


E3M: Zero-Shot Spatio-Temporal Video Grounding with Expectation-Maximization Multimodal Modulation

Peijun Bao¹, Zihao Shao², Wenhan Yang³, Boon Poh Ng¹, and Alex C. Kot¹

¹Nanyang Technological University

²Peking University ³Peng Cheng Laboratory


peijun001@e.ntu.edu.sg, zh.s@pku.edu.cn, yangwh@pcl.ac.cn

Abstract. Spatio-temporal video grounding aims to localize the spatio-temporal tube in a video according to the given language query. To eliminate the annotation costs, we make a first exploration to tackle spatio-temporal video grounding in a zero-shot manner. Our method dispenses with the need for any training videos or annotations; instead, it localizes the target object by leveraging pre-trained vision-language models and optimizing within the video and text query during the test time. To enable spatio-temporal comprehension, we introduce a multimodal modulation that integrates the spatio-temporal context into both visual and textual representation. On the visual side, we devise a context-based visual modulation that enhances the visual representation by propagation and aggregation of the contextual semantics. Concurrently, on the textual front, we propose a prototype-based textual modulation to refine the textual representations using visual prototypes, effectively mitigating the cross-modal discrepancy. In addition, to overcome the interleaved spatio-temporal dilemma, we propose an expectation maximization (EM) framework to optimize the process of temporal relevance estimation and spatial region identification in an alternating way. Comprehensive experiments validate that our zero-shot approach achieves superior performance in comparison to several state-of-the-art methods with stronger supervision. The code is available at <https://github.com/baopj/E3M>.

Keywords: Video grounding · Zero-shot learning · Vision-language model

1 Introduction

Grounding natural language in visual content is a fundamental technique to bridge the communication between humans and intelligent systems. In this work, we focus on a challenging visual grounding task named Spatio-Temporal Video Grounding (STVG) [36, 38]. Given a sentence query, as illustrated in Fig 1, the goal of STVG is to localize the target object spatially and temporally in an untrimmed video. This goes beyond aligning a global visual representation with

 Corresponding authors.

Query: A man in a suit walks into the room and sits down.



Fig. 1: 1) Given a natural language query, Spatio-Temporal Video Grounding (STVG) aims to localize the spatio-temporal video tube described by the query. We for the first time explore zero-shot STVG which eliminates the need for laborious manual annotations. 2) A significant challenge in zero-shot STVG lies in the necessity for joint spatio-temporal reasoning. Examples 2a and 2b illustrate typical grounding errors caused by failures in spatial and temporal reasoning, respectively.

a textual one, as it requires reasoning about detailed spatio-temporal visual representation and their association with natural language.

In recent years, the performance of STVG on benchmark datasets has been improved by the advancement of deep learning techniques [6, 8, 13, 15, 31] and the availability of massively annotated data [36, 38]. However, it is often expensive and time-intensive to collect the manual annotations, which consist of sentence queries and bounding box sequences. Moreover, the acquisition of training data proves to be inaccessible in numerous real-world applications (*e.g.* due to privacy concerns [18, 33]). To this end, in this paper, we propose to address the STVG task in a zero-shot manner, dispensing with the need for any ground-truth labels for training. To the best of our knowledge, this is the first attempt at zero-shot STVG in the literature.

Our core idea is to harness the zero-shot capabilities of large-scale pre-trained vision-language models (VLM) such as CLIP [20] for STVG’s input data, without the need for training on STVG-specific annotated data. Pretrained on millions of image-text pairs sourced from the internet, the VLMs have recently demonstrated impressive training-free performance across various computer vision tasks, including image-text matching [7], image grounding [26], and text-video generation [19, 35]. However, how to capitalize on the knowledge of such VLMs for spatio-temporal comprehension in untrimmed videos remains largely unsolved. A crucial challenge for STVG is that it demands not just image-level understanding but also the integrated reasoning of spatio-temporal semantics. For instance, as illustrated in Fig 1, discerning the spatio-temporal tubes “a

man in a suit walks into the room and sits down” requires reasoning across both temporal and spatial dimensions along adjacent frames.

To overcome these challenges, we propose a multimodal modulation algorithm to facilitate spatio-temporal comprehension, which augments both the visual and textual representation by integrating the spatio-temporal contexts. 1) *On the visual side*, we devise a context-based visual modulation aimed at enhancing the visual representation of object instances with contextual semantics. Specifically, we first propagate the spatial information of object instances to adjacent frames along the temporal dimension via a Kalman filter. Then we adaptively aggregate the visual features of these propagated contexts, thereby forming a spatio-temporal representation enriched with comprehensive contextual information. 2) *On the textual front*, we propose a prototype-based textual modulation that complements the textual representation with visual spatio-temporal semantics, thereby bridging their cross-modal discrepancy. We first identify semantics prototypes from spatio-temporal visual features that closely align with the sentence semantics. Subsequently, these semantics prototypes are employed to calibrate the textual features, effectively encapsulating the spatio-temporal information within the textual domain.

In addition, STVG inherently presents an intertwined spatio-temporal dilemma: spatial grounding in videos hinges on precise temporal grounding, as the videos comprise both positive and negative frames in relation to the query. Temporal missteps can inadvertently result in spatial inference within irrelevant frames, undermining the overall accuracy. Conversely, the effectiveness of temporal grounding also depends on spatial identification, given the presence of multiple irrelevant instances in a frame. To handle this intertwined dilemma, we regard the temporal relevance scores as the latent variables and introduce an expectation-maximization (EM) framework operating the temporal and spatial grounding in an alternating paradigm.

Our main contributions can be summarised as follows:

1. We propose a training-free algorithm to address the STVG task in a zero-shot manner, eliminating the necessity of any training videos or annotations. To the best of our knowledge, this is the first attempt at zero-shot STVG.
2. To facilitate spatio-temporal reasoning, we propose an Expectation Maximization Multimodal Modulation (E3M) algorithm, which comprises a context-based visual modulation, a prototype-based textual modulation, and an EM optimization in an iterative paradigm.
3. Extensive experiments verify that our zero-shot method outperforms a list of state-of-the-art methods using stronger supervision on two large-scale datasets *i.e.* HC-STVG [38] and VidSTG [36].

2 Related Works

Fully-Supervised STVG. This task of Spatio-Temporal Grounding (STVG) is first introduced by Zhang *et al.* [36] and Tang *et al.* [38]. The fully-supervised STVG approaches can be classified into two main categories: two-stage pipelines [36–

38] and one-stage pipelines [8,25,31]. Although delivering promising results, both types of fully-supervised methods heavily depend on an extensive collection of labor-intensive bounding box annotations to achieve their performance. This limits the scalability of these methods to real-world applications.

Weakly-Supervised STVG. In recent years, weakly supervised learning has achieved significant progress in various areas of computer vision [1, 3, 4, 28, 34]. Several recent works [6, 13] also tackle STVG in a weakly-supervised manner, which only uses coarse video-level descriptions for training. However, these methods still require paired video-language data, showing limited applicability in the open world. Additionally, obtaining training data can be difficult in a wide range of real-world applications, such as due to privacy concerns. In contrast, our zero-shot method obviates the requirement for any training videos, thereby significantly reducing the associated costs of data collection and annotation.

Video Understanding with CLIP. While pretrained on millions of image-text pairs, vision-language models such as CLIP [20] have recently drawn attention for video understanding [9, 16, 17, 21, 29, 30]. Wasim *et al.* [29] propose a multimodal prompt tuning for video recognition based on CLIP. Rasheed *et al.* [21] explore adapting image-level CLIP features to video data. However, these methods still rely on collecting massive training data and require time-consuming finetuning on the video data. Unlike them, our method seamlessly enables the CLIP models with the capability of spatio-temporal reasoning in a training-free manner.

Zero-Shot Multi-Modal Learning. Training deep learning models with manual supervision demands an extensive amount of annotated data. Therefore, zero-shot learning [5, 24, 26, 27, 32] has gradually drawn attention in the realm of multimodal understanding. A cross-modal hashing scheme using CLIP [20] model is developed in [32]. And Subramanian *et al.* [26] leverage CLIP model for zero-shot image grounding. To the best of our knowledge, we are the first to address the task of spatio-temporal video grounding in a zero-shot manner.

3 E3M for Zero-Shot STVG

3.1 Problem Formulation and Method Overview

Problem Formulation. Given an untrimmed video $V = \{f_t\}_{t=1}^{T_v}$ composed of T_v image frames and a sentence query Q , the goal of the Spatio-Temporal Video Grounding (STVG) is to localize the spatio-temporal tube $B = \{b_t\}_{t=t_s}^{t_e}$ described by S . Here b_t represents a bounding box in the t -th frame, t_s and t_e specify the starting and ending boundary of the retrieved object tube, respectively. Existing STVG approaches [25, 31, 36–38] have a significant drawback in that they necessitate extensive manual annotations for training. These annotations, including the spatio-temporal tubes B and the sentence queries S , are often expensive and time-consuming to collect, which limits their practicality in real-world applications. Moreover, acquiring training data can pose challenges in numerous scenarios, particularly where privacy concerns are involved. To tackle these limitations, we propose a training-free, zero-shot STVG approach that eliminates the requirement for ground-truth labels in the training phase.

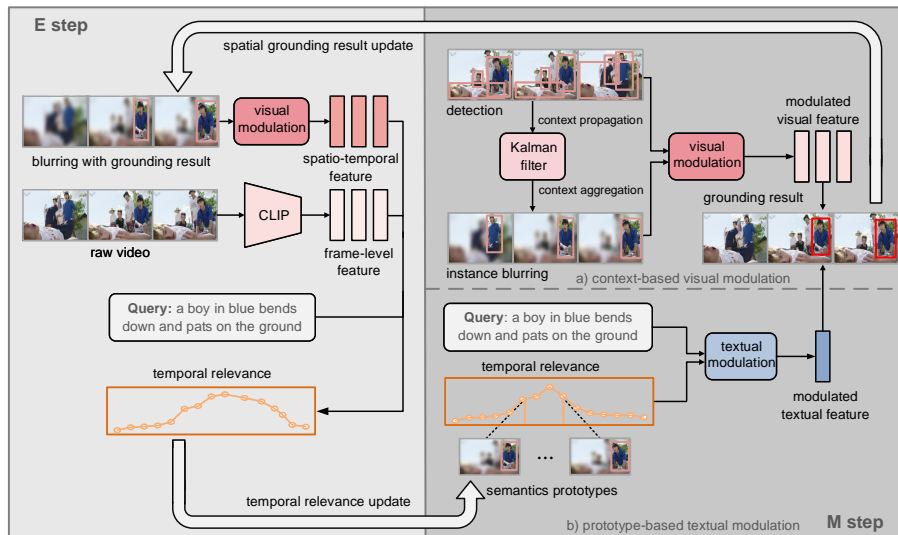


Fig. 2: An overview of Expectation-Maximization Multimodal Modulation (E3M). The E step aims to estimate temporal relevance scores between video frames and sentences, leveraging spatial grounding results from the M step. Based on the estimated temporal relevance scores, the M step optimizes spatial grounding, including: a) Context-based visual modulation, designed to facilitate reasoning among spatio-temporal contexts. This involves propagating and aggregating the spatio-temporal contexts of object instances using a Kalman filter. b) Prototype-based textual modulation to bridge the visual-textual discrepancy, where raw textual features are enriched with visual semantics prototypes.

Method Overview. Beyond merely comprehending the video frames at image level, a critical challenge for STVG involves the integrated reasoning of spatio-temporal semantics. For instance, as presented in Fig 1, to accurately localize the query “a man in a suit walks into the room and sits down” necessitates reasoning across both temporal and spatial dimensions in adjacent frames. Moreover, the spatial and temporal reasoning are inherently intertwined, with errors in one potentially propagating to the other.

To this end, as illustrated in Fig 2, we propose an Expectation-Maximization Multimodal modulation (E3M) algorithm for zero-shot STVG. Specifically, we introduce an Expectation-Maximization (EM) framework to optimize the temporal and spatial reasoning in an alternating paradigm. The temporal relevance scores, which represent the relevance between the sentence and each of the video frames, are regarded as latent variables. In the E step, we estimate the distribution of temporal relevance scores, leveraging the spatial grounding results from the M step. Based on the estimated temporal relevance scores, the M step then optimizes the spatial grounding results with a) context-based visual modulation and b) prototype-based textual modulation.

The context-based visual modulation operates *on the visual front*, aiming to enable visual reasoning among spatio-temporal contexts. The region information of object instances is first propagated along the temporal dimension via a Kalman filter. Then the visual features of the propagated regions are aggregated to enhance the object instances with the spatio-temporal contexts. Meanwhile, *on the textual side*, we devise the prototype-based textual modulation to bridge the visual-textual discrepancy. We first identify semantics prototypes from the visual features that closely align with the sentence. Afterwards, these textual features are recalibrated by incorporating these visual semantics prototypes.

3.2 M Step: Spatial Grounding with Multimodal Modulation

Assuming the distribution of the temporal relevance $\gamma \in \mathbb{R}^T$ is estimated in the Expectation (E) Step (detailed in subsection 3.3), this distribution represents the relevance between the T video frames and the query Q . Based on the temporal relevance distribution γ , the Maximization (M) step aims to optimize spatial grounding results with a joint spatio-temporal understanding. To achieve this, we devise a multimodal modulation algorithm that integrates context-based visual modulation and prototype-based textual modulation.

Context-based Visual Modulation. For each frame in the video, we first employ an off-the-shelf object detector [22] to detect object instances. Assuming that the t -th frame contains N_t object instances, their bounding boxes are denoted as b_t^i , where $i = 1 \dots N_t$.

To extract discriminative features for the i -th object instance, we first isolate the instance b_t^i from the t -th frame f_t by blurring its background. Next, we extract the visual feature v_t^i for f_t^i using the pretrained CLIP visual encoder ψ_{CLIP} , formulated as follows:

$$f_t^i = \xi(f_t, b_t^i), \quad (1)$$

$$v_t^i = \psi_{\text{CLIP}}(f_t^i). \quad (2)$$

where ξ represents the blurring operation, and f_t^i denotes the blurred version of f_t with the i -th object instance.

Note that the visual feature derived from Eq. 2 is computed on the single frame, and thus fails to capture the rich spatio-temporal semantics. To this end, we design a context-based visual modulation to facilitate visual reasoning among spatio-temporal contexts, which involves propagating and aggregating the spatio-temporal contexts of object instances using a Kalman filter.

Specifically, we first approximate the inter-frame displacements of each object instance using a linear constant velocity model and then can propagate the object instances $\{b_t^i\}$ to $2T$ neighbor frames using a Kalman filter [11]. We define the state s for each instance as $s = [u, v, a, r, \dot{u}, \dot{v}, \dot{a}]^T$. Here u and v represent the horizontal and vertical pixel locations of the instance’s center. And a and r signify the area and aspect ratio of the instance’s bounding box, respectively. The initial velocities $\dot{u}, \dot{v}, \dot{a}$ are set to zero. We then use the Kalman filter optimization [11]

to predict the correspondence probability of the object instances in adjacent frames.

With the correspondence probability derived from the Kalman filter, we finalize the association of the object instances through bipartite graph matching, a task efficiently solved by Hungarian algorithm [12]. The assignment cost matrix for the Hungarian algorithm is computed on the Intersection-over-Union (IoU) metric applied to the bounding boxes of the object instances. In the t' -th frame, assuming the bounding box associated with the i -th object in the t -th frame is denoted as $b_{t'}^i$ (with $b_{t'}^i$ possibly being empty if no bounding box is assigned).

After the propagation of object instances to $2T$ neighbors along the temporal dimension, we can obtain the spatio-temporal contexts for the i -th object instance, namely a list of bounding boxes $\{b_{t'}^i\}$ over the $(2T+1)$ frames that are associated with the i -th object instance, where $t-T \leq t' \leq t+T$ (*i.e.* covering both T frames to the left and right neighbors of the t -th frame). Subsequently, we modulate the visual features \tilde{v}_t^i for the i -th object instance at the t -th frame by aggregating visual feature of the spatio-temporal contexts $\{b_{t'}^i\}$, written as:

$$\tilde{v}_t^i = \delta(\{v_{t'}^i\}), \quad t-T \leq t' \leq t+T, \quad (3)$$

where δ represents the average pooling function over the $2T+1$ frames, and $v_{t'}^i$ is computed using Eq. 2 with the bounding box $b_{t'}^i$ on the t' -th frame.

Prototype-based Textual Modulation. Given the sentence query Q , we can extract the sentence feature q using the pretrained CLIP text encoder ϕ_{CLIP} as follows:

$$q = \phi_{\text{CLIP}}(Q). \quad (4)$$

Consider a composite event described in the sentence query, involving two visual stages. In the first stage, the visual regions exhibit a high similarity to the sentence feature q as defined in Eq. 4. However, the second visual stage might be mistakenly identified as a negative sample to the sentence query due to temporal difference, even though both stages are integral to the same event.

To handle this challenge posed by the visual-textual semantics discrepancy, we propose a novel prototype-based textual modulation that refines the textual representation by capitalizing on complementary semantics prototypes. Here, we expect that the semantics prototypes satisfy the following two properties. 1) Exemplarity: each prototype should align closely with the semantic content of the sentence. 2) Diversity: the prototypes should represent various stages of actions, so they can capture the evolving nature of visual semantics and complement the discrepancy in the textual representation.

To accomplish this, we first determine the t^* -th frame in the video, such that the visual instances at the t^* -th frame align most closely with the sentence semantics, formulated as:

$$t^* = \operatorname{argmax}_t \gamma_t \quad (5)$$

Subsequently, we calculate the instance similarities between the sentence feature q and the modulated visual features $\{\tilde{v}_{t^*}^i\}_{i=1}^{N_{t^*}}$:

$$s_{t^*}^i = \frac{q^T \tilde{v}_{t^*}^i}{\|q\| \cdot \|\tilde{v}_{t^*}^i\|}, \quad i = 1, \dots, N_{t^*} \quad (6)$$

Algorithm 1 Prototype-based Textual Modulation.

Input: Sentence query q , temporal relevance scores $\gamma_1, \gamma_2, \dots, \gamma_T$, modulated visual features $\{\tilde{v}_t^i\}_{i=1}^{N_t}$, window size W , number of prototypes K

Output: List of prototypes $P = \{p_j\}_{j=1}^K$ and modulated textual representation \hat{q}

- 1: Compute the textual representation q through Eq. 4
- 2: $P \leftarrow \emptyset$
- 3: **while** $|P| < K$ **do**
- 4: $t^* \leftarrow \operatorname{argmax}_t(\gamma_t)$ \triangleright Find the temporal index with the highest temporal relevance score
- 5: Compute $s_{t^*}^i$ and i^* by Eq.6 and Eq.7
- 6: **for** $t = \max(1, t^* - W) \rightarrow \min(T, t^* + W)$ **do**
- 7: $\gamma_t \leftarrow 0$ \triangleright Suppress scores in the neighborhood
- 8: **end for**
- 9: $P \leftarrow P \cup \{p_{i^*}\}$
- 10: **end while**
- 11: Modulate textual representation q by $P = \{p_j\}_{j=1}^K$ as \hat{q} via Eq. 9
- 12: **return** P and \hat{q}

where N_{t^*} denotes the number of the visual instances at the t^* -th frame. Then the visual instance at the t^* -th frame that showcases the maximum $s_{t^*}^i$ value is designated the semantic prototype, formulated as:

$$i^* = \operatorname{argmax}_i s_{t^*}^i \quad (7)$$

Instead of selecting a single semantic prototype, we propose formulating multiple semantic prototypes to enrich the diversity. To achieve this, we introduce a suppression strategy to suppress the temporal relevance scores within the temporal neighbors of t^* (bounded by a window size of W), written as:

$$\hat{\gamma}_t = \begin{cases} 0, & \text{if } \max(1, t^* - W) \leq t \leq \min(T, t^* + W), \\ \gamma_t, & \text{otherwise.} \end{cases} \quad (8)$$

This strategy helps mitigate redundancy by avoiding the repetitive selection of analogous semantic prototypes.

Based on the suppressed $\hat{\gamma}_t$, we subsequently repeat the above process and compute Eq. 5 to 8, until obtaining K semantics prototypes, represented as $\{p_j\}_{j=1}^K$. Based on K semantics prototypes $\{p_j\}_{j=1}^K$, we modulate the textual representation with respect to these prototypes as:

$$\hat{q} = q + \sum_{j=1}^K \psi_{\text{CLIP}}(p_j) \quad (9)$$

where ψ_{CLIP} is the pretrained CLIP visual encoder. The prototype-based textual modulation is summarized in the Algorithm 1.

Spatio-temporal Grounding Results. To identify the object instances that match the semantics of the sentence query Q , we first compute the cosine

similarities between the modulated sentence feature \hat{q} and the modulated visual feature \tilde{v}_t^i for the N_t object instances at each t frame, formulated as

$$\hat{s}_t^i = \frac{\hat{q}^T \tilde{v}_t^i}{\|\hat{q}\| \cdot \|\tilde{v}_t^i\|}, \quad i = 1 \dots N_t, \quad t = \tau_s \dots \tau_e \quad (10)$$

where τ_s, τ_e represent the start and end point of temporal prediction results obtained in the E step (detailed in subsection 3.3).

Finally, for the t -th frame, the i_t^* -th object instance is selected as the prediction result, where i_t^* is defined as:

$$i_t^* = \operatorname{argmax}_i \hat{s}_t^i \quad (11)$$

The corresponding bounding boxes are:

$$B = \{b_t^{i_t^*}\}_{t=\tau_s}^{\tau_e}. \quad (12)$$

3.3 E Step: Temporal Relevance Estimation

For each frame in the video, we first extract the frame-level feature v_t using the visual encoder ψ_{CLIP} of the CLIP model as

$$v_t = \psi_{\text{CLIP}}(f_t), \quad (13)$$

where f_t is the t -th frame of the video and $1 \leq t \leq T_v$. However, directly computing Eq. 13 as a visual representation for the frame f_t encounters two limitations. Firstly, the presence of multiple irrelevant instances within a frame can act as noise, detracting from effective semantic matching. Secondly, the lack of contextual information along the temporal dimension is a significant barrier to temporal reasoning.

To this end, we first exploit the result of spatio-temporal grounding Eq.12 to blur the irrelevant instances and highlight the positive object instances. Subsequently, the context-based visual modulation is leveraged to extract a spatio-temporal visual representation for the t -th frame as v_t^* using Eq. 3. Then the temporal relevance score γ_t between the t -frame and the language query Q can be computed as

$$\gamma_t = \frac{q^T v_t^*}{\|q\| \cdot \|v_t^*\|} + \frac{q^T v_t}{\|q\| \cdot \|v_t\|}, \quad 1 \leq t \leq T_v, \quad (14)$$

where q is the textual representation as defined in Eq. 4.

To localize the target temporal moment based on temporal relevance score γ_t , we first apply the K-Means algorithm to the visual feature of each frame to formulate C clusters $\mathcal{T} = \{(\tau_s^c, \tau_e^c)\}_{c=1}^C$, where each cluster (τ_s^c, τ_e^c) represents an atomic temporal event with the start and end point of τ_s^c and τ_e^c . Here to compute the visual feature for the t -th frame, we first compute the average of the frame-level visual feature and instance-level spatio-temporal feature as \hat{v}_t' ,

then we concatenate the normalized version of \hat{v}'_t and the timepoint t as the final visual feature v'_t , formulated as:

$$\hat{v}'_t = \frac{v_t + v_t^*}{2} \quad (15)$$

$$v'_t = \text{concat} \left(\frac{\hat{v}'_t}{\|\hat{v}'_t\|}, \frac{t}{T_v} \right) \quad (16)$$

where concat is the concatenation operation.

After obtaining the atomic event set \mathcal{T} , we predict the final temporal boundary $\tau = (\tau_s, \tau_e)$ by merging the atomic events relevant to Q . Firstly, we compute the matching score α^c for each atomic temporal event as

$$\alpha^c = \frac{1}{l^c} \sum_{t=\tau_s^c}^{\tau_e^c} \gamma_t \quad (17)$$

where l^c is the temporal length of the c -th atomic event.

Subsequently, we initialize the temporal boundary prediction τ as the atomic temporal event $c_* = \text{argmax}_c \alpha^c$ with the largest matching score. Then we gradually merge the atomic events c' that are adjacent to c_* if

$$\alpha^{c_* \cup c'} > \beta \alpha^{c_*} \quad (18)$$

where $c_* \cup c'$ is the union of the temporal boundary for c_* and c' and β is a predefined hyperparameter. We continuously update c_* until there is no adjacent atomic event eligible to Eq. 18.

4 Experiments

4.1 Datasets and Metrics

Datasets. To evaluate our zero-shot model, we adopt two widely used benchmarks for the STVG task, *i.e.* HC-STVG [38] and VidSTG [36]. **HC-STVG** dataset is collected from movie scenes and contains 5,660 untrimmed videos in multi-person scenes. This dataset is challenging for spatio-temporal grounding because 57.2% of video clips contain more than 3 people. There are 1,160 video-sentence pairs in the testing split. **VidSTG** dataset comprises a total of 99,943 sentences describing 80 types of objects appearing in 6,924 untrimmed videos. The testing subset has 10,303 video-sentence pairs. On these two datasets, our zero-shot method tackles the STVG in a training-free manner, which dispenses with the training videos and annotations, and directly performs inference on the testing videos and queries.

Metrics. We follow the standard evaluation protocol [36, 38] and use m_vIoU, and vIoU@R to assess the performance of spatio-temporal grounding. Specifically, let S_i, S_u denote the intersection and union between the predicted and ground-truth frames. The vIoU is calculated by $\frac{1}{|S_u|} \sum_{t \in S_i} \text{IoU}(\hat{b}_t, b_t)$, where \hat{b}_t

and b_t denote the detected and ground-truth bounding box at frame t respectively. The m_vIoU score represents the $vIoU$ score averaged over all testing videos. And $vIoU@R$ denotes the proportion of data samples in the testing subset with $vIoU$ greater than the threshold R where $R \in \{0.3, 0.5\}$.

4.2 Implementation Details

The proposed method for zero-shot STVG is training-free and performs grounding via directly optimizing within the input video and query during test time. We exploit Faster-RCNN [22] pretrained on COCO [14] (with the backbone of ResNet-50 [10]) as the object detector. The RGB frames of the videos are resampled with a length T_v of 128. We use the CLIP model [20] with the backbone of ViT-B/32 as the pretrained VLM. The context size $2T + 1$ for visual modulation is set to be 21 *i.e.* $T = 10$. The prototype number K for textual modulation is set to be 3 and W is set to be 3. The EM epoch number is set to 2. More implementation details can be referred to the supplement.

4.3 Performance Comparison

Existing STVG approaches are generally divided into two categories: fully and weakly supervised learning. Both types depend on extensive datasets with corresponding annotations for training. Specifically, fully-supervised methods require comprehensive annotations of sentence queries and bounding box sequences, whereas weakly-supervised ones need videos paired with sentence queries. In contrast, our proposed E3M framework dispenses the need for training videos or annotations, introducing a zero-shot approach to STVG. As we are the first to explore zero-shot STVG, we further carefully adapt the state-of-the-art zero-shot image grounding method *i.e.* ReCLIP [26] and RedCircle [24] to STVG (more details can be referred to the supplementary material) for comparative analysis. Table 1 presents a comparison of our E3M against these advanced methods on the HC-STVG and VidSTG datasets, which illustrates the following findings.

1) E3M outperforms weak supervised methods. Without specific training on STVG data, our E3M approach exceeds the state-of-the-art weakly-supervised method such as Vis-Ctx [23] and Winner [13] on HC-STVG and VidSTG (Declarative Sentence) significantly. For instance, on HC-STVG, the metric $vIoU@0.5$ of E3M is more than 75% higher than Winner. And our method achieves a similar grounding accuracy to Winner on VidSTG (Interrogative Sentence). This shows the promise to exploit the pretrained vision-language models for STVG, thereby bypassing the intensive training on annotated STVG data.

2) E3M’s superiority in zero-shot approaches. The innovative use of expectation-maximization multimodal modulation which enhances spatio-temporal understanding in video data, allows E3M to significantly outperform other leading zero-shot methods, including ReCLIP [26] and RedCircle [24]. This underscores E3M’s advanced capabilities in a zero-shot context.

3) E3M competes with some fully-supervised methods. Our E3M method also shows comparable results to some fully-supervised methods. For

Table 1: Performance comparisons of the state-of-the-art on the HC-STVG and VidSTG test set. Full, Weak, and ZS denote fully-supervised, weak-supervised, and zero-shot learning settings respectively.

Sup	Method	HC-STVG			VidSTG (Declarative Sentence)			VidSTG (Interrogative Sentence)		
		m_vIoU	vIoU@0.3	vIoU@0.5	m_vIoU	vIoU@0.3	vIoU@0.5	m_vIoU	vIoU@0.3	vIoU@0.5
Full	STGVT [38]	18.15	26.81	9.48	21.62	29.80	18.94	—	—	—
	STVGBert [25]	20.42	29.37	11.31	23.97	30.91	18.39	22.51	25.97	15.95
	TubeDETR [2]	32.40	49.80	23.50	30.40	42.50	28.20	25.70	35.70	23.20
	STCAT [8]	35.09	57.67	30.09	33.14	46.20	32.58	28.22	39.24	26.63
Weak	AWGU [6]	8.20	4.48	0.78	8.96	7.86	3.10	8.57	6.84	2.88
	Vis-Ctx [23]	9.76	6.81	1.03	9.34	7.32	3.34	8.69	7.18	2.91
	Winner [13]	14.20	17.24	6.12	11.61	14.12	7.40	10.23	11.96	5.46
ZS	Random	3.91	0.86	0.09	3.13	0.40	0.00	3.00	0.44	0.02
	RedCircle [24]	9.15	7.76	1.55	8.56	7.61	0.93	9.04	8.51	1.47
	ReCLIP [26]	14.36	18.28	4.91	14.21	17.54	7.86	8.36	7.96	2.34
	E3M (Ours)	19.11	29.40	10.60	16.21	20.47	11.91	10.61	12.20	5.44

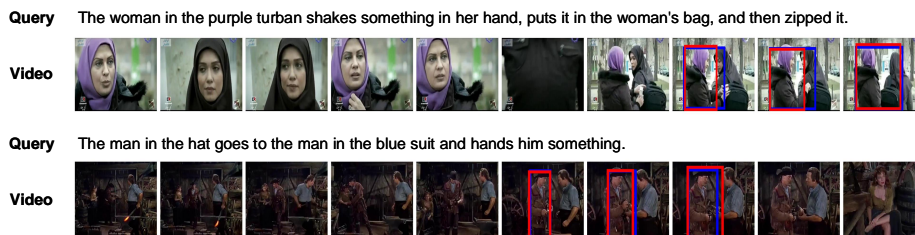


Fig. 3: Qualitative analysis on HC-STVG datasets. The red boxes denote the prediction results of our zero-shot methods, while the blue ones represent the ground truth.

instance, on HC-STVG and VidSTG (Declarative Sentence), the performance of our E3M is comparable to the fully-supervised method STGVT [38], despite STGVT employing intensive training on the fully-annotated data.

4.4 Qualitative Analysis

Fig. 3 presents the qualitative results of our E3M on HC-STVG dataset. It is noteworthy that E3M, even in a zero-shot setting, successfully identifies the target spatio-temporal tubes described by complicated sentences amidst a cluttered background. This includes challenging scenarios where the target tube occupies only a brief duration and the video frames contain unrelated objects and people.

4.5 Ablation Studies

To investigate the effectiveness of the proposed algorithms, here we conduct ablation studies on the HC-STVG dataset.

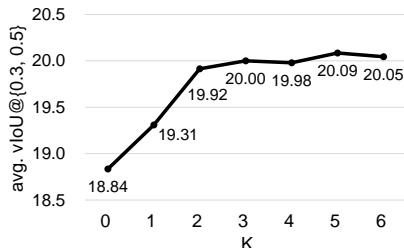
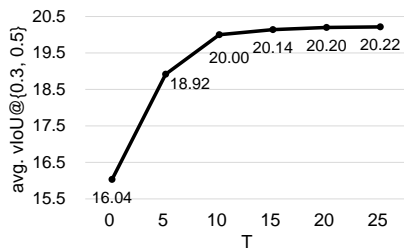
The effect of each module. Table 2 investigates the effect of the main modules by evaluating their impact when each is removed. 1) Visual modulation: To exclude visual modulation, the visual features are used directly as defined in Eq. 2 without context-based modulation. 2) Textual modulation: we use the original

Table 2: The effect of each module *i.e.* visual modulation (visual mod.), textual modulation (textual mod.), and EM.

visual mod.	textual mod.	EM	m_vIoU	vIoU@0.3	vIoU@0.5
✓	✓	✓	19.11	29.40	10.60
✓	✓	✗	17.21	25.26	8.97
✓	✗	✓	18.76	28.36	9.31
✗	✓	✓	17.05	24.66	7.41
✗	✗	✓	16.41	22.41	5.95
✗	✗	✗	15.20	19.91	5.00

Table 3: The impact of each module by replacing its variant.

visual and textual mod.	m_vIoU	vIoU@0.3	vIoU@0.5
full	19.11	29.40	10.60
full w/ short context	17.62	27.41	7.93
full w/ single prototype	18.80	28.79	9.83
full w/ past context	18.27	27.33	9.57
full w/ future context	18.09	27.16	9.14

**Fig. 4:** Ablation study on context size T . **Fig. 5:** Ablation study on proto. size K .

textual features as described in Eq. 4 to replace textual modulation. 3) EM algorithm: To drop the EM algorithm, we directly apply the temporal grounding and then use the initial temporal localization results for spatial grounding. The findings reveal that excluding any of these modules (denoted by a cross symbol in the respective column) significantly diminishes the overall performance.

Table 3 provides a deeper analysis of the impact of these modules by substituting them with different variants: 1) full with short context: configuring the context size for visual modulation to 1, *i.e.* only considering the past and future one frame, 2) full with single prototype: using solely one prototype for textual modulation. 3) full with past context: exclusively utilizing past context for visual modulation, 4) full with future context: only using future context for visual modulation. We observe that substituting the long-range visual context with the short context in the visual modulation leads to a noticeable decrease in grounding accuracy. However, it still significantly outperforms the variant lack-

Table 4: The impact of backbone architecture.

backbone	m_vIoU	vIoU@0.3	vIoU@0.5
RN-50	18.29	26.72	10.30
RN-50X4	18.66	28.28	10.43
ViT-B/32	19.11	29.40	10.60

ing visual modulation entirely. This confirms the critical role of visual context in facilitating spatio-temporal comprehension. Furthermore, when dropping either the past or future context, there is about a point decrease in each metric, which shows both the past and future context should be modeled. Replacing multiple prototypes with a single prototype in the textual modulation results in a performance decline of approximately one point, underscoring the need for textual modulation with multiple semantics prototypes.

Context size T for visual modulation. The visual modulation enhances the visual representation with $2T + 1$ visual contexts, with the hyperparameter T playing a crucial role in visual modulation. Fig. 4 presents the impact of T on the performance of STVG, where we use the average of “vIoU@ R ” with $R = \{0.3, 0.5\}$ as the evaluation metric. The grounding accuracy improves gradually as T increases when T is less than 10. After T is larger than 10, the model’s accuracy reaches a point of saturation.

Prototype size K for textual modulation. Fig. 5 studies the influence of hyperparameter K which denotes the prototype size for textual modulation. The grounding accuracy first increases as K becomes larger, which verifies the effectiveness of textual modulation with multiple semantics prototypes. When $K > 3$, the performance gradually saturates, showcasing that when setting $K = 3$ can provide sufficient spatio-temporal information for textual modulation.

Backbone architecture. Table 4 compares the performance of zero-shot STVG using CLIP of different backbone architecture *i.e.* RN50, RN50X4, and ViT-B/32. The performances of zero-shot STVG using any of them are satisfactory and evidently surpasses that of the state-of-the-art weakly-supervised methods in Table 1. This shows the adaptability of the proposed E3M method.

5 Conclusion

This paper explores spatio-temporal video grounding in a zero-shot setting for the first time. To enable spatio-temporal reasoning, we propose an Expectation-Maximization Multimodal Modulation (E3M) algorithm which integrates both prototype-based textual modulation and context-based visual modulation. Our E3M dispenses with the need for any training videos or annotations and localizes the target object by optimizing within the video and text query during the test time. Experiments demonstrate that our zero-shot method achieves competitive results in comparison to several methods that rely on stronger supervision on large-scale benchmarks.

Acknowledgements

This work was carried out at Rapid-Rich Object Search (ROSE) Lab, School of Electrical & Electronic Engineering, Nanyang Technological University. This research is supported by the NTU-PKU Joint Research Institute (a collaboration between the Nanyang Technological University and Peking University that is sponsored by a donation from the Ng Teng Fong Charitable Foundation).

References

1. Ahn, J., Kwak, S.: Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In: CVPR (2018) 4
2. Antoine Yang, Antoine Miech, J.S.I.L., Schmid, C.: Tubedetr: Spatio-temporal video grounding with transformers. In: CVPR (2022) 12
3. Bao, P., Shao, Z., Yang, W., Ng, B.P., Er, M.H., Kot, A.C.: Omnipotent distillation with llms for weakly-supervised natural language video localization: When divergence meets consistency. In: AAAI (2024) 4
4. Bao, P., Xia, Y., Yang, W., Ng, B.P., Er, M.H., Kot, A.C.: Local-global multi-modal distillation for weakly-supervised temporal video grounding. In: AAAI (2024) 4
5. Bao, P., Yang, W., Ng, B.P., Er, M.H., Kot, A.C.: Cross-modal label contrastive learning for unsupervised audio-visual event localization. In: AAAI (2023) 4
6. Chen, J., Bao, W., Kong, Y.: Activity-driven weakly-supervised spatio-temporal grounding from untrimmed videos. In: ACM MM (2020) 2, 4, 12
7. Jiang, K., He, X., Xu, R., Wang, X.E.: Comclip: Training-free compositional image and text matching. In: NAACL (2024) 2
8. Jin, Y., Li, Y., Yuan, Z., Mu, Y.: Embracing consistency: A one-stage approach for spatio-temporal video grounding. In: NeurIPS (2022) 2, 4, 12
9. Ju, C., Han, T., Zheng, K., Zhang, Y., Xie, W.: Prompting visual-language models for efficient video understanding. In: ECCV (2022) 4
10. Kaiming He, Xiangyu Zhang, S.R., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 11
11. Kalman, R.E.: A new approach to linear filtering and prediction problems. *Journal of Basic Engineering* (2011) 6
12. Kuhn, H.W.: The hungarian method for the assignment problem. *Naval Research Logistics* (1955) 7
13. Li, M., Wang, H., Zhang, W., Miao, J., Zhao, Z., Zhang, S., Ji, W., Wu, F.: Winner: Weakly-supervised hierarchical decomposition and alignment for spatio-temporal video grounding. In: CVPR (2023) 2, 4, 11, 12
14. Lin, T.Y., Maire, M., Belongie, S.J., et al.: Microsoft coco: Common objects in context. In: ECCV (2014) 11
15. Lin, Z., Tan, C., Hu, J., Jin, Z., Ye, T., Zheng, W.: Collaborative static and dynamic vision-language streams for spatio-temporal video grounding. In: CVPR (2023) 2
16. Liu, R., Huang, J., Li, G., Feng, J., Wu, X., Li, T.H.: Revisiting temporal modeling for clip-based image-to-video knowledge transferring. In: CVPR (2023) 4
17. Luo, D., Huang, J., Gong, S., Jin, H., Liu, Y.: Towards generalisable video moment retrieval: Visual-dynamic injection to image-text pre-training. In: CVPR (2023) 4
18. Mirshghallah, F., Taram, M., Vepakomma, P., Singh, A., Raskar, R., Esmaeilzadeh, H.: Privacy in deep learning: A survey. arXiv preprint arXiv:2004.12254 (2020) 2

19. Peng, B., Chen, X., Wang, Y., Lu, C., Qiao, Y.: Conditionvideo: Training-free condition-guided text-to-video generation. In: AAAI (2024) [2](#)
20. Radford, A., Kim, J.W., Hallacy, C., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021) [2](#), [4](#), [11](#)
21. Rasheed, H.A., Khattak, M.U., Maaz, M., Khan, S., Khan, F.S.: Fine-tuned clip models are efficient video learners. In: CVPR (2023) [4](#)
22. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. TPAMI (2015) [6](#), [11](#)
23. Shi, J., Xu, J., Gong, B., Xu, C.: Not all frames are equal: Weakly-supervised video grounding with contextual similarity and visual clustering losses. In: CVPR (2019) [11](#), [12](#)
24. Shtedritski, A., Rupprecht, C., Vedaldi, A.: What does clip know about a red circle? visual prompt engineering for vlms. In: CVPR (2023) [4](#), [11](#), [12](#)
25. Su, R., Xu, Q.Y.D.: Stvgbert: A visual-linguistic transformer based framework for spatio-temporal video grounding. In: ICCV (2021) [4](#), [12](#)
26. Subramanian, S., Merrill, W., Darrell, T., Gardner, M., Singh, S., Rohrbach, A.: Reclip: A strong zero-shot baseline for referring expression comprehension. In: ACL (2022) [2](#), [4](#), [11](#), [12](#)
27. Tiong, A.M.H., Li, J., Li, B.A., Savarese, S., Hoi, S.C.H.: Plug-and-play vqa: Zero-shot vqa by conjoining large pretrained models with zero training. In: EMNLP Findings (2022) [4](#)
28. Wang, Y., Zhang, J., Kan, M., Shan, S., Chen, X.: Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In: CVPR (2020) [4](#)
29. Wasim, S.T., Naseer, M., Khan, S., Khan, F.S., Shah, M.: Vita-clip: Video and text adaptive clip via multimodal prompting. In: CVPR (2023) [4](#)
30. Xing, J., Wang, M., Hou, X., Dai, G., Wang, J., Liu, Y.: Multimodal adaptation of clip for few-shot action recognition. In: CVPR (2023) [4](#)
31. Yang, A., Miech, A., Sivic, J., Laptev, I., Schmid, C.: Tubedetr: Spatio-temporal video grounding with transformers. In: CVPR (2022) [2](#), [4](#)
32. Yu, H., Ding, S., Li, L., Wu, J.: Self-attentive clip hashing for unsupervised cross-modal retrieval. In: MM Asia (2022) [4](#)
33. Zhang, G., Liu, B., Zhu, T., Zhou, A., Zhou, W.: Visual privacy attacks and defenses in deep learning: a survey. Artificial Intelligence Review (2022) [2](#)
34. Zhang, R., Wang, S., Duan, Y., Tang, Y., Zhang, Y., Tan, Y.P.: Hoi-aware adaptive network for weakly-supervised action segmentation. In: IJCAI (2023) [4](#)
35. Zhang, Y., Wei, Y., Jiang, D., Zhang, X., Zuo, W., Tian, Q.: Controlvideo: Training-free controllable text-to-video generation. ArXiv (2023) [2](#)
36. Zhu Zhang, Zhou Zhao, Y.Z.Q.W.H.L., Gao, L.: Where does it exist: Spatio-temporal video grounding for multi-form sentences. In: CVPR (2020) [1](#), [2](#), [3](#), [4](#), [10](#)
37. Zhu Zhang, Zhou Zhao, Z.L.B.H., Yuan, J.: Object-aware multi-branch relation networks for spatio-temporal video grounding. In: IJCAI (2021) [3](#), [4](#)
38. Zongheng Tang, Yue Liao, S.L.G.L.X.J.H.J.Q.Y., Xu, D.: Human-centric spatio-temporal video grounding with visual transformers. In: TCSVT (2021) [1](#), [2](#), [3](#), [4](#), [10](#), [12](#)

Supplements for E3M: Zero-Shot Spatio-Temporal Video Grounding with Expectation-Maximization Multimodal Modulation

Peijun Bao¹✉, Zihao Shao², Wenhan Yang³, Boon Poh Ng¹, and Alex C. Kot¹

¹Nanyang Technological University

²Peking University ³Peng Cheng Laboratory

peijun001@e.ntu.edu.sg, zh.s@pku.edu.cn, yangwh@pcl.ac.cn

1 Additional Ablation Studies

Convergence of EM Iteration. Fig. 1 investigates the convergence of EM iteration on the HC-STVG dataset, by evaluating the grounding accuracy as the EM epoch number increases. We employ the average vIoU@R where $R = \{0.3, 0.5\}$ as the evaluation metric. The grounding accuracy gradually improves as the EM algorithm iterates. Remarkably, the algorithm converges after just 2 epochs, with the average vIoU@R stabilizing at approximately 20.

Influence of Neighbor Size W . In Fig 2, we present ablation studies on the hyperparameter W , which denotes the neighbor size for suppression in the textual modulation. The metric used for evaluation is the average vIoU@R for $R = \{0.3, 0.5\}$. As illustrated, the grounding accuracy shows a gradual increase as W becomes larger, with satisfactory performance achieved when $W \geq 2$.

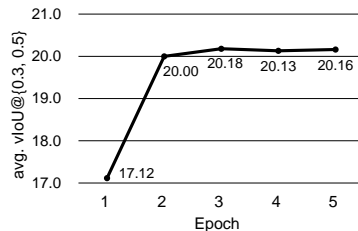


Fig. 1: Ablation studies on EM epochs.

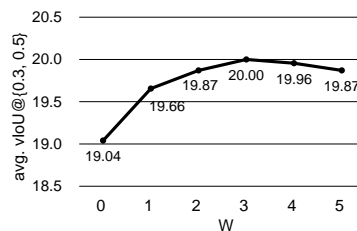


Fig. 2: Ablation studies on W .

Impact of Clustering Number C . Table 1 summarizes the impact of using varying values of clustering number C on the HC-STVG dataset. Empirically, our results indicate that utilizing a clustering number C in the range of 4 or 5 yields optimal performance. Furthermore, our experiments show that a C value outside this range leads to suboptimal performance. However, it still outperforms the state-of-the-art weakly-supervised method Winner [1] and zero-shot method ReCLIP [3] with a clear margin.

✉ Corresponding authors.

Method	m_vIoU	vIoU@0.3	vIoU@0.5
Winner [1]	14.20	17.24	6.12
ReCLIP [3]	14.36	18.28	4.91
E3M ($C = 3$)	18.29	26.12	7.76
E3M ($C = 4$)	19.11	29.40	10.60
E3M ($C = 5$)	18.21	27.67	11.12
E3M ($C = 6$)	16.92	25.26	9.66

Table 1: Ablation studies on clustering number C .

2 Additional Qualitative Analysis

Qualitative Comparisons. Fig. 3 and 4 provide qualitative comparisons among ground truth, ReCLIP [3], and our E3M method. The examples in both figures illustrate the advantages of our method in temporal-wise and spatial-wise grounding, respectively. We note that grounding errors can be attributed to two primary factors: 1) localizing the incorrect temporal boundary, such as failures to accurately pinpoint the event as illustrated in Fig. 3, and 2) localizing the wrong object, for instance, struggling to identify the correct person within a crowd described by the sentence as shown in Fig. 4. And our E3M method can effectively filter out these grounding errors that appear in the ReCLIP.

3 Additional Implementation Details

E3M. The clustering number C is set to be 4 for HC-STVG and 3 for VidSTG. We set the β in Eq.18 as 1.0 for HC-STVG and 0.9 for VidSTG dataset.

Baselines. Existing STVG approaches are generally categorized into two groups: fully-supervised and weakly-supervised learning, both relying on extensive video data with corresponding annotations for training. As our work is pioneering to explore zero-shot STVG, we carefully adapt the state-of-the-art zero-shot image grounding method, namely ReCLIP [3] and RedCircle [2], to STVG for comparative analysis. We first employ the state-of-the-art zero-shot temporal grounding method SPL [4] as the temporal grounding method. After the acquisition of the temporal grounding results, we apply the spatial grounding to each frame that is inferred as positive by the temporal grounding method. Subsequently, we select the object identified by the zero-shot spatial grounding method at each frame as the prediction results. Our implementation employs the same hyperparameter settings as outlined in the original papers [2,3].

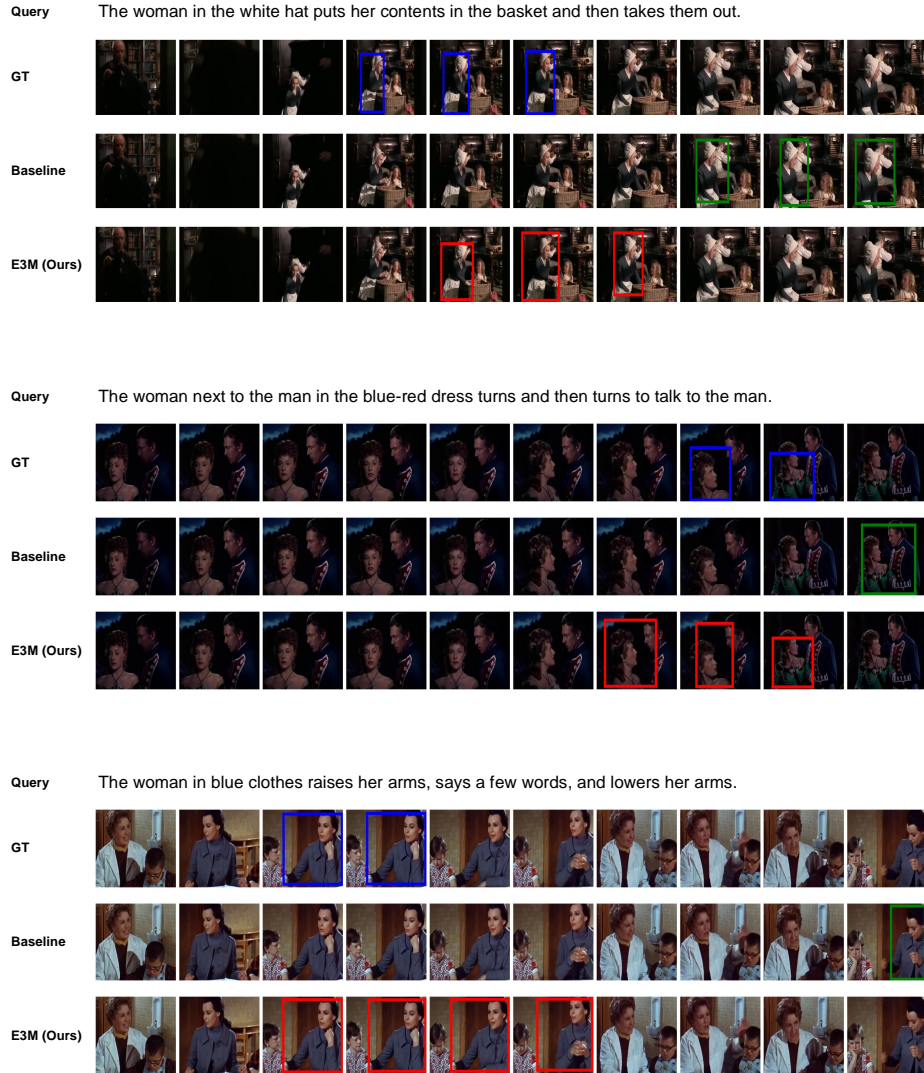


Fig. 3: Qualitative comparisons between E3M and ReCLIP highlight the superiority of our method in *temporal grounding*. The red bounding boxes represent the prediction of our zero-shot method E3M, the green ones denote the results of ReCLIP [3], and the blue ones represent the ground truth.

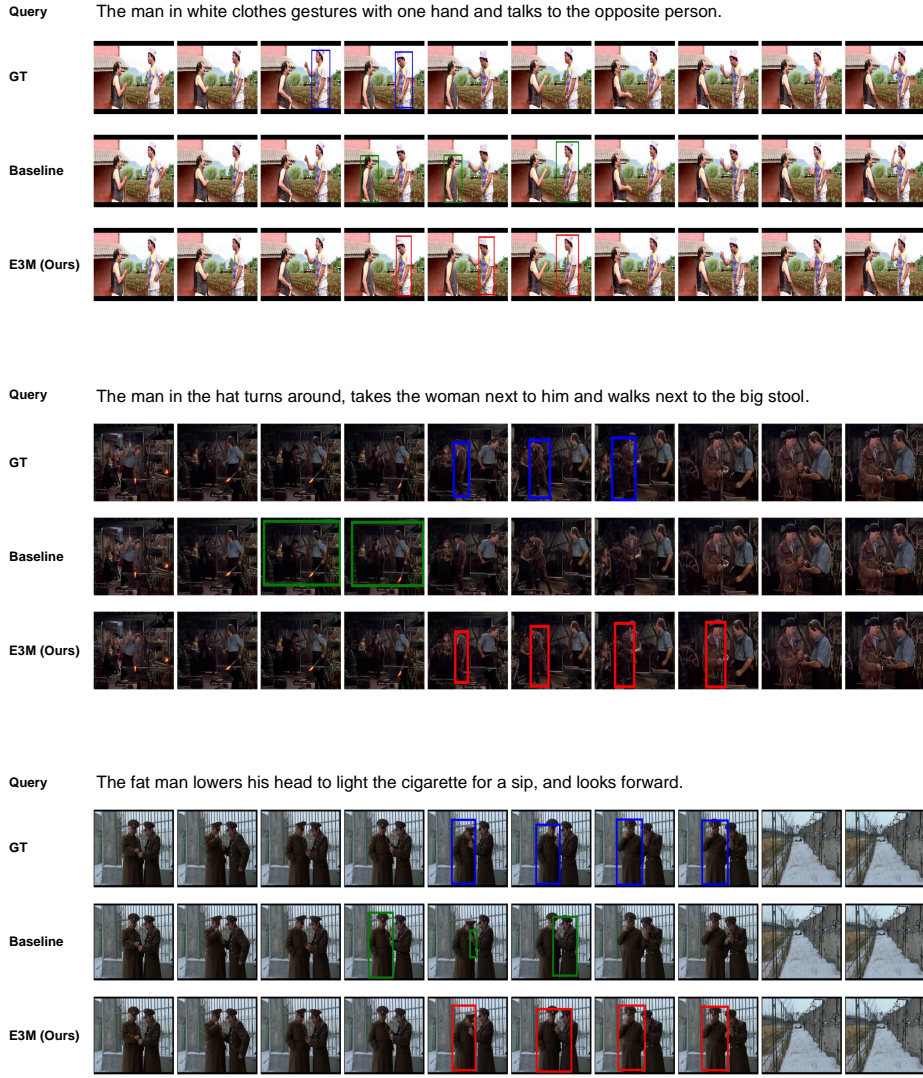


Fig. 4: Qualitative comparisons between E3M and ReCLIP highlight the superiority of our method in *spatial grounding*. The red bounding boxes represent the prediction results of our zero-shot method E3M, the green ones denote the results of ReCLIP [3], and the blue ones represent the ground truth.

References

1. Li, M., Wang, H., Zhang, W., Miao, J., Zhao, Z., Zhang, S., Ji, W., Wu, F.: Winner: Weakly-supervised hierarchical decomposition and alignment for spatio-temporal video grounding. In: CVPR (2023) [1](#), [2](#)
2. Shtedritski, A., Rupprecht, C., Vedaldi, A.: What does clip know about a red circle? visual prompt engineering for vlms. In: CVPR (2023) [2](#)
3. Subramanian, S., Merrill, W., Darrell, T., Gardner, M., Singh, S., Rohrbach, A.: Reclip: A strong zero-shot baseline for referring expression comprehension. In: ACL (2022) [1](#), [2](#), [3](#), [4](#)
4. Zheng, M., Gong, S., Jin, H., Peng, Y., Liu, Y.: Generating structured pseudo labels for noise-resistant zero-shot video sentence localization. In: ACL (2023) [2](#)