Asymmetric Relation Consistency Reasoning for Video Relation Grounding

Huan Li, Ping Wei*, Jiapeng Li, Zeyu Ma, Jiahui Shang, and Nanning Zheng

Xi'an Jiaotong University, Xi'an, China

Abstract. Video relation grounding has attracted growing attention in the fields of video understanding and multimodal learning. While the past years have witnessed remarkable progress in this issue, the difficulties of multi-instance and complex temporal reasoning make it still a challenging task. In this paper, we propose a novel Asymmetric Relation Consistency (ARC) reasoning model to solve the video relation grounding problem. To overcome the multi-instance confusion problem, an asymmetric relation reasoning method and a novel relation consistency loss are proposed to ensure the consistency of the relationships across multiple instances. In order to precisely localize the relation instance in temporal context, a transformer-based relation reasoning module is proposed. Our model is trained in a weakly-supervised manner. The proposed method was tested on the challenging video relation dataset. Experiments manifest that the performance of our method outperforms the state-of-the-art methods by a large margin. Extensive ablation studies also prove the effectiveness and strength of the proposed method.

Keywords: Video relation grounding, asymmetric relation consistency, weakly-supervised

1 Introduction

Video relation grounding (VRG) plays a crucial role in cross-modal understanding of visual scene and natural language, which has been attracting increasing attention for its significance in applications such as video caption [29] and visual question answering [12]. Given an untrimmed video and a 3-tuple query relation description $\langle subject, predicate, object \rangle$, the task is to return the spatial and temporal ranges of the *subject* and *object* in the relation connected by *predicate* [34], as shown in Fig.1 (a). Usually the spatial and temporal ranges are represented as a temporal sequence of bounding boxes containing the entities [34], e.g. the blue box sequence of $subject : person$ and the brown box sequence of $object : bicycle$ connected by $predicate : ride$ in Fig. 1 (a).

Video relation grounding is a challenging problem for two reasons. First, VRG needs to localize fine-grained spatial and temporal locations of the *subject* and the *object* in a weakly-supervised manner, which means in training only video-level labels are provided, but without the spatial or temporal locations. Second,

* Ping Wei is the corresponding author. Email: pingwei@xjtu.edu.cn.

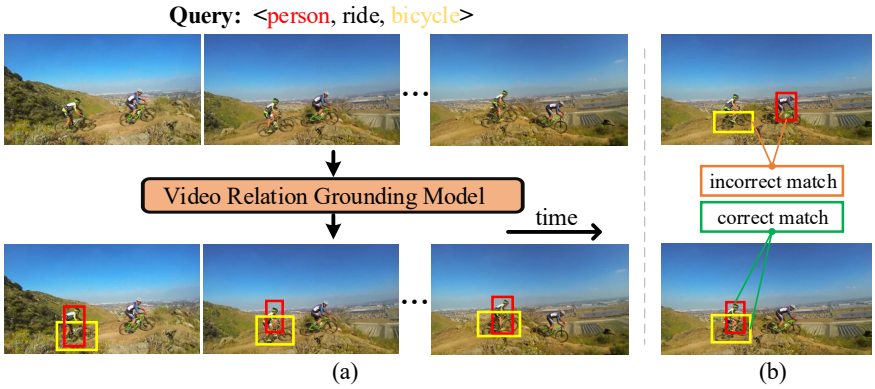

Fig. 1. Illustration of video relation grounding (VRG). (a) The VRG task requires the model to return one of the relation instance in both spatial and temporal domains. (b) Multi-instance confusion in video relation grounding.

compared with the other related video grounding tasks [4,24,37,31] which focus on the localization of temporal intervals or a single target object, VRG is defined to jointly localize a pair of object entities in both spatial and temporal domains.

The well-established VRG model [34] formulates this task as a hierarchical spatio-temporal region graph and achieved state-of-the-art results. However the *multi-instance confusion* remains to be an unsolved problem which greatly impedes the performance improvement. As shown in Fig. 1 (b), the video frame contains two instances of the relation $\langle person, ride, bicycle \rangle$. In a specific frame of the video, the grounding system may output the *subject* box engaged in one relation instance but the *object* box in another instance, which forms incorrect match pair. One of the major reasons is that the *subject* and the *object* are assumed to be symmetric or conditionally independent for relation reasoning. In this way, the semantic dependency and the spatial relationships between them would not be taken into consideration. Consequently the *subject* box and the *object* box may separately appear in different instances of the relation.

In this paper, we contend that the *subject* and the *object* are asymmetric in relation reasoning and propose a novel Asymmetric Relation Consistency (ARC) model to ground relations in videos. Different from the symmetric reasoning approaches [34,10], our model first localizes the *subject* box in each frame, and then the *object* box is searched for conditioned on the localized *subject* box. In turn, the *subject* box is sought again conditioned on the found *object* box. Intuitively, these two query results of *subject* should follow the same distribution. To model this consistency, we design a new relation consistent loss. Furthermore, to learn more precise relation semantic representation and mitigate the impact of data biases, we further propose a transformer-based [5] relation-aware reasoning module which utilizes the relation phrase of context to search for the most relevant relation duration. Similar to vRGV [34], our model is trained in a

weakly-supervised manner which means only the video-level labels are available but without the fine-grained spatio-temporal annotations.

The proposed method was tested on the challenging relation dataset ImageNet-VidVRD [22]. Experimental results manifest that our model outperforms the SOTA methods by a large margin. Extensive ablation studies also prove the effectiveness of the model.

This paper makes three contributions. Firstly, it proposes a novel asymmetric relation consistency reasoning method and designs novel loss functions for video relation grounding. Secondly, it presents a transformer-based baseline and a transformer-based relation-aware module to reason about the relationship between relation phrases and video features. Finally, the performance of the proposed method outperforms the state-of-the-art methods by a large margin.

2 Related Work

Since videos are not simple sets of separate object trajectories, modeling the interactions between two different object instances in videos enables us to deeply understand scenes and videos. Visual relation has been studied for a long time and made significant progress in recent years [15,35,36,17,14,30,6,9,26,11,33,16]. Many studies paid attention to video relation detection [22,27,25,13,19], which aims to spatio-temporally detect all the relation instances from untrimmed videos [22]. The work [22] proposes a segment-based method, where the segment-level relation class is obtained by a classifier and the final relation instance is obtained by a greedy relational association algorithm. Shang *et al.* [21] proposed a iterative inference method that effectively enhances the performance of visual video relation detection. Recently, a transformer-based method [7] is proposed to solve video relation detection task, where the relation instances are detected by set prediction. Unlike the relation detection task, Xiao *et al.* [34] proposed a more challenging task named visual relation grounding in videos, where the *subject* and *object* trajectories are localized by adopting a symmetric method with parameter-shared modules. Although this solution has achieved impress results, it overlooks the conditionally dependence between the *subject* and the *object*. In this paper, we propose an asymmetric relation reasoning method to further solve the video relation grounding problem.

The studies related to our method are vRGV [34] and SSAS [10]. The main differences are three-fold. First, these two studies adopt symmetric modes to localize *subject* and *object* with parameter-shared modules, while our model employs an asymmetric scheme. Second, the previous approaches model the relationship between *subject* and *object* by using implicit attention shifting. Instead, our model explicitly utilizes the conditional pattern. Third, vRGV takes the graph module and message passing to compute attention distribution, and SSAS uses convolution and iterative inference. Our model designs two new localizers based on transformers [5] to reason about the attention distribution.

Video grounding [4,24,37] aims to localize the temporal intervals of the targets in an untrimmed video by referring to the given sentence query. It provides

backbones for some more high-level tasks such as video caption. Considering the fact that fine-grained annotation of video is time-consuming, some studies have focused on weakly-supervised video grounding [31,3,32] that means the frame-level annotations are unavailable during training. The work [23] proposes a multi-instance learning based method, where the contextual similarity is considered to model the similarity between two frames and a visual clustering loss is proposed to learn visual features. AsyNCE-CMT [3] proposes a novel AsyNCE loss and uses a cross modal transformer block to advance the weakly-supervised video grounding task. However, these existing methods are not suitable for the video relation grounding task. The video relation grounding is required to seek a pair of objects of the relation instances and their temporal ranges. Modeling the dependency between the two object entities and the temporal continuity are key problems for video relation grounding.

3 Method

3.1 Formulation

We follow the work vRGV [34] to define the problem of video relation grounding (VRG). A video V of n frames is represented as a sequence of region proposals $V = (B_1, ..., B_n)$, where B_i is the set of regions in the ith frame. In each frame, m regions are proposed without labels or scores as $B_i = \{B_{i,j} \mid j = 1, ..., m\}$, where $B_{i,j}$ represents the jth region in the ith frame. A relation R is defined as a 3-tuple $R = \langle subject, predicate, object \rangle$, which usually means the *subject* is doing some actions described by *predicate* towards or with the *object*, such as $R = \langle person, ride, bicycle \rangle$. Relations in videos are not only dependent on features in each frame but also temporal information over a video segment span.

Given a video V of n frames and a relation $R = \langle subject, predicate, object \rangle$, video relation grounding aims to spatially and temporally localize the *subject* and the *object* connected by *predicate* in the video. With the region proposal representation of V, VRG is represented as to predict a *subject* box sequence $S = (S_k, ..., S_l)$ and an *object* box sequence $O = (O_k, ..., O_l)$, where $k, l \in [1, n]$ and $k < l$. S_i and O_i ($i \in [k, l]$) are the *subject* box and the *object* box in the ith frame, respectively, which come from the region proposal set B_i.

Following vRGV [34], VRG task is formulated as to solve the maximization problem:

$$(S^*, O^*) = \arg\max_{S,O} P(S, O|V, R)P(R|S, O, V). \tag{1}$$

The term $P(S, O|V, R)$ describes the joint posterior probability of the *subject* box sequence and the *object* box sequence. $P(R|S, O, V)$ characterizes the reconstruction of relation R given S, O and V. The final optimal output (S^*, O^*) is achieved by jointly maximizing the posterior and reconstruction terms. The above VRG definition and Eq. (1) follows the work [34]. One subtle difference is that the predicted *subject* box or the *object* box in our work represent the bounding box of the proposed region, not the region itself with the content.

Asymmetry in Relation. The previous method [34] hypothesize that the variable S and O are conditionally independent given V and R, i.e., $P(S, O|V, R) = P(S|V, R) * P(O|V, R)$. Thus the *subject* and the *object* are symmetric in the relation reasoning and the model uses a parameter-shared structure to localize the S and O respectively. However, this hypothesis overlooks two problems in VRG. First, in multi-instance scenarios, the co-occurrence of multiple $\langle subject, predicate, object\rangle$ instances may confuse the grounding system. For example, the system may return a *subject* sequence S in one relation instance but an *object* sequence O in another relation instance. Second, seeking the *subject* sequence and the *object* sequence separately ignores the semantic dependency and spatial relationships between the *subject* and *object*.

We contend that the *subject* and the *object* are asymmetric in relation reasoning and propose a novel Asymmetric Relation Consistency (ARC) model to resolve the above issues. In our model, grounding S and O are conditionally dependent given V and R. Correspondingly, our model first grounds the *subject* box in each frame according to the *subject* word embedding, and then the *object* box is searched for conditioned on the *subject* box. In turn, the *subject* box is sought again conditioned on the *object* box. Intuitively, the two query results of the *subject* should follow the same distribution, which is expressed with a novel asymmetric relation consistent loss. Since this strategy considers the semantic and spatial dependencies of the *subject* and the *object*, the multi-instance confusion can be alleviated and therefore the performance is improved.

3.2 Architecture Overview

Fig. 2 shows the overall architecture of the proposed method. It adopts a spatio-temporal detached way to conduct video relation grounding. First, we use the backbone network to extract region proposals and visual features from video frames and relation phrases. Second, the proposed asymmetric relation consistency reasoning module is utilized to localize the spatial position of the *subject* and *object* in each frame. Then the proposed relation-aware temporal reasoning module is utilized to compute the temporal boundaries of the given relation. Finally, the relation reconstruction module reconstructs the given relation based on the grounded results. The details of each part are described as follows.

3.3 Backbone Network

As shown in Fig. 2, given a video with n frames, we first use a pretrained Faster R-CNN [20] to generate m region proposals for each frame and extract the corresponding ROI-aligned regional features with ResNet [8]. Let $\boldsymbol{x}_i \in \mathbb{R}^{m\times d}$ be the features of m regions in the ith frame, where each row of $\boldsymbol{x}_i$ represents the features extracted from a region and d is the feature dimension. The spatial feature of each region is a 1×5 vector $\left[\frac{x_{min}}{W}, \frac{y_{min}}{H}, \frac{x_{max}}{W}, \frac{y_{max}}{H}, \frac{area}{W*H}\right]$, where W, H are the width and height of the frame, respectively. $(x_{min}, y_{min}, x_{max}, y_{max})$ is the bounding box position and $area$ is the area of the box. The spatial features of all the regions in the ith frame form a $m \times 5$ matrix.

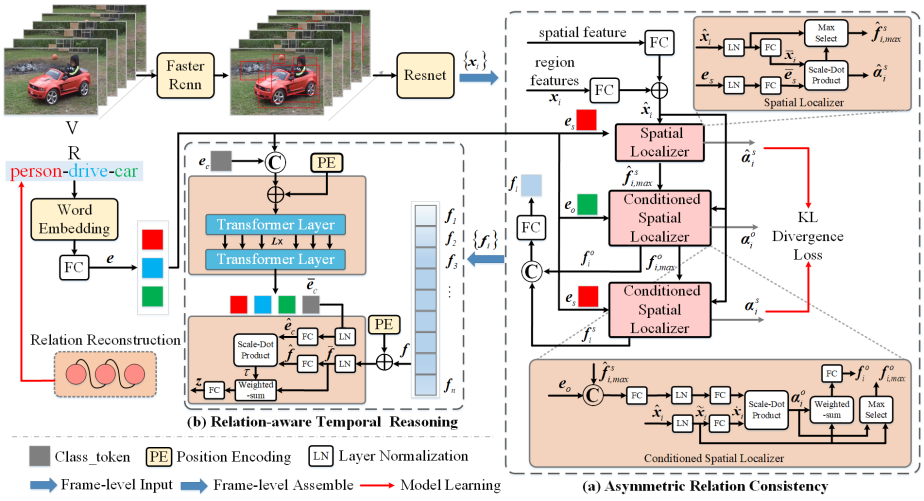

Fig. 2. The overall architecture of the proposed method.

A liner layer is used to transform the region features $\boldsymbol{x}_i$ into features of dimension $m \times D$. We learn the position embedding by mapping the $m \times 5$ spatial feature matrix into features of dimension $m \times D$. We then add the position embedding features to the transformed region features as input region features $\hat{\boldsymbol{x}}_i \in \mathbb{R}^{m \times D}$. Similar to vRGV [34], for each word in the relation phrase, we use Glove [18] to extract 300-dimensional word embeddings. Then we transform the word embedding into the same dimension with the input region features. It is represented as $\boldsymbol{e} \in \mathbb{R}^{l \times D}$, where l represents the length of the relation phrase.

3.4 Asymmetric Relation Consistency Reasoning

The asymmetric relation consistency reasoning module is designed to localize the spatial boxes of *subject* and *object* in each frame. As discussed in 3.1, grounding the subject and object boxes symmetrically ignores the spatial and semantic dependence between the subject and object, which may result in poor performance in the multi-instance scenes, as shown in Fig. 1 (b). To overcome this limitation, our model adopts an asymmetric reasoning scheme. Inspired by conditional dependency intuitively, reasoning about the object box based on the subject box is convenient for reducing the search space, thereby avoiding confusion of multi-instance, and vice versa. As shown in Fig. 2 (a), the asymmetric relation reasoning module contains two key components: spatial localizer and conditioned spatial localizer. The asymmetric reasoning process is composed of four steps: localizing *subject*, localizing *object* based on *subject*, re-localizing *subject* based on *object*, and consistency evaluation.

Step 1: Localizing *subject*. Given the input region features $\hat{\boldsymbol{x}}_i$ and relation phrase features $\boldsymbol{e}$, the task in this step is to localize the *subject* box with the

spatial localizer. In this case, we only use the *subject* word embedding feature $\boldsymbol{e}_s$. Our spatial localizer is built based on the self-attention structure [5,28], as shown in Fig. 2 (a).

We first normalize $\boldsymbol{e}_s$ and $\hat{\boldsymbol{x}}_i$ by a layer normalization layer, then following a fully-connected layer to map the normalized features into $\bar{\boldsymbol{e}}_s$ and $\bar{\boldsymbol{x}}_i$, respectively. Then the *subject* attention is computed by a cross attention [28] operation:

$$\hat{\alpha}^s_{i,j} = \frac{\exp(\beta^j_i)}{\sum_{j=1}^m \exp(\beta^j_i)}, \boldsymbol{\beta}_i = \frac{\bar{\boldsymbol{e}}_s(\bar{\boldsymbol{x}}_i)^T}{\sqrt{d_k}}, \tag{2}$$

where $\frac{1}{\sqrt{d_k}}$ is a scaling factor. $\hat{\boldsymbol{\alpha}}^s_i = \left\{\hat{\alpha}^s_{i,j}\right\}_{j=1}^m$ is the subject-aware attention distribution, which reflects the score of each region proposal in the ith frame. Based on the score, we select the most relevant region box as the current *subject* box and output the corresponding region feature $\hat{\boldsymbol{f}}^s_{i,max}$.

Step 2: Localizing *object* based on *subject*. In this step, we use the proposed conditioned spatial localizer, an shown in Fig. 2 (a), to localize the *object* box based on the localized *subject* box in Step 1. Concretely, we first concatenate $\hat{\boldsymbol{f}}^s_{i,max}$ and the word embedding of *object* $\boldsymbol{e}_o$, and input it into a fully-connected layer with relu activation function and a layer normalization layer. The input region features $\hat{\boldsymbol{x}}_i$ are also normalized as $\widetilde{\boldsymbol{x}}_i$ by a layer normalization layer, then following a fully-connected layer to map the normalized features $\widetilde{\boldsymbol{x}}_i$ into $\dot{\boldsymbol{x}}_i$. Following the similar process implemented in the spatial localizer, we can obtain the object-aware attention distribution $\boldsymbol{\alpha}^o_i = \left\{\alpha^o_{i,j}\right\}_{j=1}^m$ and the feature $\boldsymbol{f}^o_{i,max}$ of the most relevant region of *object*. The object-aware frame feature is computed as:

$$\boldsymbol{f}^o_i = g(\sum_{j=1}^m \alpha^o_{i,j}\widetilde{\boldsymbol{x}}_{i,j}), \tag{3}$$

where $\widetilde{\boldsymbol{x}}_{i,j}$ is the jth row of $\widetilde{\boldsymbol{x}}_i$, i.e. the normalized feature of the jth box. g is a fully-connected layer.

Step 3: Re-localizing *subject* based on *object*. With the feature $\boldsymbol{f}^o_{i,max}$ obtained in Step 2, we re-calculate the attention weight of the region proposals for *subject*. The $\boldsymbol{f}^o_{i,max}$ and $\boldsymbol{e}_s$ are input into the same conditioned spatial localizer to regenerate the region proposal attention weight and output $\boldsymbol{\alpha}^s_i = \left\{\alpha^s_{i,j}\right\}_{j=1}^m$ and compute the subject-aware frame feature $\boldsymbol{f}^s_i$. This regenerating operation implies that the most relevant bounding box of *subject* is searched for based on the most relevant box of *object*.

Step 4: Consistency evaluation. Intuitively, we also expect the most relevant box of *object* is searched for based on the most relevant box of *subject*. Unfortunately, one video may contain multiple relation instances, which results in smooth attention weight distribution for adjusting to multiple instances in Step 1. In this case, the most relevant box of *subject* selected in Step 1 may be inaccurate, which further causes difficulty for localizing *object* in Step 2. To overcome this problem, we expect the most relevant box selected in Step 1 is the

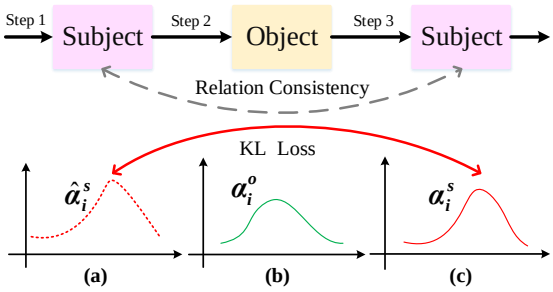

Fig. 3. Illustration of the asymmetric relation consistency. (a) Attention distribution of *subject* computed in Step 1. (b) Attention distribution of *object*. (c) Attention distribution of *subject* obtained in Step 3.

same as the box selected in Step 3. Thus, we can drive the attention distributions in Step 1 and Step 3 to follow the same distribution. As shown in Fig. 3. Specifically, we adopt the Kullback-Leibler divergence to optimize the learning process. Based on KL divergence, we propose a relation consistent loss function as follows:

$$\mathcal{L}_{arc} = \frac{1}{n}\sum_{i=1}^{n} D_{KL}(\boldsymbol{\alpha}_i^s \parallel \hat{\boldsymbol{\alpha}}_i^s), \tag{4}$$

where n is the frame number. $\boldsymbol{\alpha}_i^s$ is the attention distribution of *subject* computed in Step 3 and $\hat{\boldsymbol{\alpha}}_i^s$ is the one in Step 1. Since the grounding process employs an asymmetric and redistribution pattern, inspired by CycleGAN [38], we name this scheme as Asymmetric Relation Consistency (ARC).

Given the subject-aware frame feature $\boldsymbol{f}_i^s$ and object-aware frame feature $\boldsymbol{f}_i^o$, we concatenate these two features and then use a fully connected layer to get a final frame-level feature $\boldsymbol{f}_i$ for each frame. Thus the frame-level feature of the video can be donated as: $\boldsymbol{f} = \{\boldsymbol{f}_i\} \in \mathbb{R}^{n\times D}$.

3.5 Relation-Aware Temporal Reasoning

In the asymmetric relation reasoning process, a latent hypothesis is that all the frames contain the given relation instance. However, it is inapplicable in the VRG task since it requires not only localizing the spatial positions but also the temporal boundaries. In this section, based on the frame-level features extracted by the asymmetric relation consistency reasoning module, we propose a transformer-based relation-aware temporal reasoning (RTR) method for precisely localizing the temporal boundaries.

Since transformer [5,1] possesses the ability to extract global context information of long sequences, we use transformer to learn the relation semantics. We take the given relation as a phrase: *subject-predicate-object*. As shown in Figure 2 (b), our RTR module receives the frame-level feature and word embeddings

of relation phrases as inputs. For learning global relation semantics, an extra learnable class token $\boldsymbol{e}_c$ is inserted into the word embedding sequence. Following the work [5], we add the fixed positional encoding into the word embedding sequence. Then the word embedding sequence is input into L successive transformer layers to learn relation representations. The relation representations are learned and meanwhile the relation of *subject* and *object* is reasoned implicitly. By multi-layer passing, the updated class token $\bar{\boldsymbol{e}}_c$ is used for temporal reasoning.

For temporal reasoning, we first supplement the frame-level feature with the fixed positional encoding and then normalize it with a normalization layer, where the normalized result is denoted as $\bar{\boldsymbol{f}}$. We use a fully-connected layer to map the updated class token $\bar{\boldsymbol{e}}_c$ and the normalized feature $\bar{\boldsymbol{f}}$ into $\hat{\boldsymbol{e}}_c$ and $\hat{\boldsymbol{f}}$, respectively. Finally, the temporal attention distribution is computed as,

$$\tau_i = \frac{\exp(\sigma_i)}{\sum_{i=1}^{n} \exp(\sigma_i)}, \boldsymbol{\sigma} = \frac{\hat{\boldsymbol{e}}_c(\hat{\boldsymbol{f}})^T}{\sqrt{d_k}}. \tag{5}$$

$\boldsymbol{\tau} = \{\tau_i\}_{i=1}^{n}$ is the frame-level attention distribution. The final video-level feature $\boldsymbol{z}$ about the given relation $\langle subject, predicate, object \rangle$ is represented as,

$$\boldsymbol{z} = h(\sum_{i=1}^{n} \tau_i \bar{\boldsymbol{f}}_i), \tag{6}$$

where h is a fully-connected layer.

3.6 Train and Inference

Following [34], we train our model using phrase reconstruction of the given relation in a weakly-supervised way. The reconstruction loss is represented as:

$$\mathcal{L}_{res} = -\sum_{t=1}^{l} \log(P(R_t \mid R_{0:t-1}, \boldsymbol{z})), \tag{7}$$

where l is the number of words in the relation phrase and R_t represents each word in the phrase. The total loss is described as:

$$\mathcal{L} = \mathcal{L}_{res} + \lambda \mathcal{L}_{arc}. \tag{8}$$

$\mathcal{L}_{arc}$ is the relation consistency loss defined in Eq. (4). λ is a hyper-parameter.

In inference, we employ the similar method used in vRGV [34] to obtain the *subject* box sequence S and the *object* sequence O. We first generate candidate segments set for each video in temporal dimension based on the learned temporal frame-level distribution $\boldsymbol{\tau}$ by setting a threshold η. For each candidate segment, we then use the Viterbi algorithm to search for an optimal path for the *subject* box sequence S and the *object* sequence O, respectively. The linking cost of the successive frames is defined as:

$$c(B_{i,p}, B_{i+1,q}) = \alpha_{i,p} + \alpha_{i+1,q} + \theta \cdot IoU(B_{i,p}, B_{i+1,q}), \tag{9}$$

where $B_{i,p}$ represent the pth region in ith frame. $\boldsymbol{\alpha}_i$ is the subject-aware attention distribution $\boldsymbol{\alpha}_i^s = \left\{\alpha_{i,j}^s\right\}_{j=1}^m$ or the object-aware attention distribution $\boldsymbol{\alpha}_i^o = \left\{\alpha_{i,j}^o\right\}_{j=1}^m$. For example, to obtain the *subject* box sequence S, $\boldsymbol{\alpha}_i^s$ is used to compute the linking cost. θ is a hyper-parameter and IoU is the intersection over union. We average linking cost of the searched *subject* box sequence S and *object* sequence O as the segment score, and then we select the segment with the maximal score as the relation grounding result.

4 Experiments

4.1 Settings

Implementation details. Our model is built on the basic transformer configuration [5]. Following vRGV [34], each video is sampled $n = 120$ frames and the number of proposals for each frame is set to 40. The region features are extracted from the pretrained Faster R-CNN [20] with the backbone ResNet101 [8]. The region features and spatial features along with the word embeddings are transformed into the same dimension $D = 512$. All the experiments are conducted on 8 NVIDIA 3090 GPUs and The batch size is set to 32 for each GPU. We use the Pytorch toolbox with FP16 training. The model is trained with Adam optimizer with basic learning rate 1e-4.

Dataset and evaluation criteria. We test our model on the challenging ImageNet-VidVRD video relation dataset [22]. It consists of 1000 videos, 35 object classes, 132 predicate classes, and over 30,000 relation instances.

Our model is evaluated and compared with the previous studies using accuracy (Acc). Given a video V and the corresponding relation 3-tuple R, a result is a true positive if the tIoUs (temporal intersection over union) of *subject* box sequence S and *object* box sequence O with one of the ground-truth instance are both larger than 0.5. The tIoU is computed under three different spatial intersection over union (sIoU) thresholds (0.3, 0.5 and 0.7). Following vRGV [34], we report the whole relation accuracy (Acc_R), the subject accuracy (Acc_S) and the object accuracy (Acc_O), respectively.

4.2 Result Comparison and Analysis

Table 1 shows the experiment result comparisons on different spatial overlap thresholds. We compare our model with some previous methods: T-Rank [2], Co-occur [10], and vRGV [34], where the results of T-Rank [2] and Co-occur [10] were reported in [34]. We also report the results obtained by selecting the regions with the maximal attention scores without using IoU in inference (marked with *). We also compare our model with the baseline method which is designed based on the basic transformers. From Table 1, we can find our ARC model performs better than the baseline model under all threshold settings. And it is obvious that the ARC model outperforms all the existing methods. Specifically, compared to the SOTA method vRGV, our method gets 45.66%, 44.01%, and 32.53% performance

Table 1. Video relation grounding comparison on different spatial overlap thresholds (Acc %). * means not using IoU in inference).

Models	sIOU=0.3			sIOU=0.5			sIOU=0.7			Average		
	Acc_S	Acc_O	Acc_R	Acc_S	Acc_O	Acc_R	Acc_S	Acc_O	Acc_R	Acc_S	Acc_O	Acc_R
T-Rank V1[2]	33.55	27.52	17.25	22.61	12.79	4.49	6.31	3.30	0.76	20.27	10.68	3.99
T-Rank V2[2]	34.35	21.71	15.06	23.00	9.18	3.82	7.06	2.09	0.50	20.83	7.35	3.16
Co-occur*[10]	27.84	25.62	18.44	23.50	20.40	13.81	17.02	14.93	7.29	22.99	19.33	12.80
Co-occur[10]	31.31	30.65	21.79	28.02	27.69	18.86	21.99	21.64	13.16	25.90	25.23	16.48
vRGV*[34]	37.61	37.75	27.54	32.17	32.32	21.43	21.34	21.02	10.62	31.64	30.92	20.54
vRGV[34]	42.31	41.31	29.95	37.11	37.52	24.77	29.71	29.72	17.09	36.77	36.30	24.58
Our baseline*	40.94	38.76	29.13	35.18	33.69	23.77	27.03	25.46	13.94	34.08	32.75	22.97
Our baseline	41.41	38.86	29.84	36.75	35.14	24.78	29.66	27.78	15.43	35.58	34.60	24.38
ARC*	41.60	40.61	30.23	37.13	36.78	26.09	28.65	29.41	17.56	34.96	34.72	23.75
ARC	**45.66**	**44.01**	**32.53**	**40.99**	**40.41**	**27.83**	**33.24**	**33.39**	**20.44**	**39.66**	**39.20**	**26.42**

Table 2. Video relation grounding comparison on different temporal overlap thresholds (Acc %).

Models	tIOU=0.3			tIOU=0.5			tIOU=0.7		
	Acc_S	Acc_O	Acc_R	Acc_S	Acc_O	Acc_R	Acc_S	Acc_O	Acc_R
T-Rank V1[2]	36.51	28.67	15.05	20.27	10.68	3.99	6.15	2.67	0.55
T-Rank V2[2]	36.99	20.70	12.81	20.83	7.35	3.16	6.19	1.30	0.21
Co-occur[10]	35.30	35.50	23.23	25.90	25.23	16.48	16.81	15.04	8.94
vRGV[34]	49.97	48.98	33.16	36.77	36.30	24.58	24.27	22.11	13.69
Our baseline*	49.05	46.43	33.75	34.08	32.75	22.97	22.05	19.62	10.94
Our baseline	49.72	47.83	34.27	35.58	34.60	24.38	24.53	21.58	12.08
ARC*	49.61	49.43	35.68	34.96	34.72	23.75	24.14	25.25	14.46
ARC	**52.74**	**52.41**	**35.61**	**39.66**	**39.20**	**26.42**	**28.68**	**28.68**	**17.67**

for Acc_S, Acc_O, and Acc_R, respectively under the threshold $sIoU = 0.3$, while the SOTA method vRGV achieves 42.31% Acc_S, 41.31% Acc_O, and 29.95% Acc_R, respectively. Under the threshold $sIoU = 0.5$, our model achieves 27.83% Acc_R, 40.99% Acc_S and 40.41% Acc_O, respectively, and outperforms vRGV by a large margin. Under the setting of $sIoU = 0.7$, the performance of our ARC still has a significant improvement. These results illustrate the effectiveness of the proposed model.

To make a deeper comparison, we compare the results obtained without using IoU as the linking cost in inference. In this case, the model is required to possess the capacity of modeling temporal continuity for localizing the relation instance. In Table 1, the performance of our model still exceeds the SOTA method vRGV. We also compare our method with the previous models on different temporal overlap thresholds shown in Table 2, where our method outperform all previous models under all settings. These results prove that our model has the ability to capture the relation temporal continuity.

Table 3. Ablation study results on ImageNet-VidVRD dataset (Acc %).

Models	sIOU=0.3			sIOU=0.5			sIOU=0.7			Average		
	Acc_S	Acc_O	Acc_R	Acc_S	Acc_O	Acc_R	Acc_S	Acc_O	Acc_R	Acc_S	Acc_O	Acc_R
w/o ARC	43.03	42.19	31.99	38.70	37.89	27.08	30.18	29.59	16.13	37.08	36.43	24.86
w/o RC Loss	42.75	41.76	30.88	37.20	38.43	25.48	29.53	31.13	16.67	35.99	36.85	23.61
w/o RTR	43.86	43.97	33.03	38.04	38.42	26.25	30.34	29.32	17.39	36.59	36.58	23.86
ARC	**45.66**	**44.01**	**32.53**	**40.99**	**40.41**	**27.83**	**33.24**	**33.39**	**20.44**	**39.66**	**39.20**	**26.42**

4.3 Ablation Study

Effect of asymmetric relation consistency. In this section, we validate the effect of the asymmetric consistency reasoning and compare it with the symmetric method. Table 3 shows the ablation comparison results, where 'w/o' means 'without'. w/o ARC represents the model without the asymmetric reasoning and it achieves 16.13 $\%Acc_R$ under $sIoU = 0.7$ setting and 24.86 $\%Acc_R$ under *Average* setting, while our model ARC surpasses it by a large margin. This phenomenon manifests the asymmetric consistency reasoning can significantly improve the performance and plays an important role in relation grounding. As discussed in 3.1, the proposed asymmetric reasoning method can mitigate the multi-instance confusion problem and learn the semantic dependency between *subject* and *object*, and thus the performance is improved.

Influence of relation consistent loss. The second row in Table 3 shows the results of asymmetric relation reasoning without relation consistency KL loss. We can find that removing the consistency supervision significantly impairs the performance of the system, where the model get a 23.61% Acc_R in *Average* setting and the ability of the model nearly degenerates into the baseline level. Without the consistency loss, the model can hardly select the most relevant box for *subject* with the spatial localizer in Step 1, and further results in a false region for *object* in Step 2, which leads to the performance decline.

We visualize the attention distributions learned in Step 1 and Step 3. As shown in Fig. 4, in order to validate the ability to cope with the multi-instance cases, we implement the comparison on multi-instance videos. The first row shows the model without using the relation consistent loss (w/o RC Loss) while the second row shows the results from our ARC. When the relation consistent loss was removed, the attention distribution $\hat{\boldsymbol{\alpha}}^s$ learned in Step 1 becomes quite smooth, thus the most relevant region for *subject* is difficult to be selected. This difficulty will influences the localizing process of Step 2, thereby indirectly making the Step 3 be caught in a dilemma. On the contrary, training the model with KL loss makes the attention distribution univocal and pushes $\hat{\boldsymbol{\alpha}}^s$ and $\boldsymbol{\alpha}^s$ follow the similar distribution, which reduces the difficulty for relation grounding and thereby enhances the performance.

Effect of relation-aware temporal reasoning. Our ARC uses the proposed transformer-based relation-aware representation module (RTR). We compare ARC with the method without using the relation-aware temporal reasoning (w/o RTR), which merges the *subject* embedding and *object* embedding as query

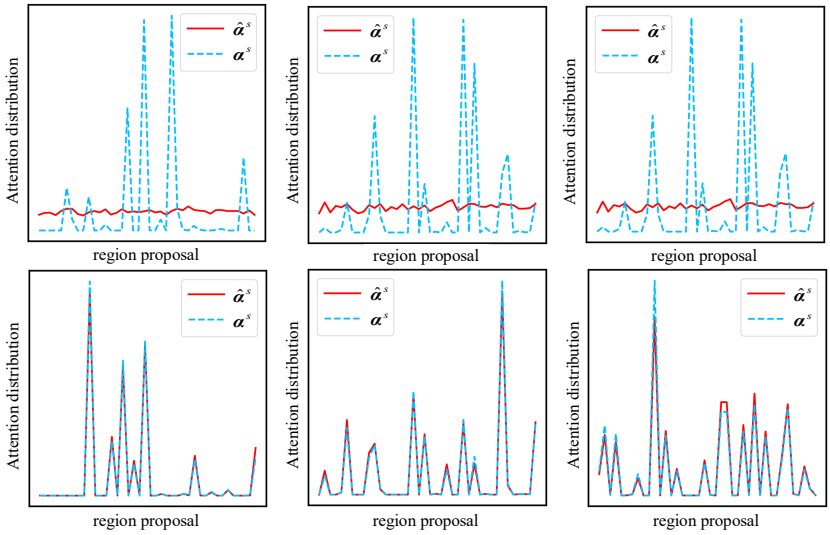

Fig. 4. Illustration of the effect of the relation consistent loss.

to localize the given relation. ARC outperforms w/o RTR by a large margin, which adequately demonstrates the validity of the proposed RTR module.

As shown in Fig. 5, we visualize video-level feature $\boldsymbol{z}$ that is used to reconstruct the given relation. All features are extracted from the relations with the same subject *person* and object *bicycle* but maybe different predicates. In Fig. 5, the same color represents the relations with same subject, object and predicate. The figure shows that the model without RTR module (w/o RTR) is apt to regard the relation with different predicates as the same relations. As shown in the red circles, the different relations closely intertwine and are inseparable. We attribute this phenomenon to data biases. For example, the video only consists of one person and one bicycle, thus the model without RTR module will neglect the predicate and directly localize the person and bicycle to get a plausible result. However this scheme may result in performance degradation in complex scenes with multiple relation instances. While our model (ARC) can effectively separate different relations and mitigate the influence of data biases, thereby improving the performance.

4.4 Zero-shot Evaluation

Due to the diversity of relations, many new relation triplets do not appear in the training set. Thus, the ability to handle the zero-shot problem is vitally important in video relation grounding. For zero-shot evaluation, we compare our model with some previous models. Table 4 shows the comparison results, where our model still outperforms the existing methods. In this case, the SOTA method vRGV achieves 10.27 % Acc_R, while our model gets 11.19% Acc_R. These

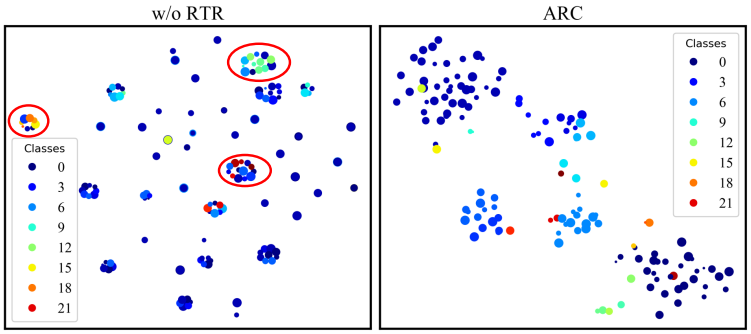

Fig. 5. Illustration of the effect of the relation-aware temporal reasoning.

Table 4. Zero-shot evaluation results on ImageNet-VidVRD (Acc %).

Model	Acc_S	Acc_O	Acc_R
T-Rank V1[2]	4.05	4.08	1.37
T-Rank V2[2]	7.09	4.13	1.37
Co-occur [10]	11.60	10.99	7.38
vRGV[34]	18.94	17.23	10.27
Our model	17.34	**19.01**	**11.19**

results verify the generalization capability and the power of our model to solve the zero-shot problem in video relation grounding task.

5 Conclusion

This paper addresses the challenging problem of weakly-supervised video relation grounding. The existing methods adopted symmetric reasoning schemes without considering the dependency between the subject and the object. We propose a novel asymmetric relation reasoning method with a relation consistency loss to overcome this weakness. A transformer-based relation-aware relation reasoning module is proposed to learn a better relation representation. The extensive experiments proved the effectiveness of the proposed method. The future work will focus on exploring the asymmetry mechanism in other grounding tasks.

Acknowledgement

This research was supported by the grants Key Research and Development Program of China (No. 2018AAA0102501), and National Natural Science Foundation of China (No. 61876149, No. 62088102).

References

1. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision (2020)
2. Chen, Z., Ma, L., Luo, W., Wong, K.Y.K.: Weakly-supervised spatio-temporally grounding natural sentence in video. In: The Annual Meeting of the Association for Computational Linguistics (2019)
3. Da, C., Zhang, Y., Zheng, Y., Pan, P., Xu, Y., Pan, C.: Asynce: Disentangling false-positives for weakly-supervised video grounding. In: ACM International Conference on Multimedia (2021)
4. Ding, X., Wang, N., Zhang, S., Cheng, D., Li, X., Huang, Z., Tang, M., Gao, X.: Support-set based cross-supervision for video grounding. In: IEEE CVPR (2021)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
6. Gao, C., Xu, J., Zou, Y., Huang, J.B.: Drg: Dual relation graph for human-object interaction detection. In: European Conference on Computer Vision (2020)
7. Gao, K., Chen, L., Huang, Y., Xiao, J.: Video relation detection via tracklet based visual transformer. In: ACM International Conference on Multimedia (2021)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)
9. Kim, B., Lee, J., Kang, J., Kim, E.S., Kim, H.J.: Hotr: End-to-end human-object interaction detection with transformers. In: IEEE CVPR (2021)
10. Krishna, R., Chami, I., Bernstein, M., Fei-Fei, L.: Referring relationships. In: IEEE Conference on Computer Vision and Pattern Recognition (2018)
11. Li, J., Wei, P., Zhang, Y., Zheng, N.: A slow-i-fast-p architecture for compressed video action recognition. In: ACM International Conference on Multimedia (2020)
12. Li, Q., Tao, Q., Joty, S., Cai, J., Luo, J.: Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions. In: European Conference on Computer Vision (2018)
13. Li, Y., Yang, X., Shang, X., Chua, T.S.: Interventional video relation detection. In: ACM International Conference on Multimedia (2021)
14. Liao, Y., Liu, S., Wang, F., Chen, Y., Qian, C., Feng, J.: Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In: IEEE CVPR (2020)
15. Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. In: European Conference on Computer Vision (2016)
16. Ma, Z., Wei, P., Li, H., Zheng, N.: Hoig: End-to-end human-object interactions grounding with transformers. In: IEEE International Conference on Multimedia and Expo (2022)
17. Mi, L., Chen, Z.: Hierarchical graph attention network for visual relationship detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
18. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Conference on Empirical Methods in Natural Language Processing (2014)
19. Qian, X., Zhuang, Y., Li, Y., Xiao, S., Pu, S., Xiao, J.: Video relation detection with spatio-temporal graph. In: ACM International Conference on Multimedia (2019)

20. 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 (2015)
21. Shang, X., Li, Y., Xiao, J., Ji, W., Chua, T.S.: Video visual relation detection via iterative inference. In: ACM International Conference on Multimedia (2021)
22. Shang, X., Ren, T., Guo, J., Zhang, H., Chua, T.S.: Video visual relation detection. In: ACM International Conference on Multimedia (2017)
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: IEEE CVPR (2019)
24. Soldan, M., Xu, M., Qu, S., Tegner, J., Ghanem, B.: Vlg-net: Video-language graph matching network for video grounding. In: IEEE/CVF International Conference on Computer Vision (2021)
25. Sun, X., Ren, T., Zi, Y., Wu, G.: Video visual relation detection via multi-modal feature fusion. In: ACM International Conference on Multimedia (2019)
26. Tamura, M., Ohashi, H., Yoshinaga, T.: Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In: IEEE CVPR (2021)
27. Tsai, Y.H.H., Divvala, S., Morency, L.P., Salakhutdinov, R., Farhadi, A.: Video relationship reasoning using gated spatio-temporal energy graph. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NIPS (2017)
29. Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., Saenko, K.: Sequence to sequence – video to text. In: IEEE International Conference on Computer Vision (2015)
30. Wang, T., Yang, T., Danelljan, M., Khan, F.S., Zhang, X., Sun, J.: Learning human-object interaction detection using interaction points. In: IEEE CVPR (2020)
31. Wang, W., Gao, J., Xu, C.: Weakly-supervised video object grounding via stable context learning. In: ACM International Conference on Multimedia (2021)
32. Wang, Y., Zhou, W., Li, H.: Fine-grained semantic alignment network for weakly supervised temporal language grounding. In: Findings of the Association for Computational Linguistics (2021)
33. Wei, P., Zhao, Y., Zheng, N., Zhu, S.C.: Modeling 4d human-object interactions for joint event segmentation, recognition, and object localization. In: IEEE Transactions on Pattern Analysis and Machine Intelligence. pp. 1165–1179 (2017)
34. Xiao, J., Shang, X., Yang, X., Tang, S., Chua, T.S.: Visual relation grounding in videos. In: European Conference on Computer Vision (2020)
35. Yu, R., Li, A., Morariu, V.I., Davis, L.S.: Visual relationship detection with internal and external linguistic knowledge distillation. In: IEEE International Conference on Computer Vision (2017)
36. Zhan, Y., Yu, J., Yu, T., Tao, D.: On exploring undetermined relationships for visual relationship detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
37. Zhao, Y., Zhao, Z., Zhang, Z., Lin, Z.: Cascaded prediction network via segment tree for temporal video grounding. In: IEEE CVPR (2021)
38. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE International Conference on Computer Vision (2017)