Contrastive Learning with Expectation-Maximization for Weakly Supervised Phrase Grounding

Keqin Chen[1], **Richong Zhang**[1,2*], **Samuel Mensah**[3], **Yongyi Mao**[4]
[1]SKLSDE, Beihang University, Beijing, China
[2]Zhongguancun Laboratory, Beijing, China
[3]Department of Computer Science, University of Sheffield, UK
[4] School of Electrical Engineering and Computer Science, University of Ottawa, Canada
{chenkq,zhangrc}@act.buaa.edu.cn
s.mensah@sheffield.ac.uk, ymao@uottawa.ca

Abstract

Weakly supervised phrase grounding aims to learn an alignment between phrases in a caption and objects in a corresponding image using only caption-image annotations, i.e., without phrase-object annotations. Previous methods typically use a caption-image contrastive loss to indirectly supervise the alignment between phrases and objects, which hinders the maximum use of the intrinsic structure of the multimodal data and leads to unsatisfactory performance. In this work, we directly use the phrase-object contrastive loss in the condition that no positive annotation is available in the first place. Specifically, we propose a novel contrastive learning framework based on the expectation-maximization algorithm that adaptively refines the target prediction. Experiments on two widely used benchmarks, Flickr30K Entities and RefCOCO+, demonstrate the effectiveness of our framework. We obtain 63.05% top-1 accuracy on Flickr30K Entities and 59.51%/43.46% on RefCOCO+ TestA/TestB, outperforming the previous methods by a large margin, even surpassing a previous SoTA that uses a pre-trained vision-language model. Furthermore, we deliver a theoretical analysis of the effectiveness of our method from the perspective of the maximum likelihood estimate with latent variables.

1 Introduction

Phrase grounding aims to localize corresponding objects in an image given a phrase in the image's caption. It is one of the most fundamental research areas in multimodal learning (Ramachandram and Taylor, 2017). This area has strong applications in other complex visual language tasks, such as visual question answering (Khan et al., 2021), cross-modal retrieval (Chen et al., 2017), etc.

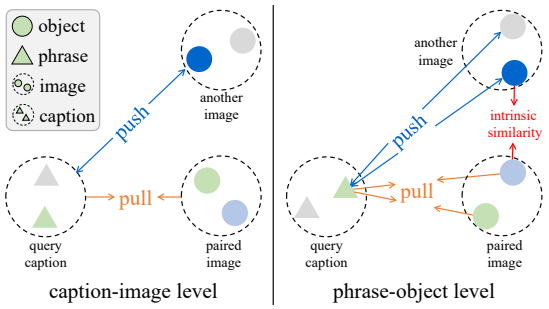

Figure 1: Previous methods neglect the intrinsic similarity which can be seen as a soft co-occurrence. Utilizing the intrinsic similarity between negative examples (dark blue circle) and false-positive examples (light blue circle) can push the false-positive examples away from the query phrase (light green triangle).

Current methods (Huang et al., 2021; Kamath et al., 2021) have achieved great success but rely heavily on bounding box annotations, which are expensive to acquire. Thus, weakly supervised phrase grounding has recently received increased attention (Liu et al., 2021; Wang et al., 2020; Dou and Peng, 2021) due to the low cost of obtaining image-text pair annotations. It aims to ground phrases using only caption-image annotations, i.e., without phrase-object annotations.

The most critical question in weakly supervised phrase grounding is how to provide sufficient phrase-object alignment supervision. Previous methods typically provide oblique phrase-object alignment signal by heuristically aggregating several fine-grained phrase-object similarities to be a coarse-grained caption-image similarity and performing contrastive learning at the caption-image level (Wang et al., 2020; Zhang et al., 2020; Gupta et al., 2020; Wang et al., 2021). While existing methods have gained some progress in the task, they are limited in several ways: this approach only provides a coarse alignment signal and has no

*Corresponding author

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 8549–8559
December 7-11, 2022

theoretical support. As shown in Figure 1 (left), A close caption-image pair does not guarantee a near distance for the corresponding phrase-object pairs and occasionally fails to disambiguate the co-occurrence objects. In contrast, we propose to conduct phrase-object contrastive learning directly.

However, contrastive learning always needs positive examples annotation. Here, they are corresponding phrase-object pairs, which we do not have under the weakly supervised setting. It seems that we are stuck in a chicken or egg dilemma: a good model needs high-quality labels to train; good labels only come from effective models. The key to get rid of the dilemma is the intrinsic similarity in data. That is, image regions sharing the same concept are close in feature space. For example, "alaskan malamute" and "husky" have the same concept (i.e., dog) and therefore their image features are expected to be close in the feature space. This intrinsic similarity is key to disambiguate the co-occurrence objects and reveal the ground truth. As shown in Figure 1 (right), for a single query phrase, when a negative example (objects in another image) is pushed away, the false-positive example (objects in the paired image but not corresponding to the query phrase) semantically similar to the negative example is pushed away as a side effect. The trick is that although we don't know which one is likely to be correct, we do have a fair estimate about which one is more likely to be wrong.

We can leverages and amplifies this effect, by gradually removing the false positives from the positives via interleaving contrastive learning and pseudo labels update. At first, we have no preference for each object and set the pseudo labels on objects as uniform distribution. After one iteration of contrastive learning, the model gains more confidence on the ground truth. Now, we retrain the model by updating the pseudo labels using the current model predictions; then since the object assignment is more correct, the retrained model is expected to predict better. This idea can be formulated as an expectation-maximization algorithm (Dempster et al., 1977) and have solid theoretical support from the MLE perspective. To the best of our knowledge, no previous work has highlighted the importance of the intrinsic similarity and developed a method accordingly.

To realize the above idea, we propose a novel contrastive learning framework based on the EM algorithm that adaptively refines the target prediction. Specifically, we treat all the objects in the paired image as positive examples for contrastive learning at the phrase-object level. And we introduce pseudo labels to describe how much each object is likely to be the correct answer, resulting in different importance for different objects in the contrastive loss. Finally, we update the pseudo labels from a moving average of model predictions. Our model can progressively refine the target prediction by iteratively minimizing the contrastive loss and updating pseudo labels.

We conduct experiments on two benchmarks, Flickr30K Entities and RefCOCO+, and achieve a new SoTA. Also, we perform a detailed ablation study and case study to show the effectiveness of each component.

In summary, the contributions of this work are threefold:

- We identify the significance of the intrinsic similarity in solving weakly supervised phrase grounding and propose a novel contrastive learning framework accordingly.
- We achieve a new SoTA on Flickr30K Entities and RefCOCO+, outperforming the previous methods by a large margin.
- We conduct extensive experiments to verify the effectiveness qualitatively and quantitatively and deliver a theoretical analysis of the effectiveness from the perspective of the maximum likelihood estimate.

2 Related Work

Weakly Supervised Phrase Grounding Weakly supervised phrase grounding has received considerable attention, as the fully supervised setting requires the labelling of phrases in image captions. Previous mainstream methods can be divided into two main categories: reconstruction-based methods (Liu et al., 2021; Dou and Peng, 2021; Rohrbach et al., 2016) and contrastive-based methods (Wang et al., 2020, 2021; Zhang et al., 2020; Gupta et al., 2020; Datta et al., 2019). Our work falls into the line of contrastive-based methods.

Previous contrastive-based methods typically conduct contrastive learning at the caption-image level by heuristically aggregating several phrase-object similarities to be caption-image similarities. Align2Ground (Datta et al., 2019) aggregates several features of objects to get a caption-conditioned

image representation, and matches it with the corresponding caption. InfoGround (Gupta et al., 2020) defines a compatibility function to measure the compatibility between images and BERT-contextualized word representation. InfoGround uses BERT to generate hard negative examples. CCL (Zhang et al., 2020) defines an aggregation function to compute the alignment score between a phrase and a set of objects, and generates counterfactual examples by the gradient. MAF (Wang et al., 2020) uses the mean of phrase-wise maximum similarity as the caption-image similarity and performs contrastive learning at the caption-image level. These works are distinctively different in their definition of the aggregation function and the approach in which they generate positive and negative examples. None of these works use contrastive learning at the phrase-object level nor do they highlight the importance of intrinsic structure as we do.

Contrastive Learning Contrastive learning has recently attracted much attention, as it has contributed to the success in unsupervised representation learning. There exist a line of work that explore contrastive learning at the instance level in computer vision, natual language processing and multimodal learning. For example, MoCo (He et al., 2020), SimCLR (Chen et al., 2020), SimCSE (Gao et al., 2021), CLIP (Radford et al., 2021), and ALBEF (Li et al., 2021a). Another line explore the learning problem in a weakly supervised setting. PiCO (Wang et al., 2022) is one of such works, which uses prototypes to address label disambiguation in partial label learning. Our work is closely related to PiCO. However, our work differs from PiCO in twofold: Firstly, our model is oriented toward a multimodal setting while PiCO only works with a unimodal setting (i.e., image). Secondly, PiCO only suits classification problems where the total number of labels is fixed while our model is unconstrained to the number of labels. Specifically, in phrase grounding, the total number of possible objects is unlimited. Hence, PiCO cannot be directly applied to this task.

3 Problem Formulation

Given a caption-image pair (S, I), a grounding model is expected to find the object k_j among m objects in image I which refers to the given phrase q_i in caption S:

$$\arg\max_{1\leq j\leq m} \log p(k_j|q_i; I, S) \tag{1}$$

It is difficult to solve this objective due to the lack of phrase-region annotations. Therefore, we treat this annotation as a latent variable z_i. That is, the phrase q_i refers to the region k_j when $z_i = j$. We then solve the following maximum likelihood estimate problem instead:

$$\begin{aligned} &\max \log p(q_i|I) \\ &= \max \log \sum_{j=1}^{m} p(q_i, z_i = j|I) \end{aligned} \tag{2}$$

Note that Zhang et al. (2018) solve $\log p(k_j|S)$. On the contrary, we solve $\max \log p(q_i|I)$. This is based on the observation that all phrases are conditioned on images but not all regions have a corresponding phrase. The maximum likelihood estimate of $p(k_j|S)$ decreases the alignment performance when k_j has no corresponding phrase in sentence S. A model that solves (2) provides a competitive estimate for latent variable z_i.

For clarity, we give several definitions here. With respect to a given query phase, an image region falls into one of the three categories: true positive, false positive, and (true) negative. All regions in an image not paired with the phase are regarded as negatives. All regions in the image paired with the phrases are positives, in which only the one that semantically corresponds to the phrase is the true positive and the rest are false positives.

4 Methodology

In this section, we introduce the proposed contrastive learning framework. We follow a 2-stage paradigm. In the first stage, we apply a pre-trained object detector to extract object features. In the second stage, we sort the objects by their similarities with phrases. As shown in Figure 2, our framework consists of three main components:

Image Encoder f_k: It extracts bounding boxes and features of objects in an image. We let $k_1, k_2, \ldots, k_m = f_k(I)$ denote object features, where I is an image and m is the number of objects in the image.

Text Encoder f_q: It takes as input a caption S and outputs n phrase features, denoted by $q_1, q_2, \ldots, q_n = f_q(S)$.

Contrastive Loss Module: It adaptively refines the alignment score between objects and phrases.

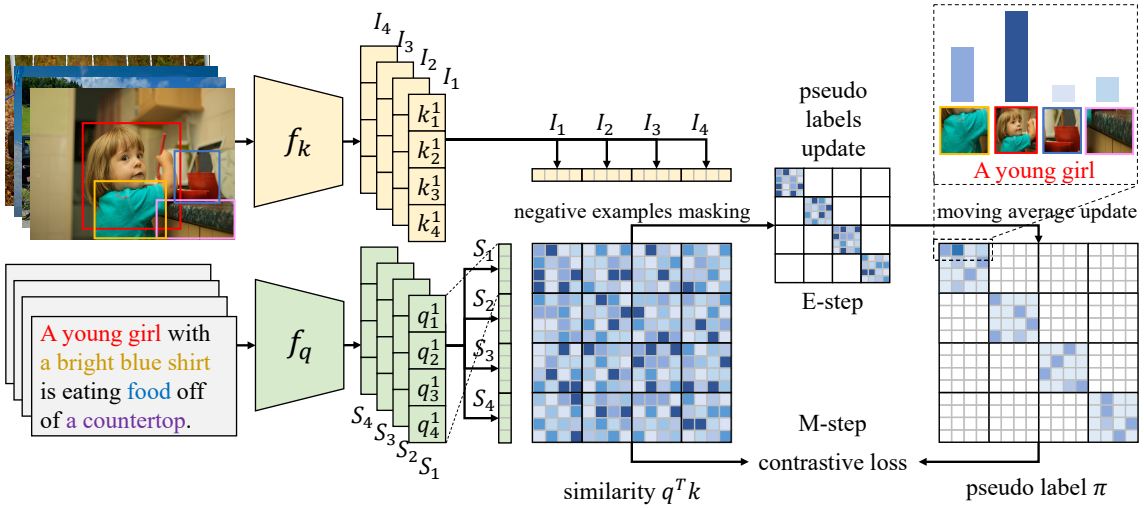

Figure 2: Illustration of our approach. Input is a batch of caption-image pairs. f_q and f_k extract features for phrases and objects. For each phrase, objects in the paired image are positive examples, and the rest are negative. In each iteration, M-step and E-step are applied successively. M-step minimizes a contrastive loss based on the pseudo label π which shows how positive one example is. E-step updates π using a moving average of model predictions.

The key idea is to apply an E-step and M-step iteratively for this purpose.

In the following subsections, we describe each component in detail.

4.1 Image Encoder

Image Encoder is responsible for extracting object features. We adopt a similar encoder as MAF (Wang et al., 2020). Specifically, the object feature is calculated as follows:

$$\begin{aligned} \bar{\boldsymbol{f}}_{\boldsymbol{i}} &= \text{dropout}(\boldsymbol{f}_{\boldsymbol{i}}) \\ \bar{\boldsymbol{k}}_{\boldsymbol{i}} &= \boldsymbol{l}_{\boldsymbol{i}} + W_f \bar{\boldsymbol{f}}_{\boldsymbol{i}} \\ \boldsymbol{k}_{\boldsymbol{i}} &= \text{dropout}(\bar{\boldsymbol{k}}_{\boldsymbol{i}}) \end{aligned} \tag{3}$$

Here, $\boldsymbol{f}_{\boldsymbol{i}} \in \mathbb{R}^{d'}$ is the feature of the i-th object calculated by a pre-trained object detector. Also, the detector predicts a text label indicating which class this object most likely belongs to. The embedding of the label is denoted as $\boldsymbol{l}_{\boldsymbol{i}} \in \mathbb{R}^d$. $W_f \in \mathbb{R}^{d*d'}$ is a projection matrix. W_f is zero-initialized to use the text-text similarity at the initial stage, providing a good initialization. During training, the model evolves gradually from a fully text-only model to a multi-modal model.

4.2 Text Encoder

The text encoder encodes each phrase in a caption. Unlike MAF (Wang et al., 2020), we find that integrating visual information into phrase features contributes little to alignment. Therefore we simply construct the phrase representation by applying a sum-pool on the phrase's GloVe embeddings (Pennington et al., 2014):

$$\begin{aligned} \bar{\boldsymbol{q}}_{\boldsymbol{i}} &= \frac{1}{\sigma} \sum_{j=1}^{n_i} \boldsymbol{h}_{\boldsymbol{ij}} \\ \boldsymbol{q}_{\boldsymbol{i}} &= \text{dropout}(W_q \bar{\boldsymbol{q}}_{\boldsymbol{i}}) \end{aligned} \tag{4}$$

Here, n_i is the number of words in the i-th phrase. $\boldsymbol{h}_{\boldsymbol{ij}} \in \mathbb{R}^d$ is the GloVe embedding for the j-th word in the i-th phrase. σ is a hyperparameter to scale the phrase representation. $W_q \in R^{d*d}$ is a projection matrix, initialized as an identity matrix to stabilize training.

4.3 Contrastive Loss Module

The contrastive loss aims to pull together query q and positive examples k^+ and push away q and negative examples k^-. InfoNCE (van den Oord et al., 2018; He et al., 2020) adopts the inner product $q^T k$ as the similarity metric between q and k:

$$\mathcal{L}_q = -\log \frac{\exp(\boldsymbol{q}^T \boldsymbol{k}^+ / \tau)}{\sum_{\boldsymbol{k}^- \in K} \exp(\boldsymbol{q}^T \boldsymbol{k}^- / \tau)} \tag{5}$$

Here, K is the set of all the negative objects, and τ is a temperature hyperparameter.

By applying InfoNCE loss at the phrase-object level, negative examples for a given phrase are straightforward to obtain: collecting objects from other images in the same training batch. However, obtaining a positive example for a phrase is challenging due to the lack of annotations. Moreover, phrases and objects live in two modalities. This

means augmentations of a phrase can not serve as positive examples.

We handle the generation of positive examples by introducing pseudo labels. Specifically, inspired by the recent progress in prototypical contrastive learning (Wang et al., 2022; Li et al., 2021b), we treat all objects in the paired image as positive examples and assign a pseudo label on each object to describe how much it is likely to be positive. So the InfoNCE loss can be rewritten as:

$$\mathcal{L}_q = -\sum_{\boldsymbol{k}^+ \in K^+} \pi_{qk^+} \log \frac{\exp(\boldsymbol{q}^T \boldsymbol{k}^+/\tau)}{\sum_{\boldsymbol{k}^- \in K} \exp(\boldsymbol{q}^T \boldsymbol{k}^-/\tau)} \tag{6}$$

Here, K^+ is the set of objects in the paired image, K is the set of all objects in the same batch, τ is a temperature hyperparameter, and π_{qk} is the pseudo label showing the confidence that the phrase $\boldsymbol{q}$ is aligned to object $\boldsymbol{k}$. Moreover, $\boldsymbol{\pi}$ is fixed during the optimization of the loss $\mathcal{L}_q$ and satisfies the following constraints:

$$\begin{aligned} \sum_{k \in K^+} \pi_{qk} &= 1 \\ \pi_{qk} &= 0 \quad \forall k \notin K^+ \end{aligned} \tag{7}$$

Initially, pseudo label $\boldsymbol{\pi}$ is assigned a uniform distribution among all positive examples due to the lack of prior knowledge about which object is more likely to be aligned.

$$\pi_{qk}^{init} = \frac{1}{|K^+|} \quad \forall k \in K^+ \tag{8}$$

During training, we interleave minimizing the loss $\mathcal{L}_q$ with updating the pseudo label π. Instead of computing π every few steps, we adopt the moving average updating strategy introduced by PiCO (Wang et al., 2022) to smoothen the training procedure. In every train step, we first minimize the contrastive loss $\mathcal{L}_q$, and then update π using the moving average strategy:

$$\boldsymbol{\pi_q}^{new} = \lambda \boldsymbol{\pi_q}^{old} + (1-\lambda)\boldsymbol{s} \tag{9}$$

where $\lambda \in (0,1)$ is a hyperparameter.

$$s_i = \begin{cases} 1 & i = \arg\max_{1 \le j \le m} \boldsymbol{q}^T \boldsymbol{k}_j^+, \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

Here, m is the total number of objects in the paired image. $\boldsymbol{s}$ can be treated as a hard version of model predictions, i.e., assigns one to the most confident object and zero to the others.

4.4 An MLE Perspective

In this section, we deliver a derivation from the perspective of the Maximum Likelihood Estimate (MLE) to illustrate the relationship between the Contrastive Loss Module and the EM algorithm, which sheds light on why it works theoretically.

Recall that the MLE problem for phrase q:

$$\begin{aligned} &\max_\theta \log P(q \mid \theta, I) \\ =&\max_\theta \log \sum_{i=1}^{m} P(q \mid z=i, \theta, I) P(z=i \mid \theta, I) \end{aligned} \tag{11}$$

z is a latent variable, indicating phrase q describes the i-th object in image I when $z=i$. Omitting some steps, we directly give the Q-function.

$$\begin{aligned} &Q(\theta, \theta^{old}) \\ =&\sum_{i=1}^{m} P(z=i \mid q, \theta^{old}, I) \log P(q, z=i \mid \theta, I) \end{aligned} \tag{12}$$

E-step aims to guess the probability of the latent variable z using θ^{old}. i.e., $\pi_i = P(z=i \mid q, \theta^{old}, I)$. In our work, we apply a softmax on the inner product of phrase q and object k_i to obtain π. i.e., $\pi = \text{softmax}_i(q^T k_i)$. Practically, we use a moving average strategy to update π to smoothen the training procedure.

M-step aims to maximize the Q-function. For convenience of derivation, we give two mild assumptions:

Assumption 1: the prior distribution of z, $P(z=i \mid \theta, I)$, is a uniform distribution. It is independent of model parameters θ. It is only relevant to the number of objects in I.

Assumption 2: $P(q \mid z=i, \theta, I)$ is a Gaussian distribution with identical variance 1.

Then we get:

$$\max_\theta Q(\theta, \theta^{old}) \tag{13}$$

$$= \max_\theta \sum_{i=1}^{m} \pi_i \cdot \log P(q \mid z=i, \theta, I) \tag{14}$$

$$= \max_\theta \sum_{i=1}^{m} \pi_i \cdot \frac{-(q-k_i)^2}{2\sigma_i^2} \tag{15}$$

$$= \max_\theta \sum_{i=1}^{m} \pi_i \cdot q^T k_i \tag{16}$$

Here, the reason for (13) $\rightarrow$ (14) is that $P(z=i \mid \theta, I)$ is a constant (Assumption 1) and has no effect on maximizing the Q-function. (14) $\rightarrow$ (15)

is due to the same reason. (15) $\rightarrow$ (16) can be explained by $-(q-k_i)^2 = -(q^2+k_i^2-2q^Tk_i) = 2q^Tk_i - 2$ when q and k_i are normalized. Note that, although we assume q and k_i are normalized here, we get a higher performance in practice when the features are not normalized. We attribute this to the magnitude of the features that learns a prior of z.

Meanwhile, the contrastive loss is:

$$\begin{aligned}\mathcal{L}_q &= -\sum_{k^+\in K^+} \pi_{k^+} \log \frac{\exp \frac{q^Tk^+}{\tau}}{\sum_{k\in K} \exp \frac{q^Tk}{\tau}} \\ &= \underbrace{(-\sum_{k^+\in K^+} \pi_{k^+} \frac{q^Tk^+}{\tau})}_{(a)} + \underbrace{\log \sum_{k\in K} \exp \frac{q^Tk}{\tau}}_{(b)}\end{aligned} \tag{17}$$

Here, (a) is regarded as an alignment item, and (b) is viewed as a uniformity item (Wang and Isola, 2020; Wang et al., 2022). The term (a) ensures that semantically similar samples are close to each other, which corresponds to maximizing Q-function in Eqn. (16). The term (b) is indispensable, although it does not appear in Eqn. (16). It is said to ensure that sample features do not collapse to a point and have rich semantics (Wang and Isola, 2020; Wang et al., 2022). Here, it also provides signals to disambiguate the co-occurrence objects and reveal the ground truth. More analysis about the term (b) is in subsection 5.2 Ablation Study.

To conclude, we can solve the original problem Eqn. (2) by minimizing the contrastive loss $\mathcal{L}_q$ and updating pseudo label π iteratively.

5 Experiments

Here, we present our experiments and results.

Datasets We adopt two widely used benchmarks Flickr30K Entities (Plummer et al., 2017) and RefCOCO+ (Yu et al., 2016; Kazemzadeh et al., 2014). Flickr30K Entities is an ectension of Flickr30K dataset (Young et al., 2014), built for the Phrase Grounding task. It contains 30k, 1k and 1k training, validation and testing images, respectively. Each image is accompanied with five captions.

RefCOCO+ is a widely-used Referring Expression dataset collected in a two-player game within a limited time. Unlike Flickr30K Entities which contains complete sentences, RefCOCO+ typically contains noun phrases. We adopt UNC split (Yu et al., 2016), which contains four parts: train, validation, testA and testB.

Evaluation Metrics Following the standard protocol in previous work (Wang et al., 2021; Gupta et al., 2020; Wang et al., 2020; Rohrbach et al., 2016), we adopt top-1 accuracy as the evaluation metric. A boundary box that overlaps the ground truth with $IoU > 0.5$ is considered to be correct. The annotation for a phrase may involve several bounding boxes. We merge them following previous work (Wang et al., 2020).

Implementation Details The image features are extracted by a Faster R-CNN (Ren et al., 2015) pre-trained on the Visual Genome (Krishna et al., 2017) dataset with a ResNet-101 (He et al., 2016) backbone. We use the same Flickr30K image features as MAF (Wang et al., 2020) and RefCOCO+ as VOLTA (Bugliarello et al., 2021), which can be obtained from their respective repositories[1]. The text features are 300-dimension GloVe embeddings (Pennington et al., 2014). Hyperparameters are tuning on the validation set. Best moving average hyperparameter $\lambda = 0.85$ and the best scale hyperparameter $\sigma = 10$. We use SGD as the optimizer, without momentum or weight decay. We train the model for 80 epochs using a batch size of 256 with a learning rate of 5e-4. The temperature τ is assigned to 1. Dropout rate is 0.1. we implement our model using PyTorch on a Linux machine with a GPU device Tesla P100 PCIE 16G.

5.1 Main Results

As shown in Table 1, our approach consistently outperforms previous methods by a large margin on both datasets. We obtain 63.05% accuracy on Flickr30K, 59.51% on RefCOCO+ TestA and 43.46% on TestB. That is, a gain of 0.95%, 11.62% and 5.26% when compared with previous SoTA, respectively. Note that previous SoTA use a complicated pre-tained visual language model to model the relationship between objects and phrases, whereas, we merely use the GloVe embedding and inner product similarity. By comparing with the prior contrastive learning methods, we also demonstrate that our expectation-maximization strategy is effective for the task.

5.2 Ablation Study

In this subsection, we analyze the contribution of each model component by conducting ablation studies. We demonstrate that the contrastive loss

[1]https://github.com/qinzzz/Multimodal-Alignment-Framework; https://github.com/e-bug/volta

Method	Backbone	Language	Proposals	Flickr30K	RefCOCO+	
					TestA	TestB
MATN (Zhao et al., 2018)	VGG16	LSTM	-	33.10	-	-
ARN (Liu et al., 2019)	RN101	LSTM	Mask-RCNN (CC)	-	34.40	36.12
Relation-aware(Liu et al., 2021)	RN101	LSTM	Faster-RCNN (VG)	59.27	-	-
W-visualBERT (Dou and Peng, 2021)	RNXT152	VL-BERT	Faster-RCNN (VG)	<u>62.10</u>	<u>47.89</u>	<u>38.20</u>
Align2Ground (Datta et al., 2019)	RN152	LSTM	Faster-RCNN (VG)	11.20	-	-
CCL (Zhang et al., 2020)	RN101	GRU	Faster-RCNN	-	36.91	33.56
InfoGround (Gupta et al., 2020)	RN101	BERT	Faster-RCNN (VG)	51.67	-	-
MAF (Wang et al., 2020)	RN101	GloVe	Faster-RCNN (VG)	61.43	17.10*	13.50*
KD+CL (Wang et al., 2021)	RN101	LSTM	Faster-RCNN (OI)	53.10	-	-
EM+CL(ours)	RN101	GloVe	Faster-RCNN (VG)	**63.05**	**59.51**	**43.46**

Table 1: Comparison of weakly supervised phrases grounding accuracy on Flickr30K Entities and RefCOCO+ test sets. Above are reconstruction-based methods and below are contrastive-based methods. The best values are in bold. The second are underlined. (VG) (CC) (OI) denote the object detector pre-trained on Visual Genome, MSCOCO, and OpenImage dataset. * are reported by W-visualBERT.

Method	Strategy				Flickr30K		RefCOCO+		
	hard	MA	update	CL	val	test	val	testA	testB
EM+CL(ours)	✓	✓	✓	✓	61.67	63.05	52.49	59.51	43.46
w/o updating	✓	✓	-	✓	38.28	39.72	38.71	41.84	36.02
w/o contrastive loss	✓	✓	✓	-	53.43	55.09	20.84	19.58	24.87
hard + non-MA	✓	-	✓	✓	60.50	61.81	52.58	60.09	43.39
soft + MA	-	✓	✓	✓	54.96	57.04	51.78	57.53	43.25
soft + non-MA	-	-	✓	✓	57.76	59.43	52.23	58.51	43.11

Table 2: Ablation study. hard, MA, update and CL mean hard prediction, moving average, updating pseudo label, and contrastive loss, respectively.

and pseudo label update are both significant to optimization, and give a geometric understanding.

The model ablation are characterized by the following, (1) **Without Pseudo Label Update**: we remove the pseudo label update step. In other words, all the positive examples have a fixed and uniformed confidence during the training; (2) **Without Contrastive Loss**: we set the number of negative objects from other pictures to zero. So the model can only use objects in the paired image; (3) **Different Updating Strategy**: we experiment with different updating strategies, following PiCO (Wang et al., 2022). We consider updating with moving average (as shown in (9)), without moving average (formulated as ${\boldsymbol{\pi}_{\boldsymbol{q}}}^{new} = \boldsymbol{s}$), hard prediction (as shown in (10)), or soft labels (formulated as $\boldsymbol{s} = \text{softmax}_j(\boldsymbol{q}^T {\boldsymbol{k}_{\boldsymbol{j}}}^+)$); (4) **Effect of Hyperparameters**: we vary the hyperparameters λ and σ to observe its effect on the model performance.

Table 2 shows our ablation results. There is a dramatic drop in performance when we ignore the update of pseudo labels or the contrastive loss. Specifically, without updating pseudo labels, the model always regards all positive samples as equally important and pulls the phrase toward each positive example with the same force. It is difficult to spot the ground truth hidden in the positive samples, even if we provide a large number of negative examples. Without the contrastive loss, that is, without using objects in other images as negative samples, the model does not have enough clues to distinguish which one is preferred and which is not. We may update the pseudo label in a wrong direction. Only combining them as a whole can bring in an excellent performance.

To further understand the role of the uniformity term (b) in Eqn. (17), we include more ablations on the contrastive loss. Specifically, we keep the batch size and vary the number of negative samples in the batch used for computing the contrastive loss. When the number equals 0, the model is the same as w/o contrastive loss; when the number equals batch size, the model is the same as the proposed one. For 0,1,4,64 negative samples, RefCOCO+ TestA/TestB accuracies are respectively 19.58/24.87, 48.57/38.13, 56.37/40.76, 59.35/43.27, and Flickr30K Test accuracies are 54.79, 62.11, 62.88, 63.01 respectively. Compar-

ing the result between 0 and 1 negative sample, we observe that the absence of negative samples causes a dramatic drop in accuracy. With an increasing number of negative samples, we observe a slow and steady increase in performance.

Since false positives are semantically similar to the negatives, when we push away a negative sample from the query phrase, we push away the false-positive samples as well. The model thus acquires some ability to distinguish the false negatives from the phrase encoding. Without negative samples, the model will fail to distinguish false-positives from the true-positive. The EM approach devised in this work just leverages and amplifies this effect, by gradually removing the false positives from the positives via down-weighting them in the pseudo label.

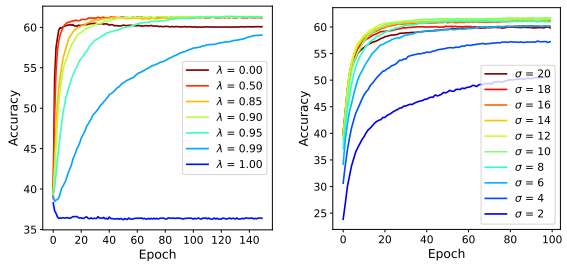

Figure 3: The validation accuracy on Flickr30K Entities for different λ and σ.

Moreover, we experiment with different updating strategies. Similar to the observation by Wang et al. (2022), all four strategies obtain competitive result. Meanwhile, the moving average strategy degenerates the performance of the soft prediction variant. Unlike the RefCOCO+ dataset, we find that Flickr30K is more sensitive to updating strategies.

We also observe the effect of the hyperparameters λ and σ. As shown in Figure 3 (left), a smaller λ results in a quicker convergence but with relatively lower performance. A similar observation is also noted in PiCO. At $\lambda = 0$, we gain competitive results. However, a large λ also hurts performance. When $\lambda = 1$, i.e., not updating pseudo label, the performance drops dramatically. It does not even go through a growth phase. We get the best results when $\lambda = 0.85$. Moreover, the sum scale parameter σ is important in this task. As shown in Figure 3 (right), scaling down the norm of phrase features results in a quicker convergence. We get best results when $\sigma = 10$.

5.3 Case Study

To better understand the capacity of our Contrastive Loss Module, we visualize the model's prediction for different epochs in Figure 4. We can observe that our method can refine the prediction progressively. Initially, the probability is nearly uniform. As training progresses, the model's confidence in the bounding box increases and gets closer to the ground truth, as shown in epoch 30.

Figure 5 also shows a number of results on RefCOCO+ validation set. We can observe that our model can handle cases with simple phrases, which occupies a large proportion in RefCOCO+. As shown in Figure 5 (a) and (b), our model has a proper understanding of gender, size, object name, etc. The results demonstrate its effectiveness.

We also show two kinds of failure in our model. Figure 5 (c) shows that our model cannot effectively handle cases which require complex reasoning. We attribute this to the fact that our method neglects the modelling of the sentence structure. Accordingly, we experiment with an LSTM text encoder to capture the sentence structure but achieve no performance gain. It is challenging to perform weakly supervised learning on cases requiring complex reasoning. We therefore leave it for future exploration. (d) also shows that our model lacks a tacit agreement with humans. It selects the proper object if mirrored from the opposite side. This tacit agreement is also hard to achieve because there is no annotation to tell the model what the tacit agreement is.

6 Conclusion

In this paper, we identify the significance of the intrinsic similarity and propose a novel contrastive learning framework based on the expectation-maximization algorithm to solve weakly supervised phrase grounding. Our method can adaptively refine the model prediction by minimizing a contrastive loss and chronologically updating pseudo targets. To validate our method, we conduct extensive experiments and deliver a theoretical analysis from the perspective of the maximum likelihood estimate. Our proposed approach achieves state-of-the-art performance on two widely used datasets, Filckr30K Entities and RefCOCO+. For future works, a more sophisticated language model is worth to be explored for phrase modeling to associate complex phrases with an object. We are also interested in investigating to what extent our

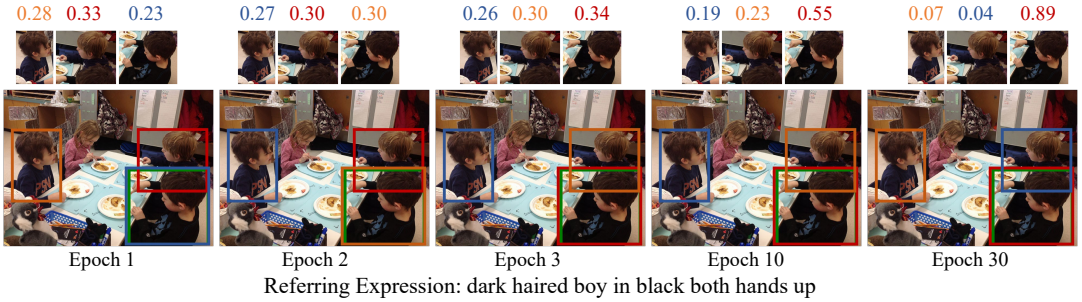

Figure 4: Model prediction for different epochs on RefCOCO+ validation set. The green box is ground truth and the red, yellow, blue box is top-1,2,3 prediction, respectively. Above is the probability for each prediction.

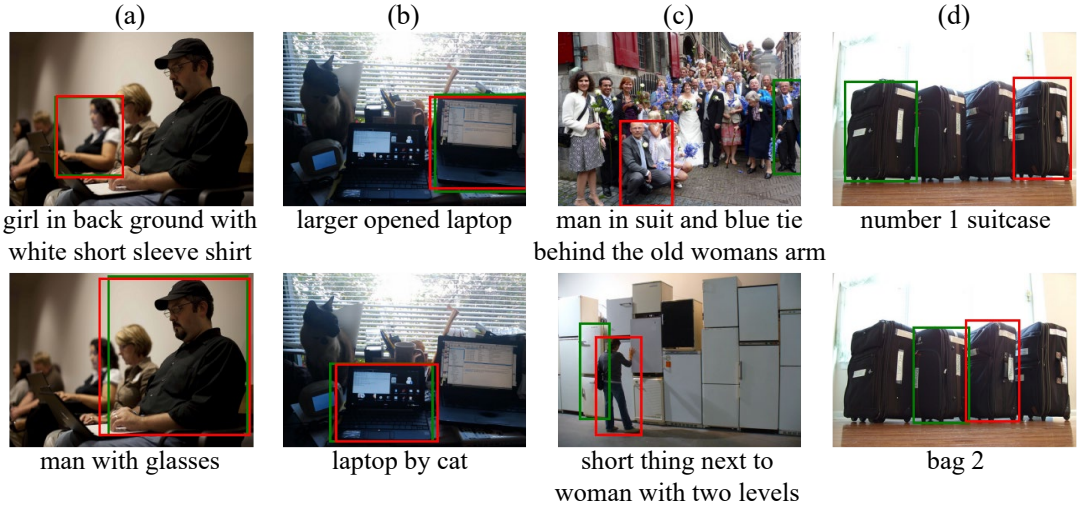

Figure 5: Visualization of weakly phrase grounding on RefCOCO+ validation set. The red box in image is the model prediction and the green box is the ground truth. (a) and (b) are results from simple phrases. (c) are results from complicated phrases which require complex reasoning. (d) shows some unclear annotations.

approach may take effect in other application problems of a "weak supervision" nature, for example, weakly supervised referring expression segmentation.

Limitations

The main limitation is that our method follows a 2-stage paradigm: object detection and then phrase grounding, which limits the scope of usage because a satisfactory pre-trained object detector may be hard to acquire. There is also a chance of error propagation to the phrase grounding stage if the object detector fails to extract the bounding boxes of the ground truths. Besides, the grounding model will require re-training when modifying the detector. It is interesting to explore how to perform weakly supervised grounding in a detector-free context.

Acknowledgements

This work was supported in part by the National Key R&D Program of China under Grant 2021ZD0110700, in part by the Fundamental Research Funds for the Central Universities, in part by the State Key Laboratory of Software Development Environment. SM is supported by a Leverhulme Trust Research Project Grant (No. RPG-2020-148).

References

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. 2021. Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language BERTs. *Transactions of the Association for Computational Linguistics*, 9:978–994.

Kan Chen, Trung Bui, Chen Fang, Zhaowen Wang, and Ram Nevatia. 2017. AMC: attention guided multi-modal correlation learning for image search. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6203–6211. IEEE Computer Society.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for

contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. 2019. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2601–2610. IEEE.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Zi-Yi Dou and Nanyun Peng. 2021. Improving pre-trained vision-and-language embeddings for phrase grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6362–6371, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. 2020. Contrastive learning for weakly supervised phrase grounding. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part III*, volume 12348 of *Lecture Notes in Computer Science*, pages 752–768. Springer.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9726–9735. Computer Vision Foundation / IEEE.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

Binbin Huang, Dongze Lian, Weixin Luo, and Shenghua Gao. 2021. Look before you leap: Learning landmark features for one-stage visual grounding. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 16888–16897. Computer Vision Foundation / IEEE.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. 2021. MDETR - modulated detection for end-to-end multi-modal understanding. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1760–1770. IEEE.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar. Association for Computational Linguistics.

Aisha Urooj Khan, Hilde Kuehne, Kevin Duarte, Chuang Gan, Niels da Vitoria Lobo, and Mubarak Shah. 2021. Found a reason for me? weakly-supervised grounded visual question answering using capsules. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 8465–8474. Computer Vision Foundation / IEEE.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven Chu-Hong Hoi. 2021a. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 9694–9705.

Junnan Li, Pan Zhou, Caiming Xiong, and Steven C. H. Hoi. 2021b. Prototypical contrastive learning of unsupervised representations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Dechao Meng, and Qingming Huang. 2019. Adaptive reconstruction network for weakly supervised referring expression grounding. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2611–2620. IEEE.

Yongfei Liu, Bo Wan, Lin Ma, and Xuming He. 2021. Relation-aware instance refinement for weakly supervised visual grounding. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 5612–5621. Computer Vision Foundation / IEEE.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word

representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2017. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *Int. J. Comput. Vis.*, 123(1):74–93.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Dhanesh Ramachandram and Graham W. Taylor. 2017. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6):96–108.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99.

Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. 2016. Grounding of textual phrases in images by reconstruction. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, volume 9905 of *Lecture Notes in Computer Science*, pages 817–834. Springer.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

Haobo Wang, Ruixuan Xiao, Yixuan Li, Lei Feng, Gang Niu, Gang Chen, and Junbo Zhao. 2022. PiCO: Contrastive label disambiguation for partial label learning. In *International Conference on Learning Representations*.

Liwei Wang, Jing Huang, Yin Li, Kun Xu, Zhengyuan Yang, and Dong Yu. 2021. Improving weakly supervised visual grounding by contrastive knowledge distillation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 14090–14100. Computer Vision Foundation / IEEE.

Qinxin Wang, Hao Tan, Sheng Shen, Michael Mahoney, and Zhewei Yao. 2020. MAF: Multimodal alignment framework for weakly-supervised phrase grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2030–2038, Online. Association for Computational Linguistics.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9929–9939. PMLR.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling context in referring expressions. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*, volume 9906 of *Lecture Notes in Computer Science*, pages 69–85. Springer.

Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. 2018. Grounding referring expressions in images by variational context. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4158–4166. Computer Vision Foundation / IEEE Computer Society.

Zhu Zhang, Zhou Zhao, Zhijie Lin, Jieming Zhu, and Xiuqiang He. 2020. Counterfactual contrastive learning for weakly-supervised vision-language grounding. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Fang Zhao, Jianshu Li, Jian Zhao, and Jiashi Feng. 2018. Weakly supervised phrase localization with multi-scale anchored transformer network. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5696–5705. Computer Vision Foundation / IEEE Computer Society.