Weakly-Supervised Spatio-Temporal Video Grounding with Variational Cross-Modal Alignment

Yang Jin[1] and Yadong Mu[1*]

[1]Peking University
jiny@stu.pku.edu.cn, myd@pku.edu.cn

Abstract. This paper explores the spatio-temporal video grounding (STVG) task, which aims at localizing a particular object corresponding to a given textual description in an untrimmed video. Existing approaches mainly resort to object-level manual annotations as the supervision for addressing this challenging task. Such a paradigm heavily constrains the scalability of processing large-scale unlabeled data. To this end, we present a novel framework that is capable of grounding the target object relying only on the video-sentence correspondence. Specifically, our model re-formulates the original STVG task as two cross-modal alignment sub-problems: *region-phrase* and *frame-sentence*. Since the absence of ground-truth alignments during the training stage, we treat them as latent variables and learn to model the joint conditional distribution by reconstructing the interactions of entities in the video. The entire framework can be effectively optimized by the variational Expectation-Maximization (EM) algorithm, which alternates between two updating steps for progressively maximizing the likelihood of query sentence, thereby approximating the real cross-modal assignment. Extensive experiments on two video benchmarks (VidSTG and HC-STVG) further show the effectiveness of the proposed method.

Keywords: visual grounding · weakly-supervised learning · multi-modality

1 Introduction

Visual grounding aims at finding a specific region associated with linguistic meaning in the visual modality, which serves as a fundamental block in the complicated multi-modal system. In the video scenario, understanding how language description relates to video content in spatial-temporal dimensions is paramount for video surveillance [10], video question answering [18], etc. Recently, spatio-temporal video grounding (STVG) was introduced in [43], which requires localizing a spatio-temporal tube (i.e., a sequence of bounding boxes) of the target object described by language query from the untrimmed video (See Figure 1). In contrast to prior grounding methods focusing on images, the STVG task poses a greater difficulty as it involves discerning nuanced characteristic changes of video instances based solely on textual semantics.

* Corresponding Author.

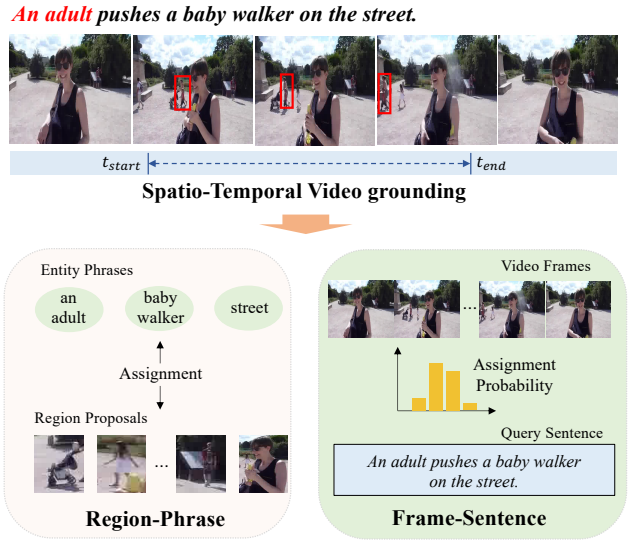

Fig. 1: The STVG task requires localizing the bounding box sequence and the temporal boundary of the target object described by a query sentence. We factorize this task into two sub-problems: *region-phrase* and *frame-sentence* alignments, which aim to match the region to the textual phrase in the query sentence and identify which frames belong to the target video segment, respectively.

To accomplish this complicated task, the prevalent approaches [16, 33, 38, 42, 43] are heavily dependent on fine-grained human annotations such as temporal boundaries and bounding boxes on each video frame. Despite achieving promising results, the labor-intensive process of annotation in each video frame makes it challenging to scale these fully-supervised methods to large-scale datasets. Accordingly, the utilization of weak supervision, where only video-sentence correspondences are available during training, is considered to be more practical in real-world applications. To this end, we focus on seeking an efficacious paradigm for weakly-supervised spatio-temporal video grounding in this work.

Although weakly-supervised grounding in images has received significant research attention [1, 11, 21, 36, 37], just a few works [8, 19] have explored this challenging setting in the STVG task. The weakly-supervised STVG task invokes several unique challenges associated with the lack of annotation and time-varying video content. Firstly, the precise alignment of region and phrase is intricate due to the existence of entities that share similar visual features or perform analogous actions within a video frame. For example, given the query of "An adult pushes a baby walker on the street" as in Figure 1, the absence of spatial supervision exacerbates the difficulty in distinguishing the target "adult" since there are four adults in the scene. In this regard, the accurate grounding of the target object is contingent upon the explicit modeling of the visual relationship (e.g., "push") between entities ("adult" and "baby walker") within the video. Secondly, unlike in images where the visual relationships are static, the interactions in untrimmed videos are dynamic and will change or disappear over time. However, the query

sentence may only refer to a short-term state of the queried object, e.g. the action "push" only occurs in a short video clip. It is of great significance to determine the temporal boundary of the queried object tubes.

To cope with these unique challenges in the weakly-supervised setting, this paper presents a novel framework for the STVG task. Specifically, given a video and its corresponding language description, we parse the text to several noun phrases that denote objects and extract a set of region proposals as candidates in each video frame. Then, the STVG task can be interpreted as two cross-modal alignment problems: *region-phrase* and *frame-sentence*. The former entails identifying the region that accurately represents the phrase, while the latter involves determining which frames correspond to the video segment conveying the semantics of the query sentence. Due to the absence of ground truth in the weakly-supervised context, we regard them as latent variables and proceed to learn the joint distribution conditioned on the correspondence between the video and sentence. The entire framework can be efficiently optimized by the variational Expectation-Maximization (EM) algorithm [26], which alternates between two learning steps. In the E-step, we introduce a variational posterior distribution to approximate the real assignment. In the M-step, we learn to reconstruct the visual relationships between the entity phrases, based on the samples drawn from the variational assignment distribution. Through multiple optimization iterations aimed at reconstructing the visual relationships between entities, the model will progressively learn a more precise cross-modal alignment. The main technical contributions of our work can be summarized as follows:

- This paper explores the STVG task in the challenging weakly-supervised setting. Specifically, we innovatively bifurcate spatio-temporal grounding into two distinct cross-modal alignment sub-problems and formulate them by effective latent variables.
- We present a methodology that employs the variational Expectation Maximization algorithm to optimize the entire framework. Notably, the proposed model successfully achieves fine-grained cross-modal alignment without manual annotations during training.
- We conduct extensive experiments on two large challenging video benchmarks, namely VidSTG [43] and HC-STVG [34]. The results indicate that our proposed approach surpasses the existing weakly-supervised approaches by a significant margin, thus proving its superiority.

2 Related Work

Fully-Supervised Video Grounding. The video grounding task can be divided into two main categories: temporal and spatio-temporal grounding. In the past few years, temporal grounding [6, 7, 25, 40] has emerged as a significant area of research interest and has garnered considerable attention from scholars. Due to the computational complexity of spatial-temporal dimensions in videos, the STVG task has been explored relatively less. The prevailing approaches [16, 33, 34, 38, 43] for STVG primarily investigate the fully-supervised

paradigm. Zhang *et al.* [43] develops a spatio-temporal graph that facilitates message passing across all frame regions. Lately, some approaches [16, 38] have proposed a one-stage approach that directly generates the bounding box at each frame and predicts the starting and ending timestamps. Notwithstanding, all of these techniques are substantially dependent on the manual annotations for both the temporal boundary and spatial bounding box in every video frame. In contrast, this work is focused on a weakly-supervised setting where the models solely rely on the video-sentence correspondence, without the annotations of the temporal and spatial location.

Weakly-Supervised Video Grounding. In the video scenario, the primary research still focuses on temporal grounding, while only a few studies [2, 8, 9, 19, 23, 32] have devoted to the weakly supervised spatio-temporal grounding task. The WSSTG [9] developed an attentive interactor that exploits the fine-grained contextual information to match the pre-extracted object tubes. However, they only pay attention to spatial grounding at each frame and ignore the temporal grounding. Chen *et al.* [8] propose novel spatial and temporal multiple instance learning frameworks for the untrimmed video grounding. Li *et al.* [19] develops a multi-modal decomposition tree to perform multi-hierarchy language-tube matching. Most of the existing works do not account for fine-grained entity interactions, thereby rendering them incapable of resolving semantic ambiguities in the query sentence and distinguishing the target object.

3 Method

3.1 Problem Formulation

The Spatio-Temporal Video Grounding (STVG) task aims to identify a spatio-temporal tube $\{b_t\}_{t=t_s}^{t_e}$ in an untrimmed video V consisting of T frames, which represents the target object described by a given query sentence S. Here, b_t denotes the bounding box in the t-th frame, while t_s and t_e indicate the starting and ending boundaries of the object tube being retrieved, respectively. In the weakly-supervised setting, only global video-sentence correspondence (V, S) is available and the model does not have any information regarding the spatio-temporal location of the queried object during the training stage.

3.2 Overview

Given the video-sentence pair (V, S), we first generate M region proposals for each video frame and parse the description S to a group of noun phrases and relationships. Thus, the video V can be denoted by $V = \{b_{i,j} | i \in [1, T], j \in [1, M]\}$. The query sentence S can be represented as (E, R), where $E = \{e_n\}_{n=1}^{N}$ is a set of noun phrases that indicate the entities in the video and $R = \{r_k\}_{k=1}^{K}$ are the visual relationships that depict the interactions between entities. Then, the STVG problem can be interpreted by two sub-problems: *region-phrase* and *frame-sentence* alignment. The former seeks to map the entity phrases with their

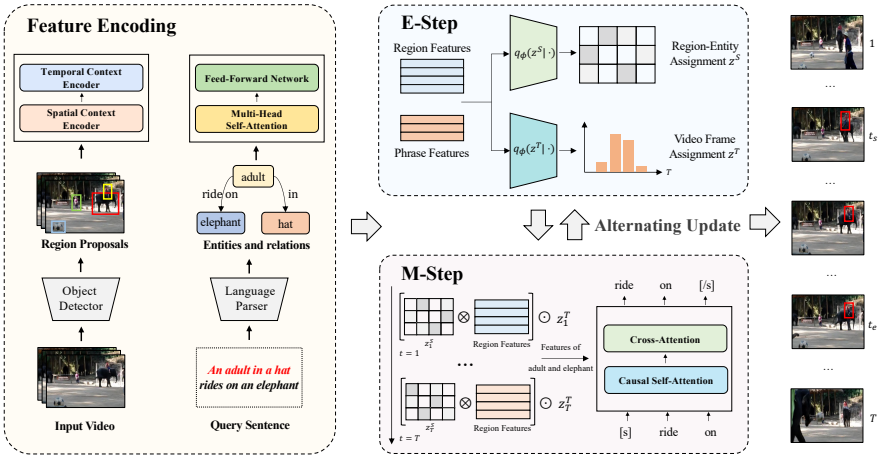

Fig. 2: The architecture of the proposed weakly-supervised spatio-temporal grounding framework. Given an input video and a query sentence, it first extracts region proposals from the video and parses the text to entity and relation phrases. Then, the feature encoding modal is responsible for aggregating the contextual information among the regions and phrases. The encoded features are then fed into the q_ϕ and p_θ networks to conduct the Variational Expectation-Maximization (EM) optimization process.

corresponding regions, while the latter concerns identifying the frames belonging to the ground-truth video segment. Since the actual assignment is unobserved during training, we model it as a latent variable $Z = (z^S, z^T)$, where $z^S = \{z_t^S\}_{t=1}^T$ with $z_t^S \in \mathcal{R}^{N\times M}$ and $z^T = \{z_t^T\}_{t=1}^T$. These two latent variables have the range of $[0, 1]$, and denote the probability of spatial or temporal assignment.

Based on the above formulation, we propose to estimate the joint conditional distribution $p_\theta(R, Z|E, V)$ in this work. Furthermore, the parameters θ can be optimized by maximizing the log-likelihood of the observed visual relationship, denoted as $\log p_\theta(R|E, V)$. However, direct maximization of this log-likelihood is not feasible due to the unavailability of real cross-modal assignment Z during training. To this end, we instead pursue optimizing the evidence lower bound (ELBO) of this log-likelihood as follows:

$$\log p_\theta(R|E, V) \geq \mathbb{E}_{q_\phi(Z|E,V)}\left[\log \frac{p_\theta(R, Z|E, V)}{q_\phi(Z|E, V)}\right], \tag{1}$$

where $q_\phi(Z|E, V)$ is a variational distribution parametrized by ϕ that approximates the real posterior distribution $p_\theta(Z|R, E, V)$ and the equations holds only when $q_\phi(Z|E, V) = p_\theta(Z|R, E, V)$. By maximizing this lower bound, the real log-likelihood $\log p_\theta(R|E, V)$ can be accordingly optimized.

Inspired by [3, 17, 28], this evidence lower bound is optimized via the variational Expectation-Maximization (EM) algorithm [26], which alternates between E-step and M-step to update ϕ and θ, respectively. To be specific, in the E-step, we fix θ and minimize the KL divergence $\mathcal{F}(\phi)$ between $q_\phi(Z|E, V)$ and

$p_\theta(Z|R,E,V)$:

$$\mathcal{F}(\phi) = \sum_Z q_\phi(Z|E,V) \log \frac{q_\phi(Z|E,V)}{p_\theta(Z|R,E,V)}. \tag{2}$$

This step yields an approximation of the real cross-modal assignment. In the M-step, we freeze ϕ and update θ to maximize the joint log-likelihood function $\mathcal{F}(\theta)$ as follows:

$$\mathcal{F}(\theta) = \mathbb{E}_{q_\phi(Z|E,V)} \left[\log p_\theta(R, Z|E, V)\right]. \tag{3}$$

Through the M-step, the model acquires the ability to reconstruct the visual relationships R and further refines the assignment Z. The insight behind is that if the interactions between entities depicted by the query sentence can be recovered perfectly, the model must learn to match the entity phrase with the regions appropriately. Consequently, the above optimization procedures will contribute to effective spatio-temporal grounding. The entire framework is illustrated in Figure 2, which consists of two individual networks parameterized by ϕ and θ. In the subsequent sections, we first present the technical details of feature encoding and then elaborate on these two updating steps, respectively.

3.3 Contextualized Feature Encoding

We will commence the overall pipeline by introducing feature extraction and context encoding. As shown in Figure 2, a pre-trained object detector [29] is employed to extract M region proposals in each video frame. For per region $b_{i,j}$, we use the RoI-Align [14] and global average pooling to obtain its region-level representation, denoted by $v_{ij} \in \mathcal{R}^D$. Additionally, the global frame features $\{f_t\}_{t=1}^T$ can also be obtained from the detector backbone. For each entity phrase, its word-level embeddings are obtained from the Glove model [27] first. Then, the linguistic representation $w_n \in \mathcal{R}^D$ for the phrase e_n is thus computed by applying the average pooling to its word-level embeddings.

Video-Context Encoding. After extracting the feature v_{ij} for each region proposal in the video, this module employs the multi-head self-attention layer [35] to capture the contextual information among all regions. For the sake of computational efficiency, the encoding of region context is factorized into spatial and temporal dimensions, respectively. Specifically, spatial context encoding is performed among regions within the same frame, thereby fusing local context information. The temporal interaction is restricted at the tube level, where a tube is a sequence of regions that may represent the same object. To obtain the object tube, a linking score $s_{\text{link}}(b_{t,i}, b_{t+1,j})$ is defined to establish the association of regions between two consecutive video frames:

$$s_{\text{link}}(b_{t,i}, b_{t+1,j}) = \cos(v_{t,i}, v_{t+1,j}) + \alpha \cdot \text{IoU}(b_{t,i}, b_{t+1,j}), \tag{4}$$

where $\cos(\cdot)$ is the cosine similarity of the region features, IoU is the intersection-over-union of two regions and β is the hyper-parameter that controls the importance of IoU metric. Based on the linking score, we can leverage the Viterbi

Algorithm [13] to generate M tubes per video by gradually selecting the path with the highest linking score. Through the message passing within the tube, each v_{ij} absorbs the temporal dynamics of objects across the video. The overall video context encoding involves alternant spatial and temporal interaction.

Language-Context Encoding. This module utilizes another multi-head self-attention layer to integrate contextual semantics across all entities within the query description S. Finally, the resulting contextualized visual and linguistic representations are fed into the subsequent modules.

3.4 E-Step: Approximating the Alignment

The exact estimation of the actual posterior distribution is intractable, thus we introduce a variational distribution $q_\phi(Z|E,V)$ to provide the approximation for the cross-modal alignment in E-step. Following the empirical practice in the variational inference [5], we utilize the mean-field approximation that assumes the independence of all the latent assignment variables to formulate $q_\phi(Z|E,V)$:

$$q_\phi(Z|E,V) = \prod_{t=1}^{T} q_\phi(z_t^S|E,V) \prod_{t=1}^{T} q_\phi(z_t^T|E,V). \tag{5}$$

Specifically, the factorized spatial assignment $q_\phi(z_t^S|E,V)$ and temporal assignment $q_\phi(z_t^T|E,V)$ can be parameterized by the following two modules.

Spatial Assignment $q_\phi(z_t^S|E,V)$. The goal of this module is to estimate the region-phrase alignment $z_t^S \in \mathcal{R}^{N\times M}$ in each video frame. For the sake of brevity in the following discussion, we omit the subscript t of the variable z_t^S. Given the contextualized region and phrase representations, two modality-dependent projection layers $g_V(\cdot)$ and $g_S(\cdot)$ are employed to map them into a joint multi-modal feature space. Here, both $g_V(\cdot)$ and $g_S(\cdot)$ are implemented by fully connected layers with ReLU activation. Let denote $z_{i,j}^S$ as the assignment variable between i-th phrase and j-th region in a video frame, the $q_\phi(z_{i,j}^S|E,V)$ is computed by:

$$q_\phi(z_{i,j}^S|E,V) = \frac{\exp\left(g_S(w_i)^\top g_V(v_j)\right)}{\sum_{j=1}^{M} \exp\left(g_S(w_i)^\top g_V(v_j)\right)}. \tag{6}$$

Temporal Assignment $q_\phi(z_t^T|E,V)$. The aforementioned spatial assignment resolves the box grounding at each frame, whereas this module endeavors to estimate the probability of individual frames appearing in the ground-truth video segment. Intuitively, since the temporal grounding requires determining the accurate video segment corresponding to the semantics of description S, it is of great significance to capture the interaction between global visual and textual information. Therefore, following [41,45], the cross-attention layers are leveraged to conduct multi-modal fusion between the frame features $\{f_t\}_{t=1}^T$ and sentence features $\{w_i\}_{i=1}^N$. The resulting cross-modal global representation f_{cls} is leveraged to produce the center c and width σ of the temporal boundary $[t_s, t_e]$ by a

prediction layer, which is a linear mapping followed by sigmoid activation. Based on c and σ, we model $q_\phi(z_t^T|E,V)$ by a Gaussian distribution:

$$q_\phi(z_t^T|E,V) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(t/T-c)^2}{2\sigma^2}\right), t \in [0,T]. \tag{7}$$

Optimization. With the above formulation, the parameter ϕ can be optimized by fixing θ and minimizing the function $\mathcal{F}(\phi)$. When θ is frozen, the $\log p_\theta(R|E,V)$ is the constant, and minimizing $\mathcal{F}(\phi)$ is equivalent to maximizing the evidence lower bound (the expectation term in Eq. 1). By substituting the mean-field formulation in Eq. 5 into the ELBO and taking the derivative with respect to each component z_t^S or z_t^T, the optimal distribution q_ϕ can be obtained by the coordinate ascent update rule:

$$q_\phi(z_t^S|E,V) \propto \exp\left(\mathbb{E}_{-z_t^S}[\log p_\theta(z_t^S|R,E,V,-z_t^S)]\right), \tag{8}$$

where $-z_t^S$ indicates all the latent components except z_t^S. The detailed proof is presented in the appendix. Based on the independence assumption for Z, the optimal solution satisfies $q_\phi(z_t^S|E,V) \approx p_\theta(z_t^S|R,E,V)$. Here, we only draw one sample from $p_\theta(\cdot)$ as the pseudo-label in the implementation. Take z^S as the example, the parameter ϕ can be updated by the cross entropy loss:

$$\mathcal{L}_{z^S,\phi} = -\sum_t p_\theta(z_t^S|E,R,V)\log q_\phi(z_t^S|E,V). \tag{9}$$

Moreover, we also introduce a contrastive objective, which leverages the video-sentence alignment as the supervision to further optimize the parameter ϕ. Concretely, the region-phrase matching score between the i-th phrase and the j-th region in the t-th frame can be obtained by $s_{t,i,j} = g_S(w_i)^\top g_V(v_{t,j})$. Then the similarity score between video V and sentence S is calculated by:

$$s_{\text{match}}(V,S) = \sum_{t=1}^{T}\sum_{i=1}^{N} z_t^T \max_{1\le j\le M} s_{t,i,j}. \tag{10}$$

The contrastive loss $\mathcal{L}_{\text{contra},\phi}$ is defined on all the matched and unmatched video-sentence pairs in a training batch:

$$\mathcal{L}_{\text{contra},\phi} = -\sum_{i=1}^{B}\log\frac{\exp(s(V_i,S_i)/\tau)}{\sum_{j=1}^{B}\exp(s(V_j,S_i)/\tau)}, \tag{11}$$

where τ is the temperature parameter and B is the batchsize.

3.5 M-Step: Reconstructing the Relationship

The objective of the M-step is to maximize the joint log-likelihood $\mathcal{F}(\theta)$ in Eq. 3. Following [28, 30], the optimization of $\mathcal{F}(\theta)$ can be further formulated by the pseudolikelihood function [4]:

$$\mathcal{F}(\theta) \approx \mathbb{E}_{q_\phi(Z|E,V)}\left[\log p_\theta(Z|E,V,R) + \log p_\theta(R|E,V,Z)\right]. \tag{12}$$

Algorithm 1: The optimization of our framework.

Input: The training dataset $\mathcal{D}$ with only the correspondence of video-sentence $\{V_i, S_i\}$

Output: The learned parameters for the q_ϕ and p_θ.

Initialize both ϕ and θ randomly.
while ϕ, θ has not converged **do**
  `// Conduct the E-Step`
  for (V_i, S_i) *in* $\mathcal{D}$ **do**
    Compute $p_\theta(Z|E, R, V)$ based on the current θ;
    Update ϕ based on $\mathcal{L}_{Z,\phi}$ and $\mathcal{L}_{\text{contra},\phi}$;
  `// Conduct the M-Step`
  for (V_i, S_i) *in* $\mathcal{D}$ **do**
    Compute $q_\phi(Z|E, V)$ based on current ϕ;
    Update θ based on $\mathcal{L}_{Z,\theta}$ and $\mathcal{L}_{\text{rec},\theta}$;
Predict the grounding result using q_ϕ and p_θ.

It consists of two conditional probability $p_\theta(Z|E, V, R)$ and $p_\theta(R|E, V, Z)$ that estimate the distributions of cross-modal assignment and the relationships between entities.

Optimization of Assignment. This learning objective is to maximize the first term in Eq.12. Here, we adopt the network implementation similar to q_ϕ to formulate $p_\theta(Z|E, V, R)$. The difference is that p_θ is also conditioned on the relationships R. Therefore, unlike q_ϕ that only utilizes entity features, the estimations of z^S and z^T integrate the representations of all entity and relation phrases. Finally, z^S and z^T yielded from q_ϕ are treated as the targets to update θ with a loss function analogous to Eq. 9.

Optimization of Reconstruction. The second optimization term of Eq.12 is obliged to reconstruct the relationships R between entities E given the cross-modal assignment Z predicted by q_ϕ. Here, a visual relationship is denoted by the triplet (e_s, r, e_o) that represents the interaction between subject e_s and object e_o. Specifically, based on the estimated $z_t^S \in \mathcal{R}^{N\times M}$ and z_t^T, the entity-relevant representation H is calculated by:

$$H = \{z_t^T \odot \tilde{h}_t\}_{t=1}^T, \quad \tilde{h}_t = z_t^S v^t \in \mathcal{R}^{N\times D}, \tag{13}$$

where $v_t \in \mathcal{R}^{M\times D}$ is the contextualized visual features of M regions in the t-th frame and z_t^S serves as the gate function that modulates each $\tilde{h}_t$ to weaken the features of irrelevant frames. Then, the feature used to reconstruct the relationship r is obtained by:

$$h_t^r = W_r[h_t^{e_s}; h_t^{e_o}] + b_r, \tag{14}$$

where $h_t^{e_s}$ and $h_t^{e_o}$ are the representations for subject and object entities, respectively. As illustrated in Figure 2, the generated $\{h_t^r\}_{t=1}^T$ in the overall video will be fed to a standard transformer decoder [35] with the causal self-attention layer

for recovering the relation r in an autoregressive way. This transformer decoder is optimized by the cross-entropy loss on the words of r:

$$\mathcal{L}_{\text{rec},\theta} = -\sum_{i=1}^{L} \log p_\theta(r_i | r_{<i}, E, V, Z), \tag{15}$$

where r_i is the i-th word in the relation r. This design facilitates the grounding, as the model can only reconstruct the relationship between the subject and predicate by correctly aligning them with the relevant region (z^S) and attending to the corresponding video segment (z^T).

3.6 Training and Inference

Training. The overall optimization algorithm is summarized in Algorithm 1. It alternates between E-step and M-step to progressively update q_ϕ and p_θ until convergence. In the E-step, θ is frozen, and we update ϕ by $\mathcal{L}_{Z,\phi}$ and $\mathcal{L}_{\text{contra},\phi}$. In the M-step, ϕ is frozen, and parameter θ is optimized by $\mathcal{L}_{Z,\theta}$ and $\mathcal{L}_{\text{rec},\theta}$. It is worth noting that, during the E-step, q_ϕ is updated based on p_θ using $\mathcal{L}_{Z,\phi}$, which brings the knowledge acquired through relation construction into q_ϕ. Similarly, in the M-step, p_θ is updated with q_ϕ using $\mathcal{L}_{Z,\theta}$, which injects the knowledge gained through contrastive learning into p_θ. Consequently, these two steps work synergistically during the optimization process.

Inference. At inference, our framework produces the bounding boxes and the starting and ending probabilities based on the prediction of both q_ϕ and q_θ. In detail, we can generate the estimated assignment variable Z by:

$$Z = \beta Z_{q_\phi} + (1 - \beta) Z_{p_\theta}. \tag{16}$$

The bounding box prediction at each frame is obtained by selecting the region proposal with the maximal assignment score between the queried object phrase and all M regions based on z^S. The temporal boundary of the object tube is determined by selecting the segment with the highest z^T.

4 Experiments

4.1 Datasets and Metrics

We conduct extensive experiments on two widely-adopted spatio-temporal video grounding benchmarks, dubbed as **VidSTG** [43] and **HC-STVG** [34]. Both these two datasets comprise manual annotations for the spatio-temporal tubes concerning a query sentence. For instance, the VidSTG contains a total of 44,808 video segments annotated with 99,943 sentences. The other dataset HC-STVG consists of 5,660 raw video files depicting scenes with multiple people in the movie, where each video has been annotated the sentence to identify a relevant human with a specific attribute or interaction with the surrounding

objects. We adopt the same dataset split for training and text following the previous works [34, 43]. To perform a quantitative evaluation of the grounding performance, we adopt the criteria proposed in [43], including **m_VIoU** and **VIoU@R**. VIoU score is calculated by: $\text{VIoU} = \frac{1}{|S_u|}\sum_{t\in S_i}\text{IoU}(\hat{b}_t, b_t)$, where S_i and S_u represent the intersection and union of the predicted and ground-truth video segments. And $\text{IoU}(\hat{b}_t, b_t)$ indicates the IoU score between the predicted $\hat{b}_t$ and ground-truth bounding box b_t at frame t. The m_VIoU score denotes the mean value of the VIoU scores across all testing videos and VIoU@R quantifies the proportion of samples in the testing subset whose VIoU score exceeds a specific threshold R.

4.2 Implementation Details

Consistent with the previous methods, the input video frames are first sampled at the rate of 5 fps. To handle the overlong videos, we conduct uniform sampling to select at most 200 frames throughout the video. Then, for each frame, a Faster-RCNN object detector pre-trained on MS-COCO [20] with ResNet-101 [15] backbone is leveraged to extract 10 regions as the candidate proposal, where each region is associated with an RoI align pooled feature. And the representation of each video frame is derived from the C4 block in the ResNet-101. In regards to the query sentence, we borrow the Stanford CoreNLP tools [24] to parse a set of entities and relationships. All the representation dimensions in the model are set to $D = 256$. Additionally, both the context encoder and reconstruction decoder have two transformer blocks with 8 heads in the attention module. We empirically set the hyper-parameters in this paper with $\alpha = 0.5$ and $\beta = 0.3$. The whole framework is trained for 20 epochs with a batch size of 32 using AdamW [22] optimizer on 4 NVIDIA V100 GPUs, where the learning rate is warmed up to 5e-5 and then decayed linearly in the remaining iterations.

4.3 Performance Comparison

To fully demonstrate the advantage of the proposed framework, this section provides extensive quantitative comparisons with the existing weakly-supervised grounding approaches. Here, we consider the following types of methods as the competitors. (1) **Object Similarity**. This type of approach predominantly employs the semantic similarity between the region label predicted by the detector and the subject phrase in the query sentence to generate the target object tube. Essentially, this is the process of grounding by utilizing the knowledge distilled from external pre-trained object detectors. (2) **Factorized Grounding**. Since the STVG problem is a compound task, an intuitive way to settle this problem is to handle the spatial and temporal sub-grounding, respectively. To this end, we first leverage the weakly-supervised image grounding methods (e.g., GroundeR [31], MATN [44], RAIR [21]) to predict the bounding-box at each video frame, and then resort to weakly-supervised temporal grounding techniques (e.g., LCNet [39], CPL [45]) to yield the predicted video segment.

Methods	Declarative Sentences			Interrogative Sentences		
	m_VIoU	VIoU@0.3	VIoU@0.5	m_VIoU	VIoU@0.3	VIoU@0.5
Object Similarity	3.41	2.78	1.31	2.87	2.95	0.98
GroundeR [31]+LCNet [12]	7.85	7.96	3.02	6.43	6.58	2.92
MATN [44]+LCNet [12]	8.16	8.03	3.59	6.97	6.64	3.05
GroundeR [31]+CPL [12]	8.28	8.35	3.68	7.16	7.28	3.23
RAIR [21]+CPL [12]	8.67	8.72	4.01	7.68	7.71	3.58
WSSTG [9]	8.85	8.52	3.87	7.12	6.87	2.96
ADWS [8]	8.96	7.86	3.10	8.57	6.84	2.88
Vis-Ctx [32]	9.34	7.32	3.34	8.69	7.18	2.91
WINNER [19]	11.62	14.12	7.40	10.23	11.96	5.46
Ours	**14.45**	**18.57**	**8.76**	**13.25**	**16.74**	**7.66**

Table 1: Performance comparisons of different methods on the VidSTG test set (%).

Methods	m_VIoU	VIoU@0.3	VIoU@0.5
Object Similarity	3.25	2.17	0.32
GroundeR [31]+LCNet	4.17	3.28	1.05
MATN [44]+LCNet	4.41	3.53	1.12
GroundeR [31]+CPL	5.23	4.18	1.25
RAIR [21]+CPL	6.88	4.87	1.36
WSSTG [9]	6.52	4.54	1.27
ADWS [8]	8.20	4.48	0.78
Vis-Ctx [32]	9.76	6.81	1.03
WINNER [19]	14.20	17.24	**6.12**
Ours	**14.64**	**18.60**	5.75

Table 2: Performance comparisons of different methods on the HC-STVG test set (%).

(3) **Video Grounding**. Only a few works have explored the weakly-supervised video grounding. We compare with them: Vis-Ctx [32], WSSTG [9], ADWS [8], WINNER [19] on the two STVG benchmarks.

For a fair comparison, all these baselines adopt the same backbone to extract features. The detailed comparisons are presented in Tables 1 and 2 for two video benchmarks, respectively. It can be seen that the grounding performance of our approach surpasses all the competitors on the VidSTG benchmark and achieves the best results in terms of evaluation metrics: **m_VIoU** and **VIoU@0.3**. Furthermore, simply grounding the target object based on the category-phrase similarity will lead to unsatisfactory results. This is mainly because there are usually multiple objects with the same category in a video scene. The model must figure out the complicated interactions between these objects to finally identify the target. The factorized grounding approaches settle the STVG task separately, which ignores the correlations between the spatial and temporal contextual information and thereby achieves inferior grounding performance. In our work, the spatial and temporal assignment works collaboratively to recover the subtle visual relationships between the entities in the video. Moreover, the proposed method also outperforms most of the existing weakly-supervised video grounding methods by a large margin. Although they consider the spatio-temporal context clues throughout the whole video with the attention mechanism by leveraging

$\mathcal{L}_{\text{contra},\phi}$	$\mathcal{L}_{\text{rec},\theta}$	m_VIoU	VIoU@0.3	VIoU@0.5
✓		10.83	13.24	5.36
	✓	12.87	16.32	7.93
✓	✓	**14.45**	**18.57**	**8.76**

Table 3: Ablation of different network combinations on the test set of VidSTG benchmark (declarative sentences).

Spatial	Temporal	m_VIoU	VIoU@0.3	VIoU@0.5
✓		13.16	17.15	7.93
	✓	14.12	18.21	8.25
✓	✓	**14.45**	**18.57**	**8.76**

Table 4: Ablation of different context encoding layer on the test set of VidSTG benchmark (declarative sentences).

the multiple instance learning framework, the lack of explicitly modeling the fine-grained interactions between entities makes them fail to distinguish the semantic ambiguities in the query sentence.

4.4 Ablation Study

In this section, we perform various ablations on the VidSTG benchmark to gain deeper insights into the contributions and design decisions made regarding the individual components of the proposed framework. More ablation results can be found in our appendix.

Effect of different components. In the network optimization E-step and M-step, the contrastive loss $\mathcal{L}_{\text{contra},\phi}$ and reconstruction loss $\mathcal{L}_{\text{rec},\theta}$ are leveraged to supervise the parameters of q_ϕ and p_θ, respectively. For inference, both the predictions of q_ϕ and p_θ are responsible for producing the final grounding results. To explore the effectiveness of these two networks, we designed experiments for different combinations on the VidSTG benchmark. Based on the quantitative ablation results presented in Table 3, one can observe that canceling either q_ϕ or p_θ will weaken the grounding performance. Notably, the network p_θ brings a more significant performance boost compared to the q_ϕ. This is basically intuitive since the p_θ aims to recover the specific interactions between the grounding target and other context objects. While the q_ϕ network only utilizes the global correspondence of video-sentence and can not benefit from the fine-grained semantics in the description. Nevertheless, the cooperation of these two networks achieves the best grounding performance.

Effect of the video context encoding. In the contextualized feature encoding module, the spatial and temporal context modeling layers are adopted to capture the global dynamics of regions across the whole video. We thus design an ablation experiment in Table 4 to investigate the function of these two layers for the subsequent grounding. From the presented comparison, we can observe a distinct performance drop without either of these two layers, which further validates the effectiveness of the spatial and temporal context encoding layer.

Fig. 3: Some illustration examples of the spatio-temporal video grounding predictions produced by our model on the VidSTG benchmark.

4.5 Visualization Analysis

In this section, we visualize some grounding results on the VidSTG benchmark to provide a qualitative analysis of the model performance. As illustrated in Figure 3, the left section displays the predicted sequence bounding boxes for the target object, while the right section presents the interactions among entities in the video. From the shown visualizations, one can clearly observe that our framework can rightly attend to the desired instance depicted by the textual description. It is worth noting that, even for the hard sample (the second one), where several instances similar to the target exist in one scene, the model can still distinguish the queried entity.

5 Conclusion

In this work, an effective framework is proposed for addressing the weakly-supervised spatio-temporal video grounding task. Specifically, we innovatively formulate the STVG task to two cross-modal alignment sub-problems. To cope with the absence of fine-grained ground-truth annotation, the assignments between phrase-region and frame-sentence are treated as latent variables optimized with the variational EM algorithm. Moreover, extensive experimental results on two video benchmarks further validate the superiority of our approach.

Acknowledgement. This work is supported by National Key R&D Program of China (2022ZD0160305), an internal grant of Peking University (2024JK28), and a grant from China Tower Corporation Limited.

References

1. Arbelle, A., Doveh, S., Alfassy, A., Shtok, J., Lev, G., Schwartz, E., Kuehne, H., Levi, H.B., Sattigeri, P., Panda, R., et al.: Detector-free weakly supervised grounding by separation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1801–1812 (2021)
2. Bao, P., Shao, Z., Yang, W., Ng, B.P., Kot, A.C.: E3m: Zero-shot spatio-temporal video grounding with expectation-maximization multimodal modulation. In: ECCV (2024)
3. Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., West, M., et al.: The variational bayesian em algorithm for incomplete data: with application to scoring graphical model structures. Bayesian statistics **7**(453-464), 210 (2003)
4. Besag, J.: Statistical analysis of non-lattice data. Journal of the Royal Statistical Society: Series D (The Statistician) **24**(3), 179–195 (1975)
5. Blei, D.M., Kucukelbir, A., McAuliffe, J.D.: Variational inference: A review for statisticians. Journal of the American statistical Association **112**(518), 859–877 (2017)
6. Chen, J., Chen, X., Ma, L., Jie, Z., Chua, T.S.: Temporally grounding natural sentence in video. In: Proceedings of the 2018 conference on empirical methods in natural language processing. pp. 162–171 (2018)
7. Chen, J., Ma, L., Chen, X., Jie, Z., Luo, J.: Localizing natural language in videos. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8175–8182 (2019)
8. Chen, J., Bao, W., Kong, Y.: Activity-driven weakly-supervised spatio-temporal grounding from untrimmed videos. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 3789–3797 (2020)
9. Chen, Z., Ma, L., Luo, W., Wong, K.Y.K.: Weakly-supervised spatio-temporally grounding natural sentence in video. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 1884–1894 (2019)
10. Collins, R.T., Lipton, A.J., Kanade, T., Fujiyoshi, H., Duggins, D., Tsin, Y., Tolliver, D., Enomoto, N., Hasegawa, O., Burt, P., et al.: A system for video surveillance and monitoring. VSAM final report **2000**(1-68), 1 (2000)
11. Datta, S., Sikka, K., Roy, A., Ahuja, K., Parikh, D., Divakaran, A.: Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2601–2610 (2019)
12. Gao, J., Sun, C., Yang, Z., Nevatia, R.: Tall: Temporal activity localization via language query. In: Proceedings of the IEEE international conference on computer vision. pp. 5267–5275 (2017)
13. Gkioxari, G., Malik, J.: Finding action tubes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 759–768 (2015)
14. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
16. Jin, Y., Yuan, Z., Mu, Y., et al.: Embracing consistency: A one-stage approach for spatio-temporal video grounding. Advances in Neural Information Processing Systems **35**, 29192–29204 (2022)

17. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
18. Lei, J., Yu, L., Bansal, M., Berg, T.L.: Tvqa: Localized, compositional video question answering. arXiv preprint arXiv:1809.01696 (2018)
19. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23090–23099 (2023)
20. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
21. Liu, Y., Wan, B., Ma, L., He, X.: Relation-aware instance refinement for weakly supervised visual grounding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5612–5621 (2021)
22. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
23. Luo, Z., Guillory, D., Shi, B., Ke, W., Wan, F., Darrell, T., Xu, H.: Weakly-supervised action localization with expectation-maximization multi-instance learning. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16. pp. 729–745. Springer (2020)
24. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S., McClosky, D.: The stanford corenlp natural language processing toolkit. In: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations. pp. 55–60 (2014)
25. Mun, J., Cho, M., Han, B.: Local-global video-text interactions for temporal grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10810–10819 (2020)
26. Neal, R.M., Hinton, G.E.: A view of the em algorithm that justifies incremental, sparse, and other variants. Learning in graphical models pp. 355–368 (1998)
27. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
28. Qu, M., Tang, J.: Probabilistic logic neural networks for reasoning. Advances in neural information processing systems **32** (2019)
29. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
30. Richardson, M., Domingos, P.: Markov logic networks. Machine learning **62**, 107–136 (2006)
31. Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T., Schiele, B.: Grounding of textual phrases in images by reconstruction. In: European Conference on Computer Vision. pp. 817–834. Springer (2016)
32. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10444–10452 (2019)
33. Su, R., Yu, Q., Xu, D.: Stvgbert: A visual-linguistic transformer based framework for spatio-temporal video grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1533–1542 (2021)

34. Tang, Z., Liao, Y., Liu, S., Li, G., Jin, X., Jiang, H., Yu, Q., Xu, D.: Human-centric spatio-temporal video grounding with visual transformers. IEEE Transactions on Circuits and Systems for Video Technology (2021)
35. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
36. Wang, L., Huang, J., Li, Y., Xu, K., Yang, Z., Yu, D.: Improving weakly supervised visual grounding by contrastive knowledge distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14090–14100 (2021)
37. Xiao, F., Sigal, L., Jae Lee, Y.: Weakly-supervised visual grounding of phrases with linguistic structures. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5945–5954 (2017)
38. Yang, A., Miech, A., Sivic, J., Laptev, I., Schmid, C.: Tubedetr: Spatio-temporal video grounding with transformers. arXiv preprint arXiv:2203.16434 (2022)
39. Yang, W., Zhang, T., Zhang, Y., Wu, F.: Local correspondence network for weakly supervised temporal sentence grounding. IEEE Transactions on Image Processing **30**, 3252–3262 (2021)
40. Zeng, R., Xu, H., Huang, W., Chen, P., Tan, M., Gan, C.: Dense regression network for video grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10287–10296 (2020)
41. Zhang, Z., Lin, Z., Zhao, Z., Zhu, J., He, X.: Regularized two-branch proposal networks for weakly-supervised moment retrieval in videos. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 4098–4106 (2020)
42. Zhang, Z., Zhao, Z., Lin, Z., Huai, B., Yuan, J.: Object-aware multi-branch relation networks for spatio-temporal video grounding. In: Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence. pp. 1069–1075 (2021)
43. Zhang, Z., Zhao, Z., Zhao, Y., Wang, Q., Liu, H., Gao, L.: Where does it exist: Spatio-temporal video grounding for multi-form sentences. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10668–10677 (2020)
44. Zhao, F., Li, J., Zhao, J., Feng, J.: Weakly supervised phrase localization with multi-scale anchored transformer network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5696–5705 (2018)
45. Zheng, M., Huang, Y., Chen, Q., Peng, Y., Liu, Y.: Weakly supervised temporal sentence grounding with gaussian-based contrastive proposal learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15555–15564 (2022)