Co-Grounding Networks with Semantic Attention for Referring Expression Comprehension in Videos

Sijie Song[1], Xudong Lin[2], Jiaying Liu[1], Zongming Guo[1]*, Shih-Fu Chang[2]
[1] Wangxuan Institute of Computer Technology, Peking University, Beijing, China
[2] DVMM Lab, Columbia University, New York, NY, USA

Abstract

*In this paper, we address the problem of referring expression comprehension in videos, which is challenging due to complex expression and scene dynamics. Unlike previous methods which solve the problem in multiple stages (i.e., tracking, proposal-based matching), we tackle the problem from a novel perspective, **co-grounding**, with an elegant one-stage framework. We enhance the single-frame grounding accuracy by semantic attention learning and improve the cross-frame grounding consistency with co-grounding feature learning. Semantic attention learning explicitly parses referring cues in different attributes to reduce the ambiguity in the complex expression. Co-grounding feature learning boosts visual feature representations by integrating temporal correlation to reduce the ambiguity caused by scene dynamics. Experiment results demonstrate the superiority of our framework on the video grounding datasets VID and LiOTB in generating accurate and stable results across frames. Our model is also applicable to referring expression comprehension in images, illustrated by the improved performance on the RefCOCO dataset. Our project is available at* https://sijiesong.github.io/co-grounding.

1. Introduction

Referring expression comprehension has attracted much attention recently. It aims to localize a region of the image/video described by the natural language. This topic is of great importance in computer vision to support a variety of research problems such as image/video captioning [2, 27], visual question answering [3] and image/video retrieval [31, 9]. It also plays a key role in machine intelligence for a wide range of applications from human-computer interaction, robotics to early education.

In the past years, most of the previous work for referring expression comprehension focus on the grounding for static images [32, 34, 38, 18, 36, 23, 24, 12, 1] and have achieved promising results. However, referring expression comprehension for videos is less explored, which is challenging yet important. Different from several threads of referring expression comprehension in videos, such as referring all mentioned entities [42], we localize the spatio-temporal tube that semantically corresponds to the whole sentence. That is, we output bounding box for each frame as shown in Figure 1.

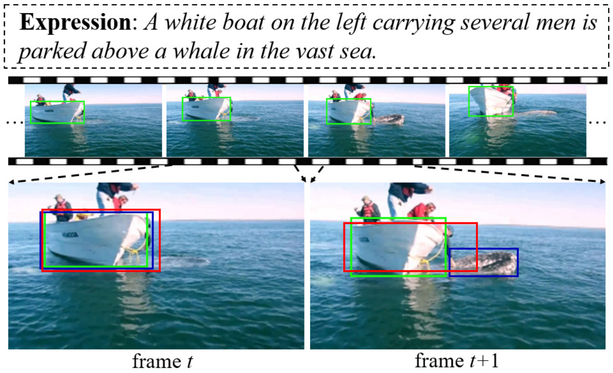

Expression: *A white boat on the left carrying several men is parked above a whale in the vast sea.*
...
...
frame *t*
frame *t*+1

Figure 1. Referring expression comprehension in videos. Due to dynamic scenes and ambiguity in the expression, per-frame inference (in blue) with state-of-the-art grounding method [36] would lead to unstable results across frames, while our co-grounding networks achieve accurate and consistent predictions (in red). Ground-truth annotations are denoted in green.

The work in [17] treats referring expression comprehension for videos as a tracking problem. We argue that it would suffer from template selection error, because it is hard to tell from the grounding results from multiple frames which is the right one to track. The other work [6] first proposes spatio-temporal tube candidates and then matches them with textual features from expression. However, the performance is limited by the proposal quality. Inspired by the advance in one-stage image grounding methods [36] that get rid of proposal detectors, another solution is to conduct per-frame inference with [36], but there are still two problems. Firstly, the entities in the expression (such as '*boat*', '*men*', '*whale*', '*sea*' in Figure 1) would cause ambiguity

[*] Corresponding author.

when encoded into textual features, making the model confused about which entity is the correct one to ground. Secondly, the dynamic scenes across frames would also interrupt the grounding process (see the drifting blue bounding boxes in Figure 1). Therefore, the key challenge is to generate robust textual and visual features to reduce ambiguity, then further achieve accurate and stable results across frames.

To tackle the aforementioned issues, we propose to solve referring expression comprehension in videos with a new perspective, *i.e.*, co-grounding, with semantic attention learning in an elegant one-stage framework. The basic structure of our model is based on YOLO [28], which predicts the bounding box and confidence simultaneously. The confidence reflects the matching score between the textual and visual features. We design a semantic attention mechanism to obtain attribute-specific features both for vision and language. Specifically, a proposal-free subject attention scheme is proposed to parse the words for subject from the expression. An object-aware location attention scheme is developed to parse the words for location from the expression. The interaction between attribute-specific textual and visual features determines the subject score and location score for each visual region (see the visualization examples in Figure 6). Besides, to improve cross-frame prediction consistency, we develop the co-grounding feature learning. Taking multiple frames as input, it utilizes the correlation across frames to enhance visual features and stabilize the grounding process in training and testing. A post-processing strategy is further employed to improve the temporal consistency during inference.

Our contributions are summarized as follows:

- We propose to solve referring expression comprehension in videos by co-grounding in an one-stage framework.
- We propose semantic attention learning to parse referring cues, including a proposal-free subject attention and object-aware location attention.
- Our networks are applicable to both video/image grounding, and achieve state-of-the-art performance on referring expression comprehension benchmarks.

2. Related Work

2.1. Referring expression comprehension

Benefiting from the advances in object detection [29, 28, 10], most methods [32, 34, 38] perform referring expression comprehension in two stages. Region candidates are proposed in the first stage with an object detector and then matched with the expression in the second stage. The best matching region is selected as the grounding result. However, these two-stage methods are limited by the proposal quality of the offline object detector. The missing of ground-truth regions in the first stage would lead to the failure of the second stage. To address the issue, more recent works are proposed to get rid of the offline object detector [18, 36]. Built upon the current one-stage object detection method, *i.e.*, YOLO-v3 [28], Yang *et al.* [36] first proposed an one-stage visual grounding framework, which extracts visual-text features and predict bounding boxes densely at all spatial locations. Liao *et al.* [18] reformulate referring expression comprehension as correlation filtering, where the filter template is generated from language features. Besides referring expression comprehension for images, there are a few works [6, 17] exploring the task in the video domain. However, both of the methods solve referring expression comprehension for videos with multiple stages (*i.e.*, tracking, proposal-based matching). The failure in the first stage would directly impact the final results. In our work, we propose an elegant one-stage framework for this task.

2.2. Attention mechanisms

Inspired by human perception, attention mechanisms have been widely studied in vision-and-language tasks, *e.g.*, visual question answering [41, 20, 14, 25], visual dialogue [33] and visual grounding [38]. In these works, attention is applied to learn the underlying correlation between different modalities [41, 20], which jointly performs language-guided visual attention and vision-guided language attention. Nguyen *et al.* [25] propose a dense symmetric co-attention to deal with every interaction between any pair of visual region and each word. To further fully understand the semantic of vision and language, a more recent work [14] presents a hypergraph model to define a common semantic space among different modalities. Attention mechanisms are also popular in the task of visual grounding to decompose the language into several components [38]. Each component focuses different attributes of the language and then trigger corresponding visual comprehension. However, the modular network designed in [38] is based on offline object proposals and requires external annotations. Our work, however, achieves semantic attention learning without the reliance of object proposals and external labels.

2.3. Temporal consistency

Generating consistent results across adjacent frames is essential for video applications.The recent works mainly focus on improving performance of per-frame result by exploiting information in the temporal domain [8, 43, 5, 16, 21, 11, 26, 4]. Some of the works aggregate local temporal context to help the inference of current frame. Optical flow is always computed by [7] to propagate features across frames [43, 4], while some methods integrate temporal context by calculating affinity matrix [8, 21] to build temporal correspondence. Nevertheless, only focusing the locality

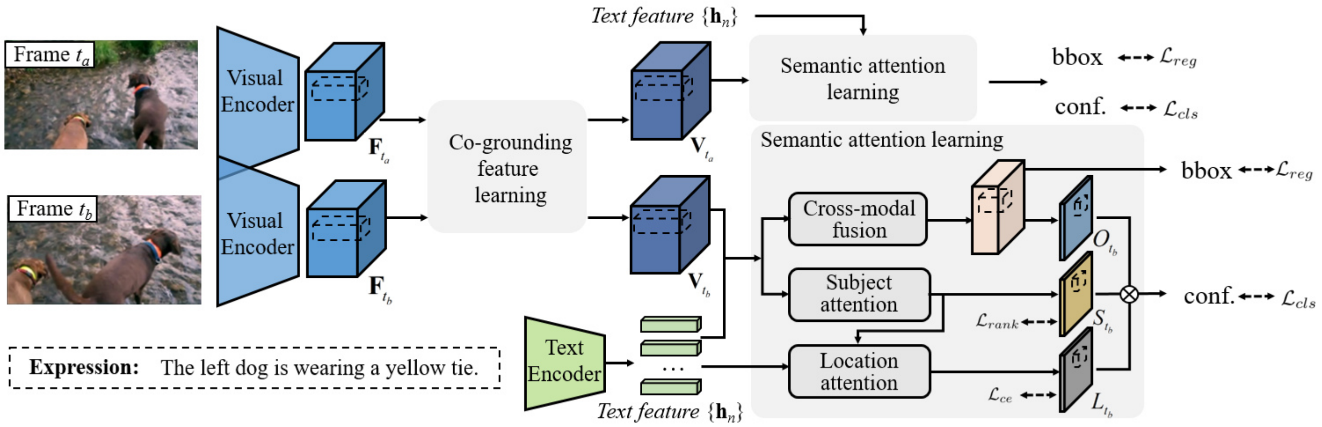


Figure 2. Our co-grounding networks with semantic attention for referring expression comprehension. The details for semantic attention learning and co-grounding feature learnng are presented in Figure 3 and Figure 4, respectively.

may lead to the lack of long-term information. To aggregate information beyond a small local range, [5] introduces a global-local aggregation network, taking full consideration of both global and local information. In addition, [16, 26] share a similar idea which leverages the merits of recurrent networks to make use of neighboring results. In our work, we design a co-grounding module to leverage correlation across frames, and further stabilize the prediction results with a post-processing strategy.

3. The Proposed Method

Given an expression Q with N words and a video $\mathbf{I}$ with T frames, our goal is to localize the object region $\{b_t\}_{t=1}^{T}$ in each frame described by Q, where b_t represents a bounding box in the t-th frame.

We build our baseline model following [36], which is based on the one-stage object detection framework, *i.e.*, YOLOv3 [28]. We conduct bounding box prediction based on cross-modal features, which are obtained by fusing visual and textual features. For each spatial position of the cross-modal features, the model outputs bounding box predictions centered at the current spatial position, with a confidence to indicate the probability of being the final grounding output. The bounding box with the highest confidence is selected as the final prediction. The basic objective function for training the model consists of the MSE (mean square error) loss $\mathcal{L}_{reg}$ to regress the bounding box towards the ground-truth, and a cross-entropy loss $\mathcal{L}_{cls}$ to select the right prediction from all the bounding boxes. We refer readers to [36, 28] for more details. Next, we elaborate the semantic attention learning and co-grounding feature learning introduced to the framework.

3.1. Semantic attention learning

To reduce ambiguity from the expression, we propose semantic attention learning to parse referring cues from the input expression. Though the input expression is usually complex, it is noticed that the words indicating subject and location play a key role to distinguish the target. Thus, we aim to decompose the expression into subject and location. With more attribute-specific textual features, we build the mapping between language and vision. Note that our semantic attention learning is different from [38] since our networks are end-to-end trainable without the reliance of offline proposal detection and external label annotations.

As shown in Figure 2, the expression Q is encoded with a text encoder consisting of bi-directional LSTM. The representation for the n-th word is the concatenation of the hidden states from both directions:

$$\mathbf{h}_n = [\overrightarrow{\mathbf{h}}_n, \overleftarrow{\mathbf{h}}_n] = \text{BiLSTM}(\mathbf{e}_n, \overrightarrow{\mathbf{h}}_{n-1}, \overleftarrow{\mathbf{h}}_{n+1}), \quad (1)$$

where $\mathbf{e}_n$ is the embedding of the n-th word. The attribute-specific textual features $\mathbf{q}_m$ $(m \in \{sub., loc.\})$ are parsed by fusing $\{\mathbf{h}_n\}_{n=1}^{N}$ with learnable weights $\mathbf{w}^m \in \mathbb{R}^N$:

$$\alpha_n^m = \frac{\exp\left(w_n^m \mathbf{h}_n\right)}{\sum_{i=1}^{N} \exp\left(w_i^m \mathbf{h}_i\right)}, \quad (2)$$

$$\mathbf{q}_m = \sum_{n=1}^{N} \alpha_n^m \mathbf{e}_n. \quad (3)$$

• **Proposal-free subject attention.** In this part, our networks learn $\mathbf{w}^{sub}$ to parse subject from Q and generate subject attention map S_t for each frame in a proposal-free manner. The subject attention map $S_t \in \mathbb{R}^{H \times W}$ reflects visual feature response to $\mathbf{q}_{sub}$ by computing cross-modal similarity as shown in Figure 3:

$$S_t(x) = \delta(\mathbf{V}_t(x), \mathbf{q}_{sub}) = \mathbf{q}_{sub}[\mathbf{V}_t(x)]^{\mathbf{T}}, \quad (4)$$

where $\mathbf{V}_t(x) \in \mathbb{R}^D$ is the feature vector at the position x from visual feature map $\mathbf{V}_t \in \mathbb{R}^{H \times W \times D}$ of the t-th frame. Rank loss is exploited to train the network, where

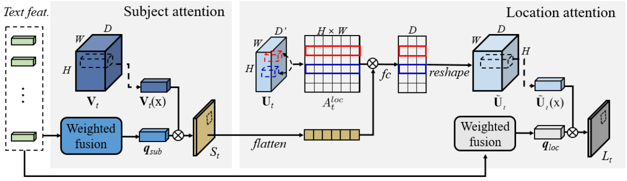


Figure 3. Details for semantic attention learning.

the score of the matched visual and textual features, *i.e.*, $(\mathbf{V}_t(x^*), \mathbf{q}_{sub})$, should be higher than unmatched ones, *i.e.*, $(\mathbf{V}_t(x^*), \mathbf{q}'_{sub})$ and $(\mathbf{V}_t(x'), \mathbf{q}_{sub})$. Therefore, the objective function is:

$$\begin{aligned}\mathcal{L}_{rank} = &\max\left(0, \Delta + \delta\left(\mathbf{V}_t(x'), \mathbf{q}_{sub}\right) - \delta\left(\mathbf{V}_t(x^*), \mathbf{q}_{sub}\right)\right) \\ &+ \max\left(0, \Delta + \delta\left(\mathbf{V}_t(x^*), \mathbf{q}'_{sub}\right) - \delta\left(\mathbf{V}_t(x^*), \mathbf{q}_{sub}\right)\right),\end{aligned} \tag{5}$$

where Δ is a margin and set to 0.5 in our experiments. During training, the visual feature vector corresponding to $\mathbf{q}_{sub}$ can be localized with the ground-truth annotation. The key issue in the training is how to select negative visual feature vectors without having object proposals. Though different schemes of hard negative sample mining have been explored, we found it is enough to tackle the problem by random sampling visual features from other training samples within the same training batch.

• **Object-aware location attention.** In this part, our networks learn $\mathbf{w}_{loc}$ to parse location from the expression, and then generate location map $L_t \in \mathbb{R}^{H\times W}$ to $\mathbf{q}_{loc}$. The key challenge in learning location attention is how to match coordinates with textual features because location is a relative concept. When we say something is '*on the left*', we have to give a reference. Thanks to the aforementioned subject attention map S which roughly identifies the subject region, we design an object-aware location representation. Specifically, we follow [36] to initially encode the coordinate feature as $\mathbf{U}_t \in \mathbb{R}^{H\times W\times D'}$. A 2D matrix $A_t^{loc} \in \mathbb{R}^{HW\times HW}$ is computed to model the relation between any two positions x and y:

$$A_t^{loc}(x, y) = [\mathbf{U}_t(x)]^{\mathbf{T}}\mathbf{U}_t(y). \tag{6}$$

With the subject attention map S_t, we inject the reference information into the location features as $A_t^{loc} \otimes \text{Flatten}(S_t)$, followed by an FC layer to shape it into $HW \times D$. Then the matrix is reshaped to $H \times W \times D$ as the final location features $\tilde{\mathbf{U}}_t$. A detailed illustration is shown in Figure 3.

Similar to Eq. 4, we obtain the location response for position x by computing cosine similarity $L_t(x) = \delta(\tilde{\mathbf{U}}_t(x), \mathbf{q}_{loc})$. We train the location attention with cross-entropy loss:

$$\mathcal{L}_{ce} = \mathbb{1}_{loc} \log \frac{\exp(L_t(x))}{\sum_y \exp(L_t(y))}, \tag{7}$$

where $\mathbb{1}_{loc} \in \{0,1\}^{H\times W}$ indicates the ground-truth location.

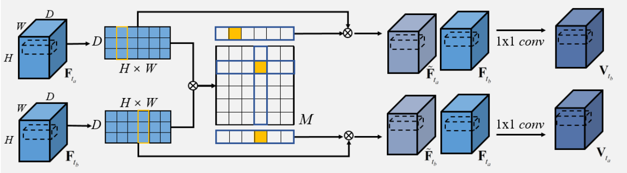


Figure 4. Details for co-grounding feature learning.

With the subject and location attention maps, the confidence map for the t-th frame is generated as $C_t = O_t \otimes S_t \otimes L_t$. Recall that we generate a bounding box prediction for each position x, $O_t(x)$ indicates how likely the predicted bounding box contains an object.

3.2. Co-grounding feature learning

For now, we have introduced how to parse referring cues from expression and build correspondence between text and visual features. However, the temporal dynamics in videos would lead to unstable visual feature representations, which may do harm to the cross-modal matching in both training and testing. To enhance the visual feature representation for a more robust learning, we propose co-grounding to integrate the temporal context by utilizing correlation across frames. As shown in Figure 2, considering two frames from the same video, we obtain the initial visual features $\mathbf{F}_{t_a} \in \mathbb{R}^{H\times W\times D}$ and $\mathbf{F}_{t_b} \in \mathbb{R}^{H\times W\times D}$ with a visual encoder. The correlation across the adjacent frames can be described by a normalized affinity matrix $M \in \mathbb{R}^{HW\times HW}$, providing the measure for similarity of spatial features:

$$M(x, y) = \frac{\exp\left([\mathbf{F}_{t_a}(x)]^{\mathbf{T}}\mathbf{F}_{t_b}(y)\right)}{\sum_y \exp\left([\mathbf{F}_{t_a}(x)]^{\mathbf{T}}\mathbf{F}_{t_b}(y)\right)}. \tag{8}$$

Then we integrate the feature vectors from $\mathbf{F}_{t_b}$ with M,

$$\tilde{\mathbf{F}}_{t_b}(x) = \sum_y M(x, y)\mathbf{F}_{t_b}(y). \tag{9}$$

The final enhance feature $\mathbf{V}_{t_a}$ is obtained by:

$$\mathbf{V}_{t_a} = \text{Conv}(\mathbf{F}_{t_a} \oplus \tilde{\mathbf{F}}_{t_b}), \tag{10}$$

where $\oplus$ indicate concatenation along the channel dimension, and $\text{Conv}(\cdot)$ denotes an 1×1 convolution operation. The visual feature $\mathbf{V}_{t_b}$ can be enhanced in the same way. Figure 4 shows the details of co-grounding feature learning.

3.3. Post processing

To further stabilize the bounding box prediction for each frame, we design a post-processing scheme based on the initial prediction results during inference. Suppose for the video, we have the initial top K bounding box predictions for each frame $\{\{b_t^i, c_t^i\}_{i=1}^K\}_{t=1}^T$, where b_t^i denotes the location for the bounding box with the i-th highest confidence c_t^i for the t-th frame. The visual feature vectors corresponding to the bounding box location for the t-th frame is $\mathcal{V}_t = \mathbf{V}_t^1(b_t^1) \oplus ... \oplus \mathbf{V}_t^K(b_t^K)$ and $\mathcal{V}_t \in \mathbb{R}^{D\times K}$. Now we consider referring the neighboring P frames as a window to stabilize the center frame t^*. For the i-th bounding box of the center frame, we stabilize its confidence score by seeking the most similar bounding box in each reference frame. The similarity is measured by the affinity matrix:

$$Z_{(t^*,t)} = \mathcal{V}_{t^*}^{\mathbf{T}}\mathcal{V}_t, t = \{t^* - \Delta t, ..., t^* + \Delta t\}, \tag{11}$$

where $\Delta t = \lfloor P/2 \rfloor$, and $Z_{(t^*,t)} \in \mathbb{R}^{K\times K}$. For each bounding box in $\{b_{t^*}^i\}_{i=1}^K$, we select the most similar bounding box from all the reference frames,

$$\mathbf{p}_t = [p_t^{(1)}, ..., p_t^{(K)}], \tag{12}$$

where

$$p_t^{(i)} = \arg\max Z_{(t^*,t)}^{(i)}, \tag{13}$$

and $Z_{(t^*,t)}^{(i)}$ denotes the i-th row of the matrix. The final scores for the initial top K bounding boxes are:

$$\tilde{\mathcal{C}}_{t^*} = \frac{1}{\mathcal{N}} \sum_{t=t^*-\Delta t}^{t^*+\Delta t} \mathbb{1}_{\mathbf{p}_t} * \mathcal{C}_t, \tag{14}$$

where $\tilde{\mathcal{C}}_{t^*} \in \mathbb{R}^K$, $\mathcal{C}_t \in \mathbb{R}^K$ and $\mathcal{C}_t = [c_t^1, ..., c_t^K]$. $\mathbb{1}_{\mathbf{p}_t} \in [0,1]^K$ can be regard as a binary mask to choose the highest score from $\mathcal{C}_t$. The bounding box with the highest score in $\mathcal{C}_{t^*}$ is treated as the final prediction result. $\mathcal{N}$ is a normalization factor.

4. Experiments

In this section, we introduce the datasets and implementation details, then report our evaluation on referring expression comprehension benchmarks. We first show the comparisons to other state-of-the-art methods, then give the ablation study and comprehensive analysis to illustrate the effectiveness of each component. Finally we discuss the failure cases and future work.

4.1. Datasets

To evaluate our model, we conduct experiments on two dynamic video datasets (*i.e.*, VID-Sentence [6], Lingual OTB99 [17]) and one static image dataset (*i.e.*, RefCOCO).

VID-Sentence (VID) [6]. This dataset consists of 7,654 trimmed videos with language descriptions, and provides the sequences of spatio-temporal bounding box annotations for each query. Following [6], the dataset is splited into 6,582/536/536 instances for training/validation/testing.

Lingual OTB99 (LiOTB) [17]. The LiOTB dataset origins from the well-known OTB100 object tracking dataset in [22]. The videos in [22] are augmented with natural language descriptions of the target object. We adopt the same protocol as [17] that 51 videos are for training and the rest are for testing.

RefCOCO [39]. The RefCOCO dataset is collected from 19,994 images in MSCOCO [19] and 142,210 natural language descriptions. RefCOCO is splited into four subsets, including train, validation, test A and test B. The images in test A are with multiple people, while those in test B are with multiple objects.

4.2. Implementation details

Training settings. Our visual encoder is based on Darknet-53 [28] pretrained on MSCOCO [19]. We adopt multi-level schemes in the grounding process, that we predict bounding boxes on three levels of feature maps, the resolution of which are 8×8, 16×16 and 32×32. The input images are resized the long edges to 256 and then padded into the size of 256×256. Following [28, 36], we adopt the data augmentation including adding randomization to the color space, horizontal flip, and random affine transformations. The network is optimized with RMSProp [30] for 100 epochs, the initial learning rate of which is set as 10^{-4} and decayed under a polynomial schedule. The batch sizes for VID, LiOTB and RefCOCO are 32, 8, 32, respectively. We set the weights for $\mathcal{L}_{rank}$ and $\mathcal{L}_{ce}$ as 100, 1, respectively. By default, the top 5 bounding boxes from the neighboring 5 frames are considered in the post-processing for the VID and LiOTB datasets during inference (*i.e.*, $K = 5$, $P = 5$). Note for the image dataset RefCOCO, we omit the co-grounding feature learning.

Evaluation metrics. We adopt different metrics to give a fair and comprehensive evaluation of our framework. Acc@0.5 is widely used to evaluate the grounding results [36, 18], where a predicted bounding box is considered correct if the IoU with the ground truth region is above 0.5. Following [37], success and precision scores are reported to evaluate the performance for videos. The success score is actually the AUC (area under curve) metric, while the precision score measures the ratio of frames where the predicted bounding box falls within a threshold of 20 pixels around the ground-truth. Besides, mIoU is also reported to show the quality of bounding boxes.

4.3. Comparison to the state-of-the-art

Table 1 shows the referring expression comprehension results on the video datasets VID and LiOTB, respectively. We first present the per-frame inference results by the state-of-the-art referring expression comprehension method [36] in the first two rows. From the 3rd to the 6th rows, we evaluate the results to see the performance when the problem is solved by per-frame tracking. Specifically, we adopt the state-of-the art tracker [15] to track the given template. With the grounding results from One-Stage LSTM [36], we conduct experiments with the first, middle, last and random frame as the tracking template, respectively. It is found that treating referring expression grounding in videos as track-

Table 1. Referring expression comprehension results on dynamic video datasets VID and LiOTB, respectively.

	VID			LiOTB		
	Accu.@0.5	Success	Precision	Accu.@0.5	Success	Precision
One-Stage BERT [36]	52.39	0.427	0.373	49.13	0.358	0.468
One-Stage LSTM [36]	54.78	0.451	0.393	49.16	0.333	0.414
First-Frame Tracking	36.97	0.334	0.250	50.93	0.391	0.482
Middle-Frame Tracking	44.00	0.384	0.307	43.08	0.356	0.421
Last-Frame Tracking	36.26	0.328	0.239	44.17	0.327	0.391
Random-Frame Tracking	40.20	0.356	0.278	25.16	0.288	0.329
WSSTG [6]	38.20	-	-	-	-	-
LSAN [17]	-	-	-	-	0.259	-
Ours	**60.25**	**0.495**	**0.462**	**52.26**	**0.392**	**0.500**

Table 2. Referring expression comprehension results (Acc.@0.5) on the static image dataset RefCOCO.

	RefCOCO			
	Visual encoder	val	testA	testB
MMI [23]	VGG16	-	64.90	54.51
Neg Bag [24]	VGG16	-	58.60	56.40
CMN [12]	VGG16	-	71.03	65.77
SLR [40]	ResNet-101	69.48	73.71	64.96
DGA [35]	ResNet-101	-	78.42	65.53
MAttN [38]	ResNet-101	76.40	80.43	69.28
CMCF [18]	DLA-34	-	81.06	71.85
One-Stage BERT [36]	Darknet53	72.05	74.81	67.59
One-Stage LSTM [36]	Darknet53	73.69	75.78	71.32
Baseline	Darknet53	73.72	76.24	71.19
S-Att.	Darknet53	77.42	**81.17**	72.77
SL-Att.(Ours)	Darknet53	**77.65**	80.75	**73.37**

Table 3. Referring expression comprehension results for ablation study on dynamic video datasets VID and LiOTB, respectively.

	VID		LiOTB	
	Acc.@0.5	mIoU	Acc.@0.5	mIoU
w/o co-grounding				
Parser	53.19	0.450	49.16	0.397
Baseline	54.41	0.448	49.11	0.405
S-Att.	58.03	0.488	49.66	0.411
SL-Att.	**59.22**	**0.490**	**50.51**	**0.418**
w/ co-grounding				
CG-Baseline	55.88	0.477	50.56	0.405
CG-S-Att.	58.74	0.497	51.67	0.412
CG-SL-Att.	59.48	0.494	50.92	0.418
CG-SL-Att. + pp.	**60.25**	**0.498**	**52.26**	**0.418**

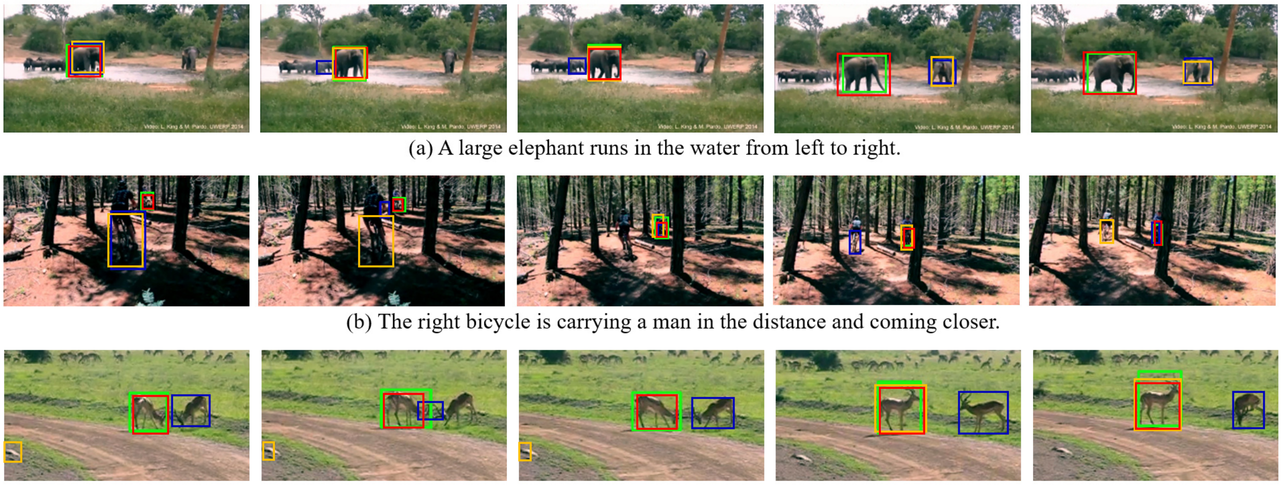
ing leads to poor results compared to per-frame grounding for almost all the cases. There are mainly two reasons to explain the poor results. On the one hand, we can not guarantee the correct template is selected during tracking. On the other hand, even given the correct template, the tracker may fail due to its limitation. The analysis also applies to the unsatisfactory results of LSAN [17]. Besides, we present the results of WSSTG [6] which relies on spatio-temporal proposal detection. Our results are shown in the last row. It can be seen our framework outperforms the compared methods by a large margin.

Table 2 shows the comparisons on the RefCOCO dataset for referring expression comprehension. Our overall results on RefCOCO are shown in the last row. To give a fair comparison, we present the backbone structure of each visual encoder. Though MAttN [38] also explores the idea of modular attention for both language and vision, it requires an offline object detector and external attribute labels. However, our model can learn the semantic attention in a proposal-free manner, which also makes our results outstanding compared to other one-stage models [18, 36].

4.4. Ablation study

To show the effectiveness of each component in our model, ablation study is conducted on VID, LiOTB and RefCOCO datasets, respectively. We explore different settings to give a comprehensive analysis. The results are presented in Table 3 and Table 2. Note that we regard the model of one-stage LSTM in [36] as our **Baseline**.

- **Semantic attention learning.** We first explore the contribution of semantic attention learning without taking co-grounding feature learning into account. For all the datasets, subject attention (**S-Att.**) brings significant improvement compared to baseline results. It is largely because the subject attention reduces ambiguity in grounding process when there are multiple entities in the expression and images. Location attention (**SL-Att.**) further improves the grounding accuracy. Overall, for the VID and LiOTB datasets, the gains from semantic attention learning over baselines are 4.81% and 1.40%, respectively. For the RefCOCO dataset, our framework outperforms the baseline by 3.93%, 4.93%, and 2.18% under different split settings, respectively. Moreover, we compare our automatic semantic attention learning with manually semantic parser [13] in Table 3 (see **Parser**). As analyzed in [38], parsing errors exist for the external parser which is not tuned for referring



(a) A large elephant runs in the water from left to right.

(b) The right bicycle is carrying a man in the distance and coming closer.

(c) The left antelope stands on the road playing with the right antelope on the grass.

Figure 5. Visualization results of video grounding on the VID dataset. We show the ground-truths, baseline results, results of SL-Att. and results of CG-SL-Att. in the green, blue, orange and red bounding boxes, respectively. The language queries are shown in the sub-captions.

Table 4. Ablation study for post-processing on the VID dataset in terms of Acc.@0.5 and mIoU.

	P=1	P=3	P=5	P=7	P=9
Acc.@0.5	59.48	59.50	60.25	60.16	59.63
mIoU	0.494	0.495	0.500	0.498	0.495

expressions. Therefore, we did not observe improvement compared to baseline results.

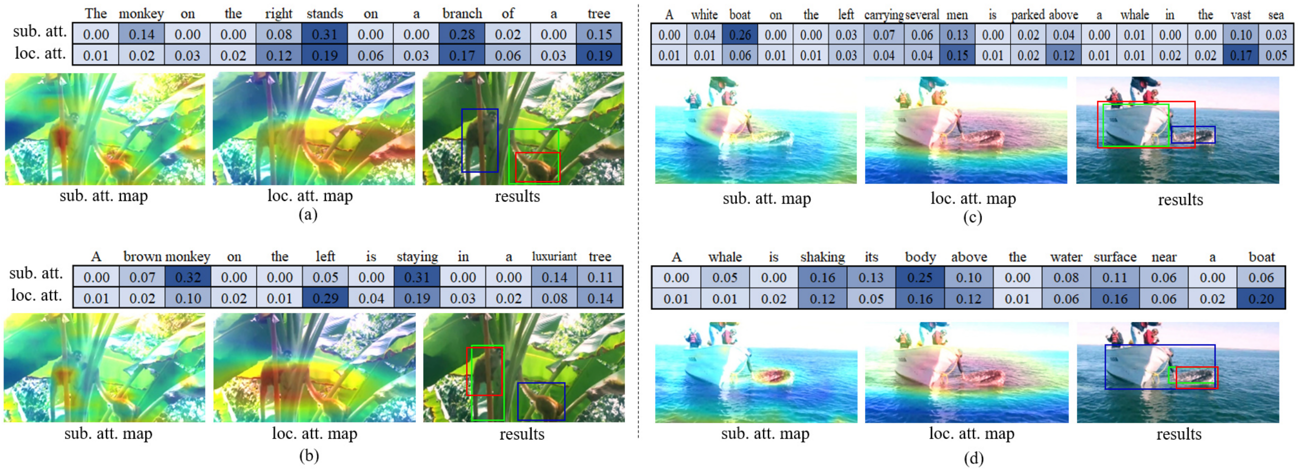
• **Co-grounding feature learning.** We further analyze the contributions of our co-grounding feature learning through Table 3. Conducting co-grounding feature learning with the baseline structure (**CG-Baseline**) brings 1.47% and 1.45% gains in terms of Acc.@0.5 over **Baseline** on the VID and LiOTB datasets, respectively. With semantic attention learning, the co-grounding feature learning helps further improve the performance on both datasets for each setting (see **CG-S-Att.** *vs.* **S-Att.**, **CG-SL-Att.** *vs.* **SL-Att.**).

• **Post processing.** Finally we illustrate the effectiveness of the post processing scheme. The results in the last row of Table 3 (**CG-SL-Att. + pp.**) show that our post-processing scheme is able to further improve the grounding results in terms of Acc.@0.5 and mIoU. In Table 4, we further explore the influence of different numbers of reference frames. While the post-processing scheme consistently improve the results compared to those without post-processing (*i.e.*, $P = 1$), we set P as 5 in all our experiments for the tradeoff between efficiency and accuracy.

4.5. Qualitative results

• **Overall grounding results.** We choose several videos from the VID dataset and visualize the grounding results in Figure 5. We compare the results of the baseline, SL-Att. and CG-SL-Att. with the blue, orange and red bounding boxes, respectively. The ground-truths are in green. From the visualization, it is observed that compared to baseline results, SL-Att. provides more accurate prediction in most cases, due to the explicitly parsed referring cues both for language and vision. However, bounding box drifting is a problem when we conduct per-frame inference without taking the temporal context into account (see the drifting orange bounding boxes). In Figure 5(b), the target bicycle is small and obscure in the first several frames, making the visual features vulnerable for SL-Att. model. In Figure 5(c), the left stone mislead SL-Att. to ground it as '*antelope*'. With co-grounding feature learning, the visual features are enhanced by integrating temporal context and become more robust. Therefore, we obtain consistent results across frames (see the red boxes). Please refer to the supplementary for more grounding results.

• **Visualization on attention.** We show the learned attention patterns for language and vision in Figure 6. For each side, different queries are given for the same frame. It is found that the semantic attention for the given expression successfully parses the words for subject entities and location from the language. And the semantic attention for the visual feature maps generate corresponding response for different attributes. In Figure 6(a) and (b), we show examples that location attention helps the model handle ambiguities in the frames and distinguish the correct bounding box. With the parsed subject '*monkey stands branch*' and '*monkey staying*', the model pays more attention on both monkeys in the frame. However, with the guidance of the parsed words describing location '*right stands branch tree*' and '*left staying*', the model shows different response on the location attention maps, providing essential cues to distinguish the correct bounding box. For the examples on the right side, there is not dominant location information in the input expressions, resulting in similar location maps in Figure 6(c)



	The	monkey	on	the	right	stands	on	a	branch	of	a	tree
sub. att.	0.00	0.14	0.00	0.00	0.08	0.31	0.00	0.00	0.28	0.02	0.00	0.15
loc. att.	0.01	0.02	0.03	0.02	0.12	0.19	0.06	0.03	0.17	0.06	0.03	0.19

	A	white	boat	on	the	left	carrying	several	men	is	parked	above	a	whale	in	the	vast	sea
	0.00	0.04	0.26	0.00	0.00	0.03	0.07	0.06	0.13	0.00	0.02	0.04	0.00	0.01	0.00	0.00	0.10	0.03
	0.01	0.01	0.06	0.01	0.01	0.03	0.04	0.04	0.15	0.01	0.02	0.12	0.01	0.01	0.02	0.02	0.17	0.05

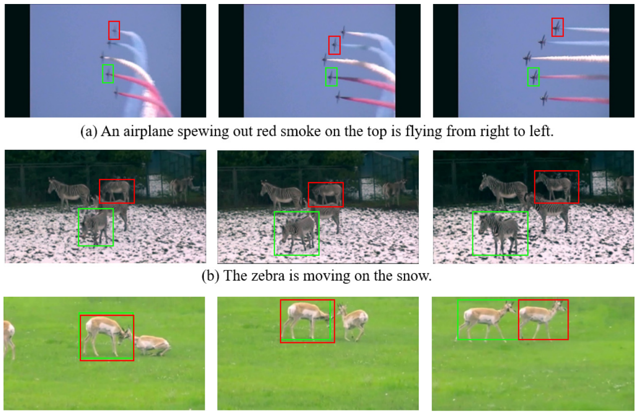
	A	brown	monkey	on	the	left	is	staying	in	a	luxuriant	tree
sub. att.	0.00	0.07	0.32	0.00	0.00	0.05	0.00	0.31	0.00	0.00	0.14	0.11
loc. att.	0.01	0.02	0.10	0.02	0.01	0.29	0.04	0.19	0.03	0.02	0.08	0.14

	A	whale	is	shaking	its	body	above	the	water	surface	near	a	boat
	0.00	0.05	0.00	0.16	0.13	0.25	0.10	0.00	0.08	0.11	0.06	0.00	0.06
	0.01	0.01	0.02	0.12	0.05	0.16	0.12	0.01	0.06	0.16	0.06	0.02	0.20

Figure 6. Visualization on attention patterns of our framework. The subject and location attention patterns for language are shown above the images. We mark the attention values for each word. The attention patterns for visual features are shown as the heat maps overlayed on the original images. The more reddish, the larger attention. Besides, we show the grounding results of baseline and SL-Att. in the blue and red bounding boxes, respectively. Ground-truths are shown in green. (Best viewed in color.)

and (d). The multiple entities such as '*whale*', '*boat*' appearing in the sentences make the baseline model confusing that which subject is correct to ground (see the blue bounding boxes in Figure 6(c)(d)). In our model, the subject attention for language effectively excludes the effect of other entities in the expression, making it clear to ground '*boat*' in Figure 6(c) and '*body*' in Figure 6(d). It further leads to corresponding high response on the subject attention maps for visual features and then satisfactory grounding predictions (see the red boxes). More visualizations can be found in our supplementary.

- **Failure case analysis and future work.** We show some typical failure cases in Figure 7, to illustrate the limitations our model, and the challenges for the topic of video grounding. (1) Multi-order reasoning is challenging for one-stage referring expression comprehension because it always involves multiple entities and relation concepts. As shown in Figure. 7(a), it is difficult for our model to locate the airplanes with red smoke and then select the top one. (2) Motion information is not explicitly explored and utilized as grounding cues. As shown in Figure 7(b), we can not determine which '*zebra*' is '*moving*' only by observing static frames. (3) The language query may not apply to all the frames. In Figure 7(c), the ground-truth is not in the '*middle*' in some frames. How to further tackle the ambiguity caused by dynamic scenes and expression is still worth exploring. We leave how to solve these failure cases as interesting future works.



(a) An airplane spewing out red smoke on the top is flying from right to left.

(b) The zebra is moving on the snow.

(c) The antelope in the middle nods at the antelope on the right.

Figure 7. Failure cases. Ground-truths are in green and our results are in red.

5. Conclusion

In this paper, we tackle the problem of referring expression comprehension in videos. We propose to solve the problem from a new perspective, co-grounding, with an elegant one-stage framework. To boost single frame results, our model learns semantic attention to decompose grounding cues into different attributes, which further contribute to the reasoning of the target described by the input expression. To boost grounding prediction consistency, we propose co-grounding feature learning by integrating neighboring features across frames to enhance visual feature representations. A post-processing scheme is conducted during inference to further stabilize the predictions. Our model is applicable to visual grounding both for videos and images. Experiments on video and image grounding benchmarks illustrate the effectiveness of our model.

Acknowledgements. This work was supported by the Beijing Natural Science Foundation under Contract No.4192025, the National Natural Science Foundation of China under Contract No.61772043. This is a research achievement of Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology). This work was done in part when Sijie Song was a visiting scholar at Columbia University, supported by China Scholarship Council.

References

[1] Hassan Akbari, Svebor Karaman, Surabhi Bhargava, Brian Chen, Carl Vondrick, and Shih-Fu Chang. Multi-level multimodal common semantic space for image-phrase grounding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12476–12486, 2019. 1

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6077–6086, 2018. 1

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Int. Conf. Comput. Vis.*, pages 2425–2433, 2015. 1

[4] Dongdong Chen, Jing Liao, Lu Yuan, Nenghai Yu, and Gang Hua. Coherent online video style transfer. In *Int. Conf. Comput. Vis.*, pages 1105–1114, 2017. 2

[5] Yihong Chen, Yue Cao, Han Hu, and Liwei Wang. Memory enhanced global-local aggregation for video object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10337–10346, 2020. 2, 3

[6] Zhenfang Chen, Lin Ma, Wenhan Luo, and Kwan-Yee Kenneth Wong. Weakly-supervised spatio-temporally grounding natural sentence in video. In *ACL*, pages 1884–1894, 2019. 1, 2, 5, 6

[7] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Int. Conf. Comput. Vis.*, pages 2758–2766, 2015. 2

[8] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Detect to track and track to detect. In *Int. Conf. Comput. Vis.*, pages 3038–3046, 2017. 2

[9] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *Eur. Conf. Comput. Vis.*, volume 5, 2020. 1

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Int. Conf. Comput. Vis.*, pages 2961–2969, 2017. 2

[11] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3588–3597, 2018. 2

[12] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. Modeling relationships in referential expressions with compositional modular networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1115–1124, 2017. 1, 6

[13] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, pages 787–798, 2014. 6

[14] Eun-Sol Kim, Woo Young Kang, Kyoung-Woon On, Yu-Jung Heo, and Byoung-Tak Zhang. Hypergraph attention networks for multimodal learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 14581–14590, 2020. 2

[15] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4282–4291, 2019. 5

[16] Dong Li, Zhaofan Qiu, Qi Dai, Ting Yao, and Tao Mei. Recurrent tubelet proposal and recognition networks for action detection. In *Eur. Conf. Comput. Vis.*, pages 303–318, 2018. 2, 3

[17] Zhenyang Li, Ran Tao, Efstratios Gavves, Cees GM Snoek, and Arnold WM Smeulders. Tracking by natural language specification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6495–6503, 2017. 1, 2, 5, 6

[18] Yue Liao, Si Liu, Guanbin Li, Fei Wang, Yanjie Chen, Chen Qian, and Bo Li. A real-time cross-modality correlation filtering method for referring expression comprehension. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10880–10889, 2020. 1, 2, 5, 6

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Eur. Conf. Comput. Vis.*, pages 740–755, 2014. 5

[20] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Adv. Neural Inform. Process. Syst.*, pages 289–297, 2016. 2

[21] Xiankai Lu, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, and Fatih Porikli. See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3623–3632, 2019. 2

[22] Yang Lu, Tianfu Wu, and Song Chun Zhu. Online object tracking, learning and parsing with and-or graphs. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3462–3469, 2014. 5

[23] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11–20, 2016. 1, 6

[24] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *Eur. Conf. Comput. Vis.*, pages 792–807, 2016. 1, 6

[25] Duy-Kien Nguyen and Takayuki Okatani. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6087–6096, 2018. 2

[26] David Nilsson and Cristian Sminchisescu. Semantic video segmentation by gated recurrent flow propagation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6819–6828, 2018. 2, 3

[27] Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei. Video captioning with transferred semantic attributes. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6504–6512, 2017. 1

[28] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 2, 3, 5

[29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Adv. Neural Inform. Process. Syst.*, pages 91–99, 2015. 2

[30] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012. 5

[31] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval-an empirical odyssey. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6439–6448, 2019. 1

[32] Peng Wang, Qi Wu, Jiewei Cao, Chunhua Shen, Lianli Gao, and Anton van den Hengel. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1960–1968, 2019. 1, 2

[33] Qi Wu, Peng Wang, Chunhua Shen, Ian Reid, and Anton Van Den Hengel. Are you talking to me? reasoned visual dialog generation through adversarial learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6106–6115, 2018. 2

[34] Sibei Yang, Guanbin Li, and Yizhou Yu. Cross-modal relationship inference for grounding referring expressions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4145–4154, 2019. 1, 2

[35] Sibei Yang, Guanbin Li, and Yizhou Yu. Dynamic graph attention for referring expression comprehension. In *Int. Conf. Comput. Vis.*, pages 4644–4653, 2019. 6

[36] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *Int. Conf. Comput. Vis.*, pages 4683–4693, 2019. 1, 2, 3, 4, 5, 6

[37] Zhengyuan Yang, Tushar Kumar, Tianlang Chen, and Jiebo Luo. Grounding-tracking-integration. *arXiv preprint arXiv:1912.06316*, 2019. 5

[38] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1307–1315, 2018. 1, 2, 3, 6

[39] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Eur. Conf. Comput. Vis.*, pages 69–85, 2016. 5

[40] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. A joint speaker-listener-reinforcer model for referring expressions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7282–7290, 2017. 6

[41] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6281–6290, 2019. 2

[42] Luowei Zhou, Nathan Louis, and Jason J Corso. Weakly-supervised video object grounding from text by loss weighting and object interaction. *arXiv preprint arXiv:1805.02834*, 2018. 1

[43] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. Flow-guided feature aggregation for video object detection. In *Int. Conf. Comput. Vis.*, pages 408–417, 2017. 2