Momentum Pseudo-Labeling for Weakly Supervised Phrase Grounding

Dongdong Kuang[1], **Richong Zhang**[1,3*], **Zhijie Nie**[1,4], **Junfan Chen**[1,2], **Jaein Kim**[1]

[1] CCSE, School of Computer Science and Engineering, Beihang University, Beijing, China
[2] School of Software, Beihang University, Beijing, China
[3] Zhongguancun Laboratory, Beijing, China
[4] Shen Yuan Honors College, Beihang University, Beijing, China
{kuangdd, zhangrc, niezj, chenjf}@act.buaa.edu.cn, jaein@buaa.edu.cn

Abstract

Weakly supervised phrase grounding tasks aim to learn alignments between phrases and regions with coarse image-caption match information. One branch of previous methods established pseudo-label relationships between phrases and regions based on the Expectation-Maximization (EM) algorithm combined with contrastive learning. However, adopting a simplified batch-level local update (partial) of pseudo-labels in E-step is sub-optimal, while extending it to global update requires inefficiently numerous computations. In addition, their failure to consider potential false negative examples in contrastive loss negatively impacts the effectiveness of M-step optimization. To address these issues, we propose a Momentum Pseudo Labeling (MPL) method, which efficiently uses a momentum model to synchronize global pseudo-label updates on the fly with model parameter updating. Additionally, we explore potential relationships between phrases and regions from non-matching image-caption pairs and convert these false negative examples to positive ones in contrastive learning. Our approach achieved SOTA performance on 3 commonly used grounding datasets for weakly supervised phrase grounding tasks.

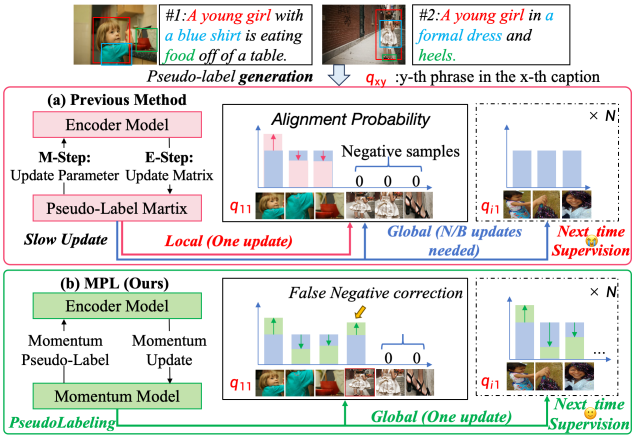


Figure 1: Comparison of pseudo-label updating mechanisms. (a) Previous method: Utilizes an encoder model to update the pseudo-label matrix locally within each batch. (b) MPL: Employs a momentum model to perform global updates on the pseudo-labels across batches and addresses the issue of potential false negatives.

Introduction

Phrase grounding is a crucial task in multimodal learning, involving the identification of specific regions within an image that corresponds to a given textual description (Liu and Hockenmaier 2019). This task has significant practical applications, including visual question answering (Chen, Anjum, and Gurari 2022) and cross-modal regions retrieval (Li* et al. 2022). Existing fully supervised methods (Huang et al. 2021; Zhang et al. 2022) depend heavily on extensive annotations of fine-grained object detection bounding boxes, which provide precise locations of objects within an image to achieve high performance. However, obtaining such detailed annotations is costly. In contrast, matched image-caption pairs, where each image is paired with a descriptive caption, are more readily available and do not require detailed spatial information about object locations. This ease of availability leads to growing interest in weakly supervised phrase grounding (WSPG), which aims to learn the correspondence between phrase and image regions without the need for extensive annotations.

Recent works in WSPG (Gupta et al. 2020; Wang et al. 2020; Chen et al. 2022; Rigoni et al. 2023; Zhang, Wang, and Liu 2023) have utilized image-caption level or phrase-region level contrastive losses based on image-caption supervision to align the representations of phrases and regions. However, under a weakly supervised setting, it is challenging to accurately identify positive samples during contrastive learning. To address this difficulty, these methods establish pseudo-label relationships between phrases and regions within matched image-caption pairs based on the Expectation-Maximization (EM) algorithm.

However, existing methods encounter two main challenges. First, they may encounter efficiency problems in E-step computation. As shown in Figure 1 (a), to quickly update the posterior probabilities of the latent variable, they adopt a simplified batch-level local update of pseudo-labels in E-step, usually on a batch of image-caption examples and fail to update the out-of-batch examples. This updating strat-

*Corresponding author

egy is sub-optimal because a typical EM optimization requires a global update that needs to compute posterior probabilities for all examples in the whole dataset. We refer to this as the **Slow Pseudo-label Update Issue**. The necessity of such global updates is verified by our empirical study (Figure 4, Table 2), which shows that, with the same number of updates, global update achieves significant improvements compared with local update. Attempting to extend existing methods to global update requires more computation.

Second, the effectiveness of M-step optimization of existing methods may be affected by potential false negative examples in contrastive learning. Existing methods focus solely on the matching relationships between phrases and regions within matched image-caption pairs, without considering potential matching relationships from other image-caption pairs. This approach treats all regions in non-matching images as negative samples.

As shown in the top of Figure 1, both pairs of example image-captions feature the phrase *a young girl* and visually similar regions depicting a girl. However, simply treating the phrase-region pairs in non-matching image-caption pairs as negative examples may lead to the issue of false negatives, thereby impacting the effectiveness of contrastive learning and the consistency performance of grounding in M-step optimization. We refer to this as the **False Negative Impact**.

To address these challenges, we first propose to introduce a Momentum Pseudo Labeling (MPL) framework to globally update pseudo-labels in E-step computation to address the **Slow Pseudo-label Update Issue**. As shown in Figure 1 (b), instead of performing a local update on a batch of examples, MPL performs global pseudo-label computation by only one update to the momentum model parameters on the fly. In this way, the pseudo-labels for any examples in the next M-step can be easily accessed from the updated momentum model. Furthermore, within the MPL framework, we model the relationships between phrases and regions in non-matching images, exploring potential connections between region-phrase pairs across unmatched image-caption pairs to mitigate the **False Negative Impact**.

In summary, our key contributions are:

- We introduce a new pseudo-label updating strategy, Momentum Pseudo Labeling (MPL), which enables us to efficiently perform global updates of pseudo-labels.
- We consider potential phrase-region relationships in non-matching image-caption pairs and treat the false negative examples as pseudo-positive samples to improve contrastive learning in MPL.
- Experimental results and intensive analysis on three commonly used phrase grounding datasets demonstrate the efficiency and effectiveness of our MPL approach.

Problem Analysis

In this section, we introduce the problem setup of WSPG and revisit the previous methods (Wang et al. 2020; Chen et al. 2022; Zhang, Wang, and Liu 2023) from the perspective of the Expectation-Maximization algorithm (Moon 1996).

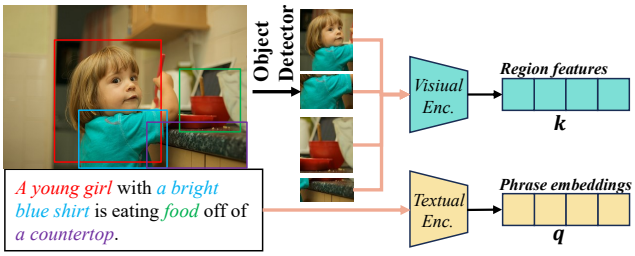


Figure 2: Diagram of the Phrase-Region Dual-Encoder

Problem Formulation

Consider an image-caption pair, (I_i, T_i). Let $Q(T_i)$ be the set of phrases in the caption T_i, and let $K(I_i)$ denote the set of regions extracted from the image I_i, namely $Q(T_i)$ and $K(I_i)$ are the partitions of the caption T_i and the image I_i, respectively. The goal of phrase grounding is to locate a region $k \in K(I_i)$ in the image I_i for each phrase $q \in Q(T_i)$ from the corresponding caption T_i, such that the phrase refers to an object within that region.

For convenience, we may also use q and k to refer to the index of the corresponding phrase in the caption and the region in the image, respectively.

EM Algorithm Perspective

In the weakly supervised scenario, the target regions in the matched image for the phrase q are ambiguous. Specifically, for each q, we introduce a variable z to represent the correspondence between phrase q and regions $k \in K(I_i)$. Since z is usually unobservable, we treat it as a latent variable. The set of latent variables for a given q is defined as:

$$Z = \{z_{qk} | k \in K(I_i)\}. \tag{1}$$

Moreover, the EM algorithm is particularly effective for estimating model parameters involving such latent variables.

EM Algorithm The EM algorithm is iterative and includes two steps: the E-step and the M-step. In the E-step of t-th iteration, the algorithm estimates the posterior probabilities of the latent variables Z given the current parameter estimates $\theta^{(t-1)}$ and the observed data (e.g., image-caption pairs), denoted as X. In the M-step, the algorithm then maximizes the expected log-likelihood (also known as the **Q**-function) to update the parameters, yielding θ^t:

$$\theta^t = \arg\max_{\theta} \underbrace{\int_Z \log P(X, Z|\theta) \cdot P(Z|X, \theta^{(t-1)}) dZ}_{\textbf{Q}\text{-function}}, \tag{2}$$

where $P(X, Z|\theta)$ represents the joint probability of the observed variables X and the latent variables Z and $P(Z|X, \theta^{(t-1)})$ is the posterior probability of the latent variables. For simplicity, given a phrase q, we denote $P(Z|X, \theta^{(t-1)})$ corresponding to region k as $\pi_{qk}^{(t-1)}$, and refer to it as the pseudo-label.

Revisiting from an EM Algorithm Perspective Existing contrastive learning-based methods (Wang et al. 2020; Chen et al. 2022; Zhang, Wang, and Liu 2023) can be summarized as implementing the E-step by estimating the pseudo-labels π_{qk} based on the current model parameters $\theta^{(t-1)}$ and a batch of B image-caption pairs data $X = \{(Q(T_i), K(I_i))\}_{i=1}^{B}$. In the M-step, the model parameters θ^t are obtained by applying a single step of stochastic gradient descent (SGD).

More precisely, in existing methods, the posterior probability of latent variables is updated according to the following equation:

$$\pi_{qk}^{t} = \lambda \cdot \pi_{qk}^{(t-1)} + (1-\lambda) \cdot s_{qk}, \tag{3}$$

where $\pi_{qk}^{(t-1)}$ is the posterior probability of latent variables from the last iteration, and s_{qk} is computed as

$$s_{qk} = \begin{cases} 1 & \text{if } k = \underset{1 \le c \le |K(I_i)|}{\arg\max} \ \langle \boldsymbol{q}, \boldsymbol{k}_c \rangle, \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where $\boldsymbol{q}$ and $\boldsymbol{k}_c$ represent the output features from a dual encoder as shown in Figure 2 of the phrase q in T_i and c-th region in I_i, respectively, with respect to current model parameters $\theta^{(t-1)}$. $\langle ., . \rangle$ denotes the dot product operation.

As concrete examples, in the context of MAF from Wang et al. (2020) and CC from Zhang, Wang, and Liu (2023), their $\mathbf{Q}$-function can be represented by the image-caption level cross-entropy, as given below:

$$\sum_{i=1}^{B} \frac{1}{B} \cdot \log \frac{\exp\left(\sum\limits_{q \in Q(T_i)} \sum\limits_{k^+ \in K(I_i)} \pi_{qk^+} \cdot \langle \boldsymbol{q}, \boldsymbol{k}^+ \rangle / |Q(T_i)|\right)}{\sum\limits_{j=1}^{B} \exp\left(\sum\limits_{q \in Q(T_i)} \max\limits_{k \in K(I_j)} \langle \boldsymbol{q}, \boldsymbol{k} \rangle / |Q(T_i)|\right)}. \tag{5}$$

Similarly, for CLEM as discussed in Chen et al. (2022), the pseudo-label estimation formula during the E-step is identical to that of MAF and CC. However, it replaces the $\mathbf{Q}$-function with the alignment term from the InfoNCE loss:

$$\sum_{i=1}^{B} \sum_{q \in Q(T_i)} \sum_{k^+ \in K(I_i)} \frac{\pi_{qk^+}}{B \cdot |Q(T_i)|} \cdot \log \frac{\exp(\langle \boldsymbol{q}, \boldsymbol{k}^+ \rangle)}{\sum\limits_{k \in K(\mathtt{I})} \exp(\langle \boldsymbol{q}, \boldsymbol{k} \rangle)}, \tag{6}$$

where $K(\mathtt{I})$ in the denominator represents the set of regions from all images in the batch.

Motivating Factors

Efficiency in E-step Computation: Ideally, the E-step in the EM algorithm requires updating the posterior probabilities of *pseudo-labels* π_{qk} for all N samples $(Q(T_i), K(I_i))$ in the dataset. However, this is usually computationally expensive because N is very large in practice, often exceeding $10K$. Therefore, as demonstrated in Equation (5) and (6), to quickly update the posterior probabilities π_{qk}, existing works adopt a local update strategy in the E-step. Specifically, they only update π_{qk} of a mini-batch of $(Q(T_i), K(I_i))$ samples and ignore the update of pseudo-labels outside of that batch. This strategy may lead to sub-optimal estimation of the posterior probabilities π_{qk} and degrade the EM performance under for **Slow Pseudo-label Update Issue**, as evidenced in Figure 4. However, applying global updates to all samples $(Q(T_i), K(I_i))$ with existing methods is inefficient, requiring at least N/B updates.

Effectiveness in M-step Optimization: Existing methods transform the maximization of the $\mathbf{Q}$-function in M-step into minimizing a contrastive loss, where the positive and negative examples are determined by the pseudo-labels estimated with π_{qk}. Therefore, treating one of these similar regions as a positive sample and the others as negatives can potentially confuse the model (Huynh et al. 2022), thereby influencing the parameter updates in the M-step and the latent variable estimation in the next E-step.

Methodology

We introduce our MPL-based weakly supervised contrastive learning framework, as illustrated in Figure 3. Our framework is structured into three main components: the Dual Encoder, the Momentum Pseudo Labeling Module, and the Pseudo-label Guided Contrastive Loss. The Dual Encoder comprises text and image encoders responsible for encoding regions and phrases. The MPL module ensures stable and efficient pseudo-label updates leveraging the momentum model. Finally, the Pseudo-label Guided Contrastive Loss incorporates strategies to handle false negative region samples, improving the model's consistency.

Dual Encoder

Our framework utilizes a dual encoder architecture similar to previous work (Gupta et al. 2020; Wang et al. 2020; Chen et al. 2022). This dual encoder consists of a text encoder and an image encoder, as shown in Figure 2. In the following section, we will refer to the dual encoder as the base model, corresponding to the momentum model.

Text Encoder Our text encoder, similar to the one used in Chen et al. (2022), uses pre-trained GloVe embeddings (Pennington et al. 2014) to obtain word vector $\boldsymbol{w}$ for each word. Given a phrase q, the representation of a phrase is computed by summing the hidden states that constitute it:

$$\begin{aligned} \{\boldsymbol{h_i}\}_{i=1}^{n_q} &= F_t(\{\boldsymbol{w}_i\}_{i=1}^{n_q}), \\ \boldsymbol{q} &= W_q(\sum_{i=1}^{n_q} \boldsymbol{h}_i / \sigma), \end{aligned} \tag{7}$$

where F_t denotes one layer of LSTM network, $\boldsymbol{h}_i$ represents the i-th output hidden states, W_q is a linear mapping, σ is a hyperparameter scaling the text representation, and n_q is the number of words in the current phrase.

Image Encoder Our image encoder consists of two functional components. The first is a frozen pre-trained object detector responsible for identifying salient regions in the image and extracting visual features from these regions. The second component focuses on establishing relationships between these regions and mapping their features across spaces. For instance, given an image I, the object detector f_D outputs m pairs (l_i, v_i), where l_i represents the probable object label of the i-th region, and v_i denotes the high-dimensional features output from the object detector. Formally, $\{(l_i, v_i)\}_{i=1}^{m} = f_D(I)$.

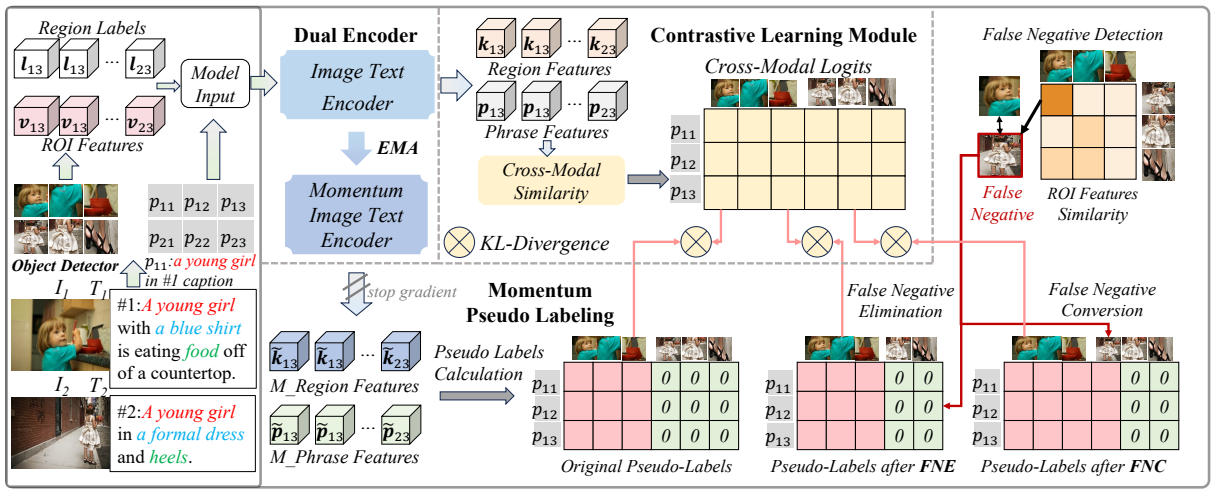


Figure 3: Illustration of our approach. Our MPL framework comprises three modules: Dual Encoder, Momentum Pseudo Labeling Module, and Contrastive Learning Module. Through the FNE and FNC strategies, the pseudo-label calculation in MPL considers the implicit relationships between phrases and regions in non-matching image-text pairs.

In the remaining part of the image encoder, similar to (Gupta et al. 2020; Chen et al. 2022), we use the outputs l_i to incorporate prior knowledge from object detectors into the region features. We utilize transformer encoder (Vaswani et al. 2017) to establish relationships among different region features, enhancing the regional perception capabilities within each image, as illustrated in the following equations:

$$\begin{aligned} \hat{\boldsymbol{k}}_i &= W_k(v_i) + \boldsymbol{l}_i, \\ \{\boldsymbol{k}_i\}_{i=1}^m &= E(\{\hat{\boldsymbol{k}}_i\}_{i=1}^m), \end{aligned} \tag{8}$$

where W_k is a linear projection applied to each region's visual features, $\boldsymbol{l}_i$ represents the word embedding corresponding to the object label of each region k, and E denotes the transformer encoder operation.

Momentum Pseudo Labeling Module

As mentioned earlier, previous methods, from the perspective of the EM algorithm, update pseudo-labels only within the current batch during the E-step due to efficiency considerations. To tackle this limitation, we propose the Momentum Pseudo Labeling (MPL) method. MPL uses a momentum model to accumulate updates to the dual encoder parameters during the M-step. This approach computes pseudo-labels for different batches in a more stable and smooth manner, providing a cost-effective and efficient way to approximate a global update of pseudo-labels.

Notably, our MPL method distinguishes itself from previous methods such as (Li et al. 2020; He et al. 2020; Li et al. 2021). Specifically, Li et al. (2020) apply a momentum model during the E-step to globally maintain certain cluster centers for latent variables, while He et al. (2020) utilize the momentum model to maintain queues of negative examples. Li et al. (2021) use a momentum model to compute pseudo-labels, reducing noise in the actual image-caption labels. Our approach utilizes the momentum model to compute pseudo-labels of phrase-region pairs to guide the parameter update of the base model.

Pseudo Labels Calculation Unlike ALBEF (Li et al. 2021), we only have image-caption level supervision in the current weakly supervised scenario, but finer-grained phrase-region alignment is needed. Under this constraint, we use the momentum model to compute pseudo-labels for phrase-region matches in matched image-caption pairs, serving as fine-grained supervision. These pseudo-labels guide the contrastive learning module in our framework, detailed in the next section. Given a phrase $q \in Q(T_i)$ and all regions within the batch $k^+ \in K(\mathtt{I})$, pseudo-labels are computed as follows:

$$\pi_{qk^+} = \begin{cases} 0 & \text{if } k^+ \in K(\mathtt{I}) \setminus K(I_i), \\ \frac{\exp(\langle \tilde{\boldsymbol{q}}, \tilde{\boldsymbol{k}}^+ \rangle / \tau_E)}{\sum_{k \in K(I_i)} \exp(\langle \tilde{\boldsymbol{q}}, \tilde{\boldsymbol{k}} \rangle / \tau_E)} & \text{otherwise.} \end{cases} \tag{9}$$

Where the pseudo-label π_{qk^+} represents the correspondence probability between q and k^+. The terms $\tilde{\boldsymbol{q}}$ and $\tilde{\boldsymbol{k}}$ denote the phrase and region features output by the momentum model, respectively. Additionally, τ_E is a temperature parameter. Here we simply set the matching probability of phrases and regions across different image-caption pairs to 0.

Momentum Model Update The update strategy for the momentum model follows a similar approach to that used in MoCo (He et al. 2020):

$$\tilde{\theta}^t = \gamma \cdot \tilde{\theta}^{(t-1)} + (1-\gamma) \cdot \theta^t, \tag{10}$$

where θ and $\tilde{\theta}$ represent the parameters of the base model and the momentum model, respectively, with t indicating a specific time step. γ is the momentum coefficient of the model. The initialization of the momentum model is identical to the base model initialization.

It is worth noting that our momentum model provides more consistent feature output for pseudo-label calculation by leveraging exponential moving averaging (EMA) and benefits from its low update overhead. Specifically, after each update of the base model parameters, the synchronously updated momentum model is used to calculate the pseudo-labels for the positive pairs in the next mini-batch of image-caption pairs. This can be viewed as implicitly completing the global update of pseudo-labels in the E-step after the M-step.

Pseudo-label Guided Contrastive Loss

Unlike traditional binary relationships in conventional image-text contrastive learning, our approach uses pseudo-labels between phrases and regions provided by the momentum model to establish positive relationships, which we term as pseudo-label guided contrastive learning. While π_{qk^+} is calculated using our MPL method, the contrastive loss for a phrase q can be expressed in the form of a KL divergence:

$$\mathcal{L} = -\sum_{k^+ \in K(\mathtt{I})} (\pi_{qk^+} \cdot \log \underbrace{\frac{\exp(\langle \boldsymbol{q}, \boldsymbol{k}^+ \rangle / \tau)}{\sum_{k \in K(\mathtt{I})} \exp(\langle \boldsymbol{q}, \boldsymbol{k} \rangle / \tau)}}_{\xi_{qk^+}} - \pi_{qk^+} \cdot \log \pi_{qk^+}). \tag{11}$$

Where ξ_{qk^+} represents the matching probability between the phrase q and the region k^+ as computed by the base model, since the last term is gradient-free, we will omit it in the following loss function.

About **False Negative Impact** mentioned earlier, we highlight how current methods (Wang et al. 2020; Chen et al. 2022) based on contrastive learning overlook the issue of false negatives under weakly supervised settings, potentially affecting the consistency of grounding. Considering the possible presence of false negatives, we propose two strategies to build connections between phrases and regions in non-matching image-caption pairs.

False Negative Elimination To mitigate the impact of false negative samples under a weak supervision setting, we retrieve potential false negative samples based on the similarity of regional features in the visual modality, as shown in Figure 3. When a given phrase $q \in Q(T_i)$, This set of potential false negatives can be represented as:

$$\begin{aligned} \mathcal{F}_q = \{ k' \mid k' \in K(\mathtt{I}) \setminus K(I_i), \exists k^+ \in K(I_i) \\ \text{s.t. } \cos(v_{k'}, v_{k^+}) > \phi \}, \end{aligned} \tag{12}$$

where v represents the high-dimensional visual features outputted by the detector, $\cos(.,.)$ denotes cosine similarity, and ϕ is the similarity threshold used to retrieve false negatives.

We attempt to ignore these potential false negative region samples in the loss calculation. We define the filtered set of remaining regions $\mathcal{K}^e(\mathtt{I}) = K(\mathtt{I}) \setminus \mathcal{F}_q$. The modified pseudo-label calculation is as follows:

$$\pi^e_{qk^+} = \begin{cases} 0 & \text{if } k^+ \in \mathcal{K}^e(\mathtt{I}) \setminus K(I_i), \\ \frac{\exp(\langle \tilde{\boldsymbol{q}}, \tilde{\boldsymbol{k}}^+ \rangle / \tau_E)}{\sum_{k \in K(I_i)} \exp(\langle \tilde{\boldsymbol{q}}, \tilde{\boldsymbol{k}} \rangle / \tau_E)} & \text{otherwise.} \end{cases} \tag{13}$$

The loss function is also modified as

$$\mathcal{L}^e_q = -\sum_{k^+ \in \mathcal{K}^e(\mathtt{I})} \pi^e_{qk^+} \cdot \log \frac{\exp(\langle \boldsymbol{q}, \boldsymbol{k}^+ \rangle / \tau)}{\sum_{k \in \mathcal{K}^e(\mathtt{I})} \exp(\langle \boldsymbol{q}, \boldsymbol{k} \rangle / \tau)}. \tag{14}$$

By designing the loss in this manner, we adjust the negative sampling under weakly supervised contrastive learning. This approach is referred to as False Negative Elimination.

False Negative Conversion To further investigate the role of the false negative examples detected and explore the potential relationships of phrase-region pairs from non-matching image-caption pairs, we design to convert region k' within $\mathcal{F}_q$ into potential positive sample to strengthen the grounding consistency of model.

We utilize the regions in $\mathcal{F}_q$ to expand $K(I_i)$, defining a new expanded set of regions $\mathcal{K}^c(I_i) = K(I_i) \cup \mathcal{F}_q$. Given a phrase $q \in Q(T_i)$, the calculation of the pseudo-labels is:

$$\pi^c_{qk^+} = \begin{cases} 0 & \text{if } k^+ \in K(\mathtt{I}) \setminus \mathcal{K}^c(I_i), \\ \frac{\exp(\langle \tilde{\boldsymbol{q}}, \tilde{\boldsymbol{k}}^+ \rangle / \tau_E)}{\sum_{k \in \mathcal{K}^c(I_i)} \exp(\langle \tilde{\boldsymbol{q}}, \tilde{\boldsymbol{k}} \rangle / \tau_E)} & \text{otherwise.} \end{cases} \tag{15}$$

After converting false negatives, the contrastive learning loss can be expressed as:

$$\mathcal{L}^c_q = -\sum_{k^+ \in K(\mathtt{I})} \pi^c_{qk^+} \cdot \log \frac{\exp(\langle \boldsymbol{q}, \boldsymbol{k}^+ \rangle / \tau)}{\sum_{k \in K(\mathtt{I})} \exp(\langle \boldsymbol{q}, \boldsymbol{k} \rangle / \tau)}. \tag{16}$$

To recap, within our framework, we designed the False Negative Elimination to modify the original negative sample sampling method, using the False Negative Conversion to transform negative samples into potential positive samples. These two methods are both used to mitigate the impact of potential false negatives on contrastive learning in the current weakly supervised setting. For clarity, we refer to our method incorporating the FNC strategy as MPL.

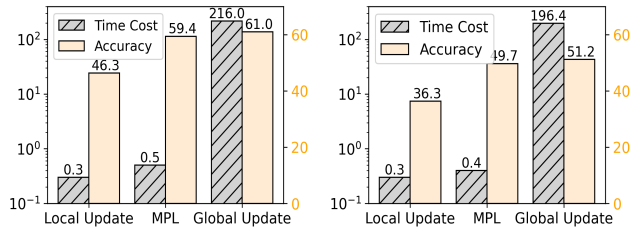
Experiments

Datasets and Metric Our main experimental results are derived from benchmarks on three publicly used datasets for phase grounding task, including the Flickr30k Entities (Plummer et al. 2015), RefCOCO and RefCOCO+ (Kazemzadeh et al. 2014; Yu et al. 2016). For the RefCOCO/+ dataset, we employ the UNC split (Yu et al. 2016), dividing both datasets into four parts: train, validation, testA, and testB. We evaluate our method using the **IoU@0.5** as utilized in previous works (Wang et al. 2020; Jin et al. 2023).

Implementation Details Following prior work, we extracted regions and their features using Faster R-CNN (Ren et al. 2016), which is pre-trained on Visual Genome (Krishna et al. 2017). For the Flickr30k Entities dataset, we used the image features as in MAF (Wang et al. 2020), while for RefCOCO/+ datasets, we used our image features aligned with CLEM (Chen et al. 2022). Regarding the hyperparameter settings: the momentum update coefficient γ is set to 0.99. For FNE, the threshold ϕ is set to 0.85, and for FNC, ϕ is set to 0.95. *For more training and evaluation details, please refer to our supplementary materials.* Our code can be accessed via https://github.com/Kuangdd01/MPL.

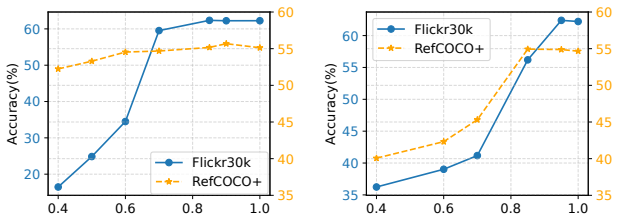
Method	Backbone	LM	Proposals	Flickr30k	RefCOCO		RefCOCO+	
					TestA	TestB	TestA	TestB
ARN (Liu et al. 2019)	RN101	LSTM	Faster-RCNN	-	35.27	36.47	34.40	36.12
W-visualBERT (Dou et al. 2021)	RNXT152	VL-BERT	Faster-RCNN (VG)	62.10	-	-	47.89	38.20
Pseudo-Q (Jiang et al. 2022)	RN101	BERT	Faster-RCNN (VG)	60.41	58.25	54.13	45.06	32.13
CPL (Liu et al. 2023)	RN101	BERT	Faster-RCNN (VG)	63.87	69.77	63.44	55.30	45.52
CCL (Zhang et al. 2020)	RN101	GRU	Faster-RCNN	-	37.64	32.59	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	51.76†	34.86†	32.20†	38.27†
KD+CL (Wang et al. 2021)	RN101	LSTM	Faster-RCNN (OI)	53.10	-	-	-	-
CLEM (Chen et al. 2022)	RN101	GloVe	Faster-RCNN (VG)	63.05	66.63†	54.60†	59.51	43.46
RefCLIP (Jin et al. 2023)	Darknet-53	GRU	YOLOv3(VG)	-	58.58	**57.13**	40.45	38.86
MPL$_{\text{FNC}}$ (ours)	RN101	GloVe	Faster-RCNN (VG)	**64.15**	**70.19**	55.74	**63.59**	**45.20**

Table 1: Comparison of three mainstream WSPG methods across three datasets. The top section pertains to methods based on modal reconstruction. The middle gray section represents methods that enhance weak supervision through additional data and model knowledge, while the bottom section and our method employ contrastive learning. (†) denotes our reproduction under identical settings. (VG) (CC) (OI) denote the object detector pre-trained on Visual Genome, MSCOCO, and OpenImage dataset.



(a) Flickr30K Entities (b) RefCOCO+

Figure 4: Comparison of performance and time cost between different pseudo-label update strategies on the validation set.



(a) Accuracy of FNE. (b) Accuracy of FNC.

Figure 5: The impact of different similarity threshold values on the effectiveness of FNE and FNC. Accuracy refers to the performance on the validation set across different datasets.

Update Method	Flickr30k	RefCOCO	RefCOCO+
Local Update	62.01	61.27	53.85
Global Update	62.32	63.57	54.49
MPL (ours)	62.44	63.22	55.81

Table 2: The impact of different strategies for pseudo-label accuracy on the training set.

Main Result

As shown in Table 1, we demonstrate the top-1 accuracy of our method on three datasets. Our method outperforms others that are also based on contrastive learning and show comparable performance to methods (Jiang et al. 2022; Liu et al. 2023) that require additional pre-training and knowledge of multimodal models. Specifically, it outperformed the previous SOTA, CLEM (Chen et al. 2022), which was based on weakly-supervised contrastive learning, by 1.1% on the Flickr30k, 3.6%/1.1% on the RefCOCO testA/testB, and 4.1%/1.7% on the RefCOCO+ testA/testB.

Ablation Study

Comparison of Pseudo-label Updates To demonstrate the advantages of our MPL method in pseudo-label updating, we designed comparative experiments comparing Local pseudo-label updating, Global pseudo-label updating, and our methods, as shown in Figure 4. Both the Local and Global updates of pseudo-labels follow Equation (3) with settings consistent with those used in Chen et al. (2022). The difference is that the global update refreshes the pseudo-labels across the entire training dataset after a batch-based parameter update, whereas the local update only updates the pseudo-labels within the current batch. To ensure fairness, three models were trained for the same number of steps with the same scale of trainable parameters, differing only in the form of pseudo-label updates.

From Figure 4, methods that update pseudo-labels globally yield better results but incur greater time costs. Conversely, local pseudo-label updating methods reduce time costs during training but suffer from slower pseudo-label updates, leading to poorer performance under insufficient training steps. MPL enables the modeling of global pseudo-labels without significant additional time costs, achieving faster convergence and effectiveness comparable to global updating methods. In Table 2, we compared the pseudo-label accuracy at the end of model training under different update strategies on three datasets. Our method outperforms the Lo-

Strategy			Flickr30k		RefCOCO		
EMA	FNE	FNC	val	test	val	testA	testB
-	-	-	61.59	62.47	62.07	68.31	54.80
-	✓	-	61.18	62.45	62.69	68.41	55.23
-	-	✓	62.09	63.72	63.66	69.68	55.07
✓	-	-	62.25	64.15	63.82	69.77	55.53
✓	✓	-	62.35	64.02	63.45	70.05	**56.11**
✓	-	✓	**62.53**	**64.15**	**64.07**	**70.19**	55.74

Table 3: Ablation study on EMA and False Negatives handling strategies. "FNE" stands for False Negative Elimination, and "FNC" denotes False Negative Conversion.

cal Update approach regarding pseudo-label accuracy, providing better supervision for contrastive learning.

Ablation of Weakly Supervised Training We further examined the effect of EMA on the pseudo labeling and false negative handling strategies within the MPL framework with results shown in Table 3. The findings demonstrate that EMA helps improve model performance, confirming the effectiveness of EMA for the pseudo-labeling module. Additionally, converting false negatives within batches can also enhance model performance consistently.

To further explore the effect of the similarity thresholds ϕ on FNE and FNC, we recorded the performance of MPL with different threshold values as shown in Figure 5.

Ablation of momentum Referring to the settings of the momentum coefficient in MoCo (He et al. 2020), we considered the following momentum values γ as shown in Table 4. The results represent the accuracy of the model on the Flickr30K validation set under different momentum values.

momentum γ	0	0.9	0.99	0.999
Accuracy(%)	61.59	62.22	62.25	62.20

Table 4: The impact of the momentum coefficient on model performance on the Flickr30k validation set.

Case Study

To more vividly demonstrate the effectiveness of our method in the pseudo-labeling and prediction, we visualized some of the pseudo-labels during the training phase. The results in the upper part of Figure 6 are from CLEM, while the lower part shows the results from our method. As shown in Figure 6, the left images depicts the top three confidence target regions for a phrase during training; the right images show the predictions of our model. The dark blue box represents the golden box (invisible during training). The orange box in the images on the right represents prediction results. Where Orange, yellow, and sky blue boxes indicate regions with the top 1, 2, and 3 confidence levels, respectively.

Related Work

Weakly Supervised Phrase Grounding In WSPG tasks, previous studies have been divided into three primary classifications. The first strategy employs a multimodal mask-reconstruction loss to enhance the model's capacity in comprehending intricate connections between images and captions (Li et al. 2019, 2021; Dou et al. 2021; Zhao et al. 2023). The second strategy incorporates knowledge distillation from multimodal models like BLIP (Li et al. 2022) to utilize their captioning abilities and transform weakly supervised assignments into fully supervised ones (Jiang et al. 2022; Liu et al. 2023) or use the attention-based heatmaps generated by multimodal models to achieve weakly-supervised grounding (Shaharabany, Tewel, and Wolf 2022; Shaharabany and Wolf 2023; Lin et al. 2024). The final approach implements the Expectation-Maximization (EM) algorithm (Moon 1996) to allot and regularly update pseudo-labels for phrase-region pairings, contributing to contrastive learning.

(a) a backpacker (b) bride just married

Figure 6: Visualization of pseudo-labels (left) on the Flickr30k and prediction (right) on the RefCOCO+.

False Negative Detection The effectiveness of contrastive learning is limited by negative example sampling, making false negative detection crucial for self-supervised representation learning and cross-modal retrieval. Huynh et al. (2022) applied strategies to filter similar images and reduce false negatives in visual representation learning. In weakly supervised visual-audio tasks, Sun et al. (2023) utilized a uni-modal similarity matrix to mitigate the influence of false negatives and enhance true negatives to improve visual-audio alignment. Some works (Li et al. 2023, 2024) attempt to improve the image-text retrieval performance by correcting false negatives and selecting negative examples.

Conclusion

We introduce a novel method called Momentum Pseudo Labeling (MPL) that leverages a momentum model to compute pseudo-labels. Building on this foundation, we explore and model the relationships between phrases and regions in both matching and non-matching image-caption pairs. Empirical experiments demonstrate that our MPL method provides more effective guidance during the training stages.

Acknowledgments

This work was supported by the National Science and Technology Major Project under Grant 2022ZD0120202, in part by the National Natural Science Foundation of China (No. U23B2056 and No. 62306026), in part by China Postdoctoral Science Foundation (No. 2023M740184), in part by the Fundamental Research Funds for the Central Universities, and in part by the State Key Laboratory of Complex & Critical Software Environment.

References

Chen, C.; Anjum, S.; and Gurari, D. 2022. Grounding answers for visual questions asked by visually impaired people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19098–19107.

Chen, K.; Zhang, R.; Mensah, S.; and Mao, Y. 2022. Contrastive Learning with Expectation-Maximization for Weakly Supervised Phrase Grounding. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 8549–8559.

Dou, Z.-Y.; Peng, N.; et. al; and et. al. 2021. Improving pre-trained vision-and-language embeddings for phrase grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6362–6371.

Gupta, T.; Vahdat, A.; Chechik, G.; Yang, X.; Kautz, J.; and Hoiem, D. 2020. Contrastive learning for weakly supervised phrase grounding. In *European Conference on Computer Vision*, 752–768. Springer.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

Huang, B.; Lian, D.; Luo, W.; and Gao, S. 2021. Look Before You Leap: Learning Landmark Features for One-Stage Visual Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16888–16897.

Huynh, T.; Kornblith, S.; Walter, M. R.; Maire, M.; and Khademi, M. 2022. Boosting Contrastive Self-Supervised Learning with False Negative Cancellation. arXiv:2011.11765.

Jiang, H.; Lin, Y.; Han, D.; Song, S.; and Huang, G. 2022. Pseudo-q: Generating pseudo language queries for visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15513–15523.

Jin, L.; Luo, G.; Zhou, Y.; Sun, X.; Jiang, G.; Shu, A.; and Ji, R. 2023. Refclip: A universal teacher for weakly supervised referring expression comprehension. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2681–2690.

Kazemzadeh, S.; Ordonez, V.; Matten, M.; and Berg, T. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 787–798.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123: 32–73.

Li, H.; Bin, Y.; Liao, J.; Yang, Y.; and Shen, H. T. 2023. Your negative may not be true negative: Boosting image-text matching with false negative elimination. In *Proceedings of the 31st ACM International Conference on Multimedia*, 924–934.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705.

Li, J.; Zhou, P.; Xiong, C.; and Hoi, S. C. 2020. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Li*, L. H.; Zhang*, P.; Zhang*, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; Chang, K.-W.; and Gao, J. 2022. Grounded Language-Image Pre-training. In *CVPR*.

Li, Z.; Guo, C.; Feng, Z.; Hwang, J.-N.; and Du, Z. 2024. Integrating Language Guidance Into Image-Text Matching for Correcting False Negatives. *IEEE Transactions on Multimedia*, 26: 103–116.

Lin, P.; Yu, Z.; Lu, M.; Feng, F.; Li, R.; and Wang, X. 2024. Visual Prompt Tuning for Weakly Supervised Phrase Grounding. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7895–7899. IEEE.

Liu, J.; and Hockenmaier, J. 2019. Phrase Grounding by Soft-Label Chain Conditional Random Field. arXiv:1909.00301.

Liu, X.; Li, L.; Wang, S.; Zha, Z.-J.; Meng, D.; and Huang, Q. 2019. Adaptive Reconstruction Network for Weakly Supervised Referring Expression Grounding. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2611–2620.

Liu, Y.; Zhang, J.; Chen, Q.; and Peng, Y. 2023. Confidence-aware Pseudo-label Learning for Weakly Supervised Visual Grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2828–2838.

Moon, T. K. 1996. The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6): 47–60.

Pennington, J.; Socher, R.; Manning, C. D.; et al. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2016. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6): 1137–1149.

Rigoni, D.; Parolari, L.; Serafini, L.; Sperduti, A.; and Ballan, L. 2023. Weakly-Supervised Visual-Textual Grounding with Semantic Prior Refinement. arXiv:2305.10913.

Shaharabany, T.; Tewel, Y.; and Wolf, L. 2022. What is where by looking: Weakly-supervised open-world phrase-grounding without text inputs. *Advances in Neural Information Processing Systems*, 35: 28222–28237.

Shaharabany, T.; and Wolf, L. 2023. Similarity maps for self-training weakly-supervised phrase grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6925–6934.

Sun, W.; Zhang, J.; Wang, J.; Liu, Z.; Zhong, Y.; Feng, T.; Guo, Y.; Zhang, Y.; and Barnes, N. 2023. Learning audio-visual source localization via false negative aware contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6420–6429.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, L.; Huang, J.; Li, Y.; Xu, K.; Yang, Z.; and Yu, D. 2021. Improving Weakly Supervised Visual Grounding by Contrastive Knowledge Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14090–14100.

Wang, Q.; Tan, H.; Shen, S.; Mahoney, M. W.; and Yao, Z. 2020. Maf: Multimodal alignment framework for weakly-supervised phrase grounding. *arXiv preprint arXiv:2010.05379*.

Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, 69–85. Springer.

Zhang, H.; Zhang, P.; Hu, X.; Chen, Y.-C.; Li, L.; Dai, X.; Wang, L.; Yuan, L.; Hwang, J.-N.; and Gao, J. 2022. Glipv2: Unifying localization and vision-language understanding. *Advances in Neural Information Processing Systems*, 35: 36067–36080.

Zhang, R.; Wang, C.; and Liu, C.-L. 2023. Cycle-Consistent Weakly Supervised Visual Grounding With Individual and Contextual Representations. *IEEE Transactions on Image Processing*.

Zhang, Z.; Zhao, Z.; Lin, Z.; He, X.; et al. 2020. Counterfactual contrastive learning for weakly-supervised vision-language grounding. *Advances in Neural Information Processing Systems*, 33: 18123–18134.

Zhao, Z.; Guo, L.; He, X.; Shao, S.; Yuan, Z.; and Liu, J. 2023. MAMO: Fine-Grained Vision-Language Representations Learning with Masked Multimodal Modeling. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1528–1538.