence for subsequent event-state reasoning and reduces the influence of query-conditioned semantic bias.

```
    You are describing the current content of a video observation.

    Describe only what is directly visible in the current video
    content.

    Focus on:
    - the visible person or subject,
    - the current action or body state,
    - relevant objects and their visible states,
    - visible interactions between the subject and objects.

    Do NOT infer the target event.
    Do NOT assume what happened before or what will happen next.
    Do NOT use any text query.

    Return one concise factual description.
```

A.3.3 EVENT-STATE REASONING PROMPT

The following prompt is used to perform event-state reasoning at each observation position. The VLM jointly considers the target event, its structured representation, the current visual observation,

and the query-independent auxiliary caption. Rather than relying only on semantic relevance, the model explicitly checks whether the subject, action, objects, and their relations are satisfied. It then predicts one of four temporal states: BEFORE, DURING, AFTER, or UNCLEAR.

```
You are performing event-state reasoning for video temporal
grounding.

Target event:
"{QUERY}"

Structured event information:
- Subject: {SUBJECT}
- Core action: {ACTION}
- Relevant objects: {OBJECTS}
- Relation: {RELATION}
- Action type: {ACTION_TYPE}

Current visual observation:
{VIDEO_INPUT}

Additional event evidence:
- Video-text semantic evidence: {SEMANTIC_SCORE}
- Human motion evidence: {MOTION_EVIDENCE}
- Query-independent caption: "{AUXILIARY_CAPTION}"

Judge whether the COMPLETE target event is satisfied at the
current
observation. Do not make a positive judgment only because a
related
person or object is visible.

Determine the temporal state as one of:
BEFORE, DURING, AFTER, or UNCLEAR.

Return ONLY a valid JSON object:

{
  "subject_match": "YES/PARTIAL/NO",
  "action_match": "YES/PARTIAL/NO",
  "object_match": "YES/PARTIAL/NO/NONE",
  "relation_match": "YES/PARTIAL/NO",
  "target_match": "YES/PARTIAL/NO",
  "phase": "BEFORE/DURING/AFTER/UNCLEAR",
  "evidence": "brief concrete visual evidence"
```

A.4 Computational Complexity and Inference Efficiency

We analyze the computational characteristics of the main components of STEER in Table 7. VideoCLIP and pose features can be extracted offline and cached before inference, while query-independent auxiliary captions are also generated offline. During online inference, the main computational cost comes from VLM-based event-state reasoning at the sampled observation positions. Importantly, all components remain frozen and no task-specific training or fine-tuning is required.

Table 7: Computational characteristics of different components.

Component	Offline / Online	Trainable	Main Cost
VideoCLIP features	Offline / cached	No	Visual encoding
Pose features	Offline / cached	No	Pose extraction
Auxiliary captions	Offline	No	VLM inference
Event-state reasoning	Online	No	VLM inference

A.5 Implementation Details of Video Sampling and Temporal Alignment

To construct temporally aligned event evidence over the full video, we first establish a discrete observation grid along the video timeline. Since Charades-STA and ActivityNet Captions differ substantially in video duration and event length, we adopt different sampling resolutions while using the same temporal alignment strategy.

For Charades-STA, we uniformly sample the full video at 1 FPS, corresponding to one observation every second. This relatively dense sampling preserves fine-grained temporal changes in short actions. At each observation timestamp t_i, the corresponding frame is used for video-text semantic matching and VLM-based reasoning. Query-independent captions and human motion cues are aligned to the same temporal position, allowing different evidence sources to be jointly analyzed on a shared observation grid.

For ActivityNet Captions, we use a sparser sampling rate of 0.5 FPS, corresponding to one observation every 2 seconds. This reduces redundant computation on longer videos while maintaining coverage of the full video timeline. We further adopt a visual-change-aware VLM update strategy, where previous judgments can be reused when the visual content remains stable and updated when necessary.

Precomputed evidence sources, including VideoCLIP features, human pose features, and query-independent captions, may have timestamps that differ from the observation grid. We therefore align them by nearest-timestamp matching. For an observation timestamp t_i, the aligned feature from the m-th evidence source is selected as

$$\widetilde{x}_i^{(m)} = x_{j^*}^{(m)}, \qquad j^* = \arg\min_j \left| t_i - \tau_j^{(m)} \right|, \tag{12}$$

where $\tau_j^{(m)}$ denotes the timestamp of the j-th feature from the m-th evidence source.

This procedure maps semantic, motion, and caption evidence with different temporal resolutions onto a shared observation grid. All aligned evidence is subsequently used for event-state reasoning and temporal localization.

Table 8: Dataset statistics and sampling settings.

Benchmark	#Videos	#Queries	Video Len. (s)	Moment Len. (s)	Sampling Rate
Charades-STA	1.3k	3.7k	29	7.8	1.0 FPS
ANet-Captions	4.9k	17.1k	118	40.2	0.5 FPS

A.6 Limitations

Although STEER achieves competitive training-free temporal grounding performance, several limitations remain. First, the framework relies on frozen pretrained models, including the VLM, VideoCLIP, and the pose estimator. Therefore, errors in visual recognition, semantic matching, or pose estimation may propagate to subsequent event-state reasoning. This is particularly challenging when the target action is visually subtle, heavily occluded, or involves objects that are difficult to recognize.

Second, human motion evidence is most effective for actions with clear body movement. For events dominated by subtle object-state changes or scene-level transitions, pose cues may provide limited information. STEER reduces this issue by treating motion as complementary evidence rather than an independent decision signal, but such cases remain challenging.

Third, repeated or visually similar actions within the same video may produce multiple temporally separated regions with similar semantic and motion evidence. In these cases, distinguishing the target occurrence from other related actions can still be difficult, especially when the visual differences between event stages are small.

Finally, although STEER does not require task-specific training or fine-tuning, VLM-based reasoning introduces additional inference cost for long videos. Future work could investigate more efficient adaptive observation strategies, lightweight state reasoning, or stronger temporal representations to reduce computation while preserving localization accuracy.

A.7 AI USE STATEMENT

Generative AI tools were used to assist with language translation and polishing, literature retrieval, manuscript drafting and revision, and to provide feedback on research methodology, experimental design, and code development and debugging. All AI-assisted suggestions and outputs were carefully reviewed and verified by the authors. The authors conducted and validated all experiments and take full responsibility for the methodology, results, claims, and final content of this work.