Diffusion-Assisted Progressive Learning for Weakly Supervised Phrase Localization

Pengyue Lin, Yanyang Hu, Xinjing Liu, Wenqi Jia, Fangxiang Feng, Ruifan Li*

School of Artificial Intelligence, Beijing University of Posts and Telecommunications
{linpengyue, yyh, liuxj_ai, jiawenqi, fxfeng, rfli}@bupt.edu.cn

Abstract

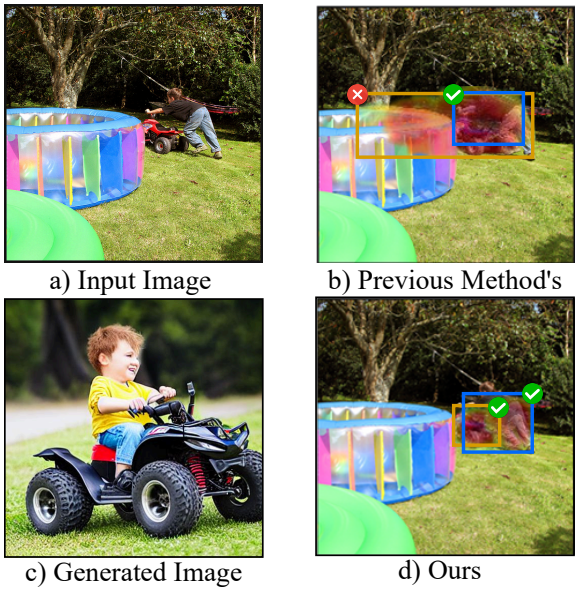
Weakly supervised phrase localization (WSPL) aims to localize visual objects mentioned by given phrases, but it learns without human-annotated bounding boxes. Previous works struggle in multi-object scenarios where objects in the background often appear simultaneously with the target objects. To this end, we propose a Diffusion-Assisted PrOgressive learning framework (i.e., DAPO) for WSPL task in this paper. Specifically, we score the difficulty of training samples based on the quantity of objects and the level of semantic alignment. These samples are then used progressively during training, in an order by their difficulty scores. To address the sample imbalance problem, we propose a Generation-Assisted Tuning method for the grounding network. First, to enrich the samples from few-object scenarios, we leverage Stable Diffusion (SD) to generate images with phrases. Second, we introduce an attention-driven scheme to direct SD's attention on the mentioned objects. Finally, we design a diffusion-guided loss, which helps the grounding network learn the objects' layouts. Extensive experiments show that our DAPO framework outperforms the strong baselines on benchmark datasets.

Code — https://github.com/LinPengyue/DAPO

1 Introduction

Given an image and a phrase, weakly supervised phrase localization (WSPL) aims to localize the visual objects mentioned in the phrase. An example is shown in Figure 1a). This task does not require any bounding box annotation during the model's training, which saves human labor for manual annotation. WSPL connects vision with language, which contributes to various downstream tasks, such as visual question answering (Xiao et al. 2024; Peng et al. 2024; You et al. 2024) and vision-language navigation (Barthel et al. 2019; Wu et al. 2022; Eftekhar et al. 2024).

Generally, we group WSPL models into two categories, detector-based models (Liu et al. 2021; Wang et al. 2024) and auxiliary-task-based models (Shaharabany and Wolf 2023; Zeng et al. 2024). The former group extracts regional proposals from pretrained object detectors, and then ranks them based on object-phrase similarity measures. The latter group designs some auxiliary vision-language tasks with a finer-grained understanding of object categories.



a) Input Image b) Previous Method's c) Generated Image d) Ours

Figure 1: The WSPL task and general results' comparison. a) For WSPL task, an image with a phrase is given without bounding boxes when the model is trained. Here, the phrase is "a young boy playing his toy quad". b) Previous methods correctly localize *boy*, but fail for *toy quad*. c) An image generated by SD clearly shows the two objects. d) Our model's result is consistent with ground truth. Here, *boy* is enclosed in a blue bounding box and *toy quad* in an orange one.

However, the two groups of methods face challenges in multi-object scenarios. Here, objects in the background often simultaneously appear with targeted objects. Thus, objects in the background could be strong yet misleading candidates. If models choose misleading candidates, incorrect correlations between vision and language could emerge. This issue is even more pronounced under weak supervision, where precise localization annotations are absent.

To this end, we need to consider the following two questions. **First**, how to establish precise entity-object semantic alignment in multi-object scenarios? Take Figure 1b) as an

*Corresponding author

Subset	S1	S2	S3	S4	S5
# Object	2	3	4	5	6–8
# Entity	2	≤3	≤4	≤5	≤6–8
# Image	917	5,277	8,147	7,286	8,027
# Phrase	4,585	26,385	40,735	36,430	40,135

Table 1: Statistics of training sample in Flickr30K Entities (subsets S1–S5).

example; for the entity *toy quad*, the objects *toy pool* and *toy ball* create visual ambiguity, making it difficult to identify the correct referent. In contrast, when such objects in the background are absent, the semantic alignment becomes significantly clearer (See Figure 1c)). This observation suggests that semantic alignment learned in simpler scenarios could be leveraged in more complex settings. Inspired by the progressive nature of human learning, a natural solution is to begin with simple samples and gradually introduce more challenging ones. To this end, we need to assess the difficulty of training samples and schedule their incorporation accordingly. **Second**, how to balance the sample distribution to support progressive learning? We observe that the training dataset exhibits an imbalanced distribution (See Table 1). Images with multi-object scenarios dominate the dataset, whereas few-object ones are significantly underrepresented. This sample imbalance problem hinders progressive learning. Most entity-relevant objects rarely appear in few-object scenarios, limiting opportunities to establish semantic alignment early in training. Thus, samples with few-object scenarios should be effectively enriched. In short, we need to resample phrases describing few-object scenarios, and generate images conditioned on these phrases.

In this paper, we propose a Diffusion-Assisted PrOgressive learning framework (i.e., DAPO), to teach the grounding network to locate target objects in multi-object scenarios. **Firstly**, we score the difficulty of training samples by quantifying the number of objects in each sample. We then progressively incorporate samples in ascending order of object count. To refine the ordering when quantity-based scores are identical, we integrate semantic-based scores to measure the level of semantic alignment. Furthermore, we design a self-paced adapting method with semantic regularization to dramatically select well-aligned samples. **Secondly**, we propose a Generation-Assisted Tuning (GAT) method to address the sample imbalance problem. Specifically, we leverage Stable Diffusion (SD) (Rombach et al. 2022) to generate images conditioned on the phrases, which describe few-object scenarios. However, we observe that SD often fails to faithfully render the mentioned objects due to missing entity semantics. To mitigate this, we propose an attention-driven scheme to enhance the attention value correlated with entity tokens. Furthermore, we design a diffusion-guided loss, directly leveraging diffusion attention as weak supervision for localization. It guides the grounding network to learn the object's layouts from the generated images.

Our main contributions are summarized as follows.

- We propose a novel framework, DAPO for WSPL task. DAPO scores the difficulty of training samples and uses them progressively in an order of difficulty scores.
- We design a GAT method to fine-tune the grounding network with SD. We design an attention-driven scheme to enforce SD's attention on the mentioned objects. We design a diffusion-guided loss to learn the object's layouts.
- We conduct extensive experiments on five WSPL benchmark datasets to demonstrate the effectiveness of our proposed DAPO framework.

2 Preliminary

Diffusion Model. Denoising diffusion probabilistic models (Ho, Jain, and Abbeel 2020) learn the desired data distribution by defining a Markov chain of length T. In the forward pass, this chain gradually adds noise to a given data sample x_0 to obtain a sequence of noisy samples x_t, $t \in T$. In the reverse process, a model ϵ_θ parameterized by θ is learned to predict the noise added for each step t. Specifically, SD as a latent diffusion model applies the denoising diffusion process to the latent representation z of x in the latent space of a variational auto-encoder. Its learning objective predicts the added noise at each time step t as:

$$L_{LDM} = \mathbb{E}_{\Phi(x),\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t)\|^2\right], \quad (1)$$

where z_t represents the noised latent representation at time step t. During inference, the reverse process starts with a random noise $x_T \sim \mathcal{N}(0, I)$ and gradually generates an image sample by evolving the noise from step T to 0.

Cross-Attention in SD. SD introduces text guidance via a cross-modal mechanism. The denoising UNet network in the latent space of SD consists of self-attention layers followed by cross-attention layers at resolutions $C \in \{64, 32, 16, 8\}$. Given a text prompt P composed of M tokens, a textual vector $\Phi_{txt}(P)$ is then obtained via CLIP text encoder Φ_{txt}. $\Phi_{txt}(P)$ is then mapped to intermediate feature maps of the SD model $\epsilon(\theta)$ through each cross-attention layer,

$$A = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \quad (2)$$

where A, Q, K, V denote the attention, query, key, and value matrices. d represents the dimension of Q and K.

3 Problem Formulation

The task of WSPL aims to ground the entities of a textual phrase in an associated image, while the correspondences between entities and image regions are not available for training. Given an input image I and a phrase P, we utilize entity annotations (Plummer et al. 2015) and extract noun chunks from the phrase, obtaining a group of entities $\mathcal{E} = \{e_k\}_{k=1}^K$. A grounding network G learns to predict the corresponding heatmaps $\mathcal{H} = \{h_k\}_{k=1}^K$ for these entities. Here, the heatmaps serve as a bridge to obtain the bounding boxes $\mathcal{B} = \{b_k\}_{k=1}^K$.

4 Methodology

Our DAPO framework is shown in Figure 2. First, GAT helps to build object-entity semantic alignment in few-object scenarios. Second, we assign difficulty scores for images, and a training scheduler is used for progressively learning.

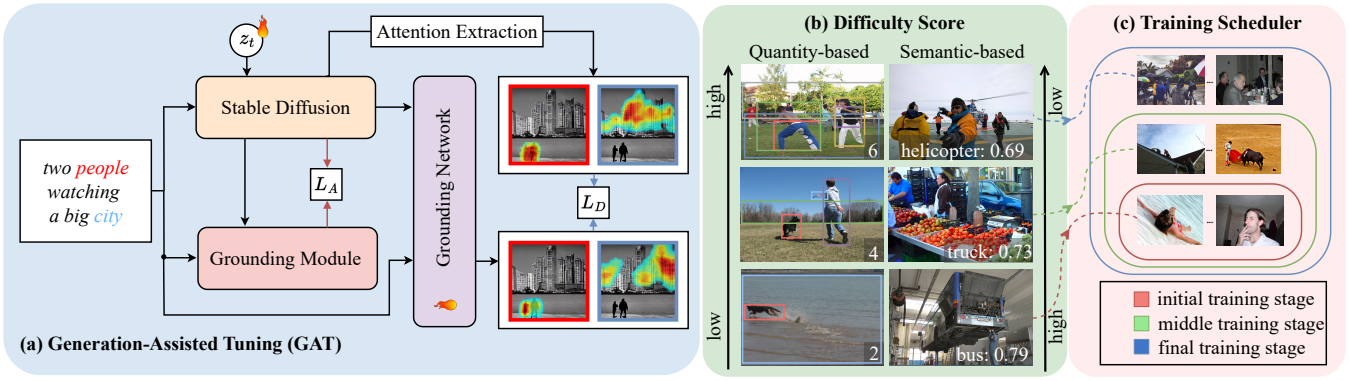


Figure 2: The overview of our DAPO framework. a) GAT uses SD to generate an image by the phrase. With a pretrained grounding module, SD generates mentioned objects without missing entity semantics. GAT also fine-tunes the grounding network with the phrase, the generated image, and the extracted attention maps. b) We apply scoring methods based on both quantity and semantic to all of the images. c) We incorporate training samples progressively in an order by their difficulty scores.

4.1 Generation-Assisted Tuning (GAT)

In this section, we design GAT to fine-tune the grounding network. Specifically, we resample phrases describing few-object scenarios. Then, we leverage a pre-trained SD to generate images conditioned on these phrases. Due to the challenging nature of generating all mentioned objects, some desired entity semantics are often neglected.

To overcome this challenge, we introduce an attention-driven scheme to enforce the generation for the mentioned objects. **Firstly**, we aggregate the attention maps from SD for subsequent optimization. As shown in Sec. 2, SD introduces text guidance using a cross-attention mechanism. Given a phrase P composed of K entities, at each inference step t, we obtain a collection of attention maps $\mathcal{A}_t = \{(a_t)_1, ..., (a_t)_K\}$, where $a_t \in \mathbb{R}^{C\times C}$. These attention maps capture the cross-correlation with tokens in entities $\mathcal{E} = \{e_k\}_{k=1}^K$, where $(a_t)_k$ represents the probability assigned to the k-th entity. Following Chefer et al. (2023), we average all the attention maps. **Secondly**, during the generation process, at each time step t, we optimize the intermediate latent representation z_t by improving the attention value correlated with entity tokens. We then compute the gradient update for z_t by applying a loss function to a_t. Specifically, given the latent representation z_t, we obtain the decoded generation result $x_t = \mathcal{D}(z_t)$. Then, we utilize a pretrained grounding module to predict heatmaps $\mathcal{H} = \{(h_t)_k\}_{k=1}^K$ from entities $\mathcal{E}$ and x_t. We define the loss function with mean squared error (MSE) as follows,

$$L_A = \frac{1}{|\mathcal{K}|} \sum_{(i,j)\in\mathcal{K}} \left[\|(h_t)_i - (a_t)_i\|_F^2 - \|(a_t)_i - (a_t)_j\|_F^2 \right], \tag{3}$$

where $\mathcal{K} = \{(i,j) \mid i,j \in \{1,\dots,K\}, i \neq j\}$ denotes the set of index pairs. K means the number of entities associated with the current phrase. $|\mathcal{K}| = K(K-1)$ denotes the total number of such pairs. $\|\cdot\|_F$ represents the Frobenius norm of a matrix. This loss improves the attention value by the guidance of heatmaps. It also encourages the attention regions of two entities to separate. **Finally**, we optimize the current latent encoding z_t in the following way,

$$z_t' \longleftarrow z_t - \alpha_t \bigtriangledown_{z_t} L_A, \tag{4}$$

where α_t is a scalar defining the step size of the gradient update. We perform another forward pass through SD using z_t', to compute z_{t-1} for the next denoising step. This update process is repeated for a subset of time steps $t = T, ..., 2, 1$.

Training Objective. To help the grounding network G learn object layouts of generated images, we consider cross-attention maps from SD as training targets. At the final denoising step $t = 0$, we extract these maps corresponding to entity tokens, $\mathcal{A} = \{a_k\}_{k=1}^K$. Given the predicted heatmap h_k and the cross-attention map a_k for the k-th entity, we define a diffusion-guided loss as follows,

$$L_D = \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{K} \|h_k^{(n)} - a_k^{(n)}\|_F^2, \tag{5}$$

where N denotes the total number of phrases. K means the number of entities per phrase. n indicates the corresponding map for the n-th phrase. This loss function leverages extracted attention maps referred to by text inputs, thereby applying only weakly supervised guidance to the heatmaps.

Inspired by prior WSPL works (Shaharabany, Tewel, and Wolf 2022; Lin et al. 2024a,b), we extract attention maps from VLMs (i.e., pretrained Vision-Language Models) based on real images. We also follow their task losses L_{task}, respectively. Thus, to train the grounding network, we design the total loss as follows,

$$\mathbb{L} = \mu L_{task} + (1-\mu) L_D \tag{6}$$

where the indicator $\mu \in \{0, 1\}$. It takes 1 for a real-world image and 0 for a generated one. Note that the total loss does not include MSE loss L_A (Eq. 3), because no gradients are propagated between SD and the grounding network.

4.2 Difficulty Scorer and Training Scheduler

In this section, we present our progressive learning details for WSPL task. This learning process includes two key components: a difficulty scorer and a training scheduler.

Difficulty Scorer. We propose two scoring methods based on the quantity of objects and the level of semantic alignment. **First**, we assign higher scores to training samples containing more visual objects, referred to as the quantity-based score (ScoreQ). Specifically, we consider that samples containing multiple objects and entities have inherently complex scenarios. Therefore, we define ScoreQ as the object count. As shown in Table 1, we divide the training dataset into difficulty-level subsets based on the number of objects per image. **Second**, we compute a semantic-based score (ScoreS) for each image-entity pair using CLIP (Radford et al. 2021), a vision-language semantic matching model. A higher ScoreS means stronger alignment between the entity's semantics and its corresponding visual object. In other words, precise semantic alignment benefits entity grounding. Both scoring techniques are applied consistently across all training samples to estimate their learning difficulty.

Training Scheduler. Inspired by Yang et al. (2022), we utilize a baby step training scheduler for ScoreQ. Specifically, the entire training dataset $\mathcal{D}$ is divided into different subsets, i.e., $\{\mathcal{D}_1, ..., \mathcal{D}_s\}$. Those samples with the same difficulty score are categorized into the same subset. The training runs from the easiest subset. After convergence, the next subset is merged into the current training set. Finally, all the subsets are merged and used. Note that when multiple samples have the same quantity-based scores, we incorporate the semantic-based score as a difficulty metric to select samples. However, the hard sample selection based solely on CLIP is suboptimal. This is because the semantic-based scores of training samples are close (e.g., 0.73 vs. 0.78), leading to an over-sensitivity to predefined hyperparameter.

To address this problem, we propose a self-paced adapting method with semantic regularization. We adjust the sample selection to be dynamic rather than rule-based. Specifically, in an iteration for each training sample (indexed by s), a latent weight variable $v_s \in [0, 1]$ is assigned for the training loss $\mathbb{L}_s(\theta)$. It indicates whether such a sample is fit to optimize the grounding network G. The parameter θ belongs to G network. To find the proper θ^* and the weight v^*, our objective is given as follows,

$$\theta^*, v^* = \underset{v,\theta}{\operatorname{argmin}} \sum_{s=1}^{N} v_s \mathbb{L}_s(\theta) + f(v), \tag{7}$$

where N is the total number of samples. $f(v)$ is the self-paced regularizer, which is designed as follows,

$$f(v) = \frac{1}{\gamma} \sum_{s=1}^{N} \left(\frac{1}{2} v_s^2 - v_s \cdot \mathrm{SREG_s} \right) \tag{8}$$

where γ ($\gamma > 0$) is a hyperparameter that increases iteratively in order to estimate the parameters of G via self-paced learning. SREG_s represents the normalized ScoreS of the s-th sample. Note that smaller values of γ require lower v_s, thus focusing on easier samples initially, while harder samples are gradually included as γ increases. Furthermore, we do not adopt a discrete binary assignment for v (i.e., either 0 or 1), but instead use a continuous and learnable weight. This soft weighting strategy avoids overly rigid sample selection, enabling better adaptation to complex training scenarios.

5 Experiment and Analysis

5.1 Datasets

Our DAPO follows a pretraining and fine-tuning paradigm. To pretrain our grounding network, we follow the widely adopted setup in MG (Akbari et al. 2019). Specifically, we use either the MSCOCO (Lin et al. 2014) or Visual Genome (VG) (Krishna et al. 2017) training splits for pretraining. During fine-tuning, we build (image, phrase, entity) triplets from the training split of Flickr30K Entities (Plummer et al. 2015). Each image is on average associated with five phrases, and each phrase corresponds to 2$\sim$8 entities.

We evaluate the performance of WSPL on five benchmark datasets: VG, Flickr30K Entities, ReferIt (Grubinger et al. 2006; Chen, Kovvuri, and Nevatia 2017), as well as two constructed subsets, i.e., VG-M and Flickr30K-M. The first three benchmarks follow the test splits defined in Akbari et al. (2019) for consistency. The other subsets, VG-M and Flickr-M, are specifically designed for multi-object scenarios. These subsets are derived from the original test sets of VG and Flickr30K Entities, respectively.

5.2 Baselines and Metrics

We compare our DAPO framework with state-of-the-art WSPL **baselines**. The first group includes task-specific models, such as MG (Akbari et al. 2019), Gbs (Arbelle et al. 2021), WWbL-g (Shaharabany, Tewel, and Wolf 2022), WWbL-g++ (Shaharabany and Wolf 2023), BBR (Gomel, Shaharbany, and Wolf 2023), TAS (Lin et al. 2024a), and VPT (Lin et al. 2024b). The second group includes foundation models, such as SelfEQ (He et al. 2024a), APR (Zeng et al. 2024), SynGround (He et al. 2024b), and HIST (Luo et al. 2025). Moreover, we use two **metrics**, the "pointing game" accuracy (Cinbis, Verbeek, and Schmid 2016) and the bounding box accuracy (Shaharabany, Tewel, and Wolf 2022). "Pointing game" accuracy measures the percentage of predicted maximum points of the heatmap that lie within the bounding box ground truth. Bounding box accuracy measures the percentage of heatmap bounding boxes that have an IoU greater than 1/2.

5.3 Implementation Details

In our DAPO framework, we select three representative grounding networks as our backbones, including WWbL-g, TAS and VPT. We use SGD optimizer with a batch size of 32 and an initial learning rate of 0.012. The optimizer momentum is 0.9 and the weight decay is 0.0001. We increase ScoreQ from 2 to 8 (See Table 1), and introduce different subsets into the training dataset. Based on ScoreS, we initialize the hyperparameter γ to 0.5 at the beginning of training. As training proceeds, γ is linearly decayed to 0.1. We resample 10^4 phrases in total, where bi-entity phrases are 7,000 and tri-entity phrases are 3,000. For SD, we utilize the DDPM scheduler for 50 runs, and generate images conditioned on these resampled phrases. Grounding networks are fine-tuned on one A6000 GPU, with training on each subset lasting 12 epochs. It takes approximately 30 hours.

Model	VG Pre-trained						MSCOCO Pre-trained					
	Point Acc.			Bbox Acc.			Point Acc.			Bbox Acc.		
	VG	Flickr	ReferIt	VG	Flickr	ReferIt	VG	Flickr	ReferIt	VG	Flickr	ReferIt
APR	75.04	84.49	69.26	–	–	–	–	–	–	–	–	–
SelfEQ	–	81.90	67.40	–	–	–	–	84.07	62.75	–	–	–
SynGround	–	80.73	–	–	–	–	–	–	–	–	–	–
HIST	–	83.60	69.50	–	–	–	–	85.30	63.40	–	–	–
MG	48.76	60.08	60.01	14.45	27.78	18.85	47.94	61.66	47.52	15.77	27.06	15.15
Gbs	53.40	70.48	59.44	–	–	–	52.00	72.60	56.10	–	–	–
WWbL-g++	66.63	79.95	70.25	30.95	45.56	38.74	62.96	78.10	61.53	29.14	46.62	32.43
BBR	63.51	78.32	67.33	31.02	42.40	35.56	60.05	77.19	63.48	28.77	47.26	30.63
WWbL-g	62.31	75.63	65.95	27.26	36.35	32.25	59.09	75.43	61.03	27.22	35.75	30.08
TAS	58.07	76.69	70.86	27.31	45.63	35.70	60.31	77.85	62.63	29.58	45.46	33.41
VPT	62.72	80.03	68.21	27.40	45.60	34.76	60.74	81.15	64.14	27.65	45.09	31.14
WWbL-g†	62.17	75.92	66.07	25.23	36.92	32.71	59.33	75.66	61.85	26.73	35.88	30.22
DAPO-g	65.17	78.61	68.04	31.46	42.89	35.53	62.45	79.12	63.53	30.11	38.72	32.26
TAS†	56.21	76.81	70.93	27.58	45.80	35.68	59.55	77.92	62.81	29.03	45.53	33.48
DAPO-TAS	62.25	79.63	72.98	32.08	48.82	38.96	63.18	80.22	64.25	30.97	47.66	34.13
VPT†	61.08	80.17	68.55	27.51	45.71	34.89	60.82	81.25	64.33	27.72	45.14	31.29
DAPO-VPT	66.87	85.43	72.11	30.79	48.25	37.56	63.60	84.35	66.90	31.63	48.57	33.68

Table 2: WSPL results on the test set: "pointing game" and bounding box accuracy. Results marked with † are fine-tuned on Flickr30K Entities training data.

5.4 Main Results

In this section, we compare our method with other WSPL methods. To keep fair and effective of progressive learning, we choose to fine-tune three baselines on the Flickr30K Entities training set. The experimental results are reported in Table 2. Our approach works with different pre-training data (i.e., VG and MSCOCO) and the testing data (i.e., VG, Flickr30K Entities, and ReferIt). We observe a consistent enhancement in the grounding performance across all compared methods. It demonstrates that our approach has seamless compatibility with existing WSPL approaches. DAPO-VPT improves point accuracies on Flickr30K from 61.08%, 80.17% and 68.55% to 66.87%, 85.43% and 72.11%, respectively. The results outperform other classical WSPL methods. This shows the superiority of our framework in grounding entities in multi-object scenarios. Also, APR and HIST achieve high accuracies on some datasets. In fact, these two methods are based on continual training on ALBEF (Li et al. 2021). Note that they aim to locate the key regions of target objects, instead of delineating their full extent with bounding boxes.

Furthermore, we evaluate DAPO with two built datasets, VG-M and Flickr-M. Both datasets contain only images with multiple objects. Here, we consider the number of objects larger than 16 and 6, respectively. Table 3 reports the experimental results. These results demonstrate that our approach outperforms current SOTA WSPL methods on these two datasets. Compared to WWbL-g, our DAPO framework improves the bounding box accuracy by 8.08% and 6.35% on VG-M. In addition, our DAPO achieves the best point accuracy of 54.50%. On the other hand, our DAPO achieves a significant improvement on Flickr-M. Compared to other methods, the point accuracy and bounding box accuracy of DAPO improved by $5.25 \sim 6.71\%$ and $2.04 \sim 6.77\%$, respectively. These results demonstrate that our method is effective in multi-object scenarios.

	Model	Point Acc.		Box Acc.	
		VG-M	Flickr-M	VG-M	Flickr-M
VG	WWbL-g†	42.05	70.10	17.85	29.75
	DAPO-g	51.45	76.33	25.93	36.47
	TAS†	41.02	68.77	17.35	34.15
	DAPO-TAS	52.61	75.36	24.65	36.44
	VPT†	43.28	70.30	19.92	36.13
	DAPO-VPT	54.50	75.55	26.22	38.17
COCO	WWbL-g†	42.20	71.30	18.03	29.93
	DAPO-g	53.58	77.84	24.38	36.70
	TAS†	41.12	71.71	17.45	32.14
	DAPO-TAS	52.86	78.08	24.83	39.32
	VPT†	42.71	71.30	20.08	36.69
	DAPO-VPT	53.73	78.01	26.40	39.04

Table 3: Performance of DAPO vs. previous models on two datasets, VG-M and Flickr-M.

5.5 Ablation Study

In this section, we empirically investigate how the performance of our framework is affected by different model settings. All ablation experiments are based on DAPO-TAS.

Resampled Phrases. We show the performance of our framework with different numbers of resampled phrases.

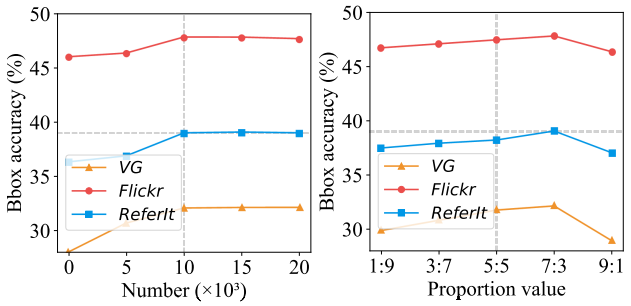


Figure 3: The impact of different resampled numbers and proportion values on the effectiveness of DAPO-TAS.

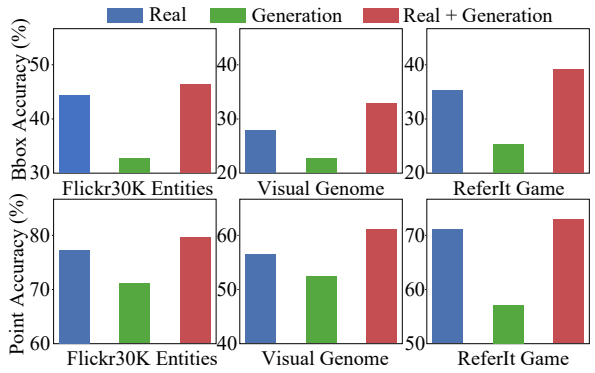


Figure 4: Performance with different data sources.

The experimental results are reported in Figure 3. As the number of resampled phrases increases, the model achieves improved performance, demonstrating the benefit of resampling for semantic alignment. However, once the number of phrases reaches 10^4, the performance gains become marginal. It indicates that the semantic alignment learned from few-object scenarios fails to generalize effectively to multi-object settings. Furthermore, we investigate the impact of the resampled phrase ratio on grounding performance. With the total number of resampled phrases fixed, we define five ratios of bi-entity to tri-entity phrases in the training dataset. Results show that the model achieves optimal performance when the ratio is 7:3, under which the image count in each subset reaches 8,000. A balanced distribution of samples appears to benefit progressive learning.

Generated Images. We further investigate the impact of incorporating generated images into training. Specifically, we evaluate three settings: only real images, only generated images, and a mix of both. The results are shown in Figure 4. This shows that our DAPO achieves gains by mixing generated and real images. This is because generated images effectively balance the training distribution. However, relying solely on generated images yields the worst results. Due to a domain shift between generated and real images, the model struggles to generalize to real-world scenarios.

Scoring Method. We investigate the effect of difficulty scoring methods. The first one does not use any scoring method. The second one only uses ScoreQ, denoted as QSM. The third one uses ScoreS, denoted as SSM. The fourth merges the aforementioned two methods. The experiments are reported in Table 4. Compared to QSM and SSM, our framework achieves the best performance using both methods. Actually, weak supervision makes it difficult to select appropriate learnable samples at each step. Therefore, relying solely on a single scoring method may result in misalignment between objects and entities due to incomplete or biased difficulty estimation.

QSM	SSM	Point Acc.			Bbox Acc.		
		VG	*Flickr*	*ReferIt*	*VG*	*Flickr*	*ReferIt*
✗	✗	56.44	77.03	71.15	27.09	46.00	35.92
✓	✗	56.67	77.25	71.36	28.02	46.23	36.15
✗	✓	59.41	78.52	72.13	30.76	47.54	37.83
✓	✓	**62.25**	**79.63**	**72.98**	**32.08**	**48.82**	**38.96**

Table 4: The performance of our DAPO-TAS using various scoring methods. QSM means quantity-based scoring method, while SSM means semantic-based scoring method.

L_A	L_D	L_{task}	Test Point Accuracy			Test Bbox Accuracy		
			VG	*Flickr*	*ReferIt*	*VG*	*Flickr*	*ReferIt*
✗	✗	✓	58.22	76.54	70.71	27.46	45.78	35.55
✗	✓	✗	48.31	65.41	53.06	18.88	28.65	20.11
✗	✓	✓	58.35	76.67	70.83	27.59	45.91	35.68
✓	✗	✓	52.65	71.23	57.91	22.84	32.76	25.02
✓	✓	✗	61.73	78.11	71.46	31.57	47.30	38.04
✓	✓	✓	**62.25**	**79.63**	**72.98**	**32.08**	**48.82**	**38.96**

Table 5: The performance of our DAPO-TAS framework using various loss terms.

Loss. We evaluate the contributions of three loss terms. Specifically, three losses include attention-driven loss L_A, diffusion-guided loss L_D and original task loss L_{task}. As shown in Table 5, each loss term comprehensively enhances the performance. Using only the task loss, DAPO-TAS achieves 76.54% "pointing game" accuracy over the Flickr30K test set. Adding attention-driven loss improves accuracy to 78.11%, suggesting that attention-driven loss enhances the semantic alignment between language and vision. This loss ensures that the referred objects exist within generated images. This is a key assumption in the phrase localization setting. Further incorporating diffusion-guided loss increases the accuracy to 79.63%, demonstrating its effectiveness in regularizing object layout. This is because SD is capable of modeling the relation between entities when generating objects. In contrast, VLMs have limited capability in capturing such fine-grained relational semantics, leading to incorrect localization for similarly described entities.

5.6 Qualitative Analysis

Figure 5 visualizes our grounding results from Flickr30K Entities. Phrase in (a) requires the localization of the three people except the guard, demonstrating our framework's ability to distinguish target objects. In (b), the phrase calls

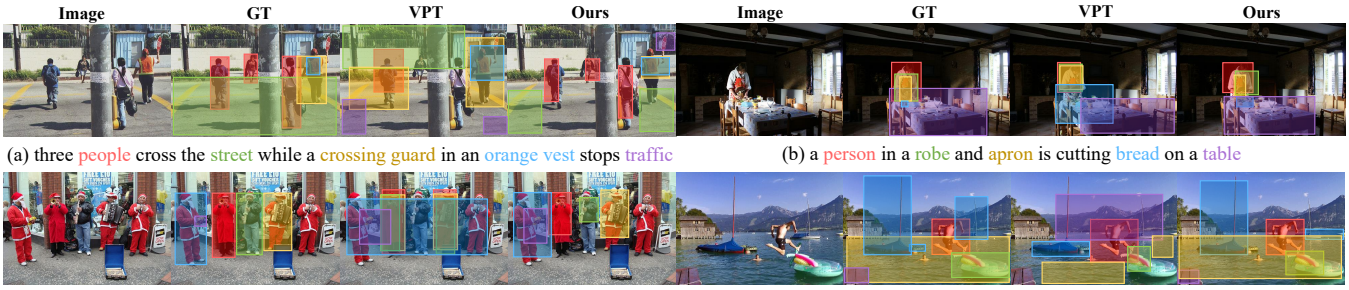


Figure 5: Qualitative results of our DAPO-VPT and their comparison to those of VPT.

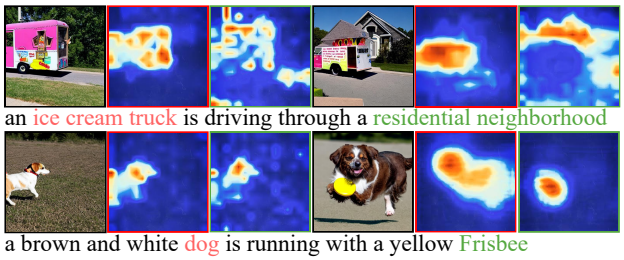


Figure 6: Examples of generated images and attention maps. The left is produced by original SD. The right is ours.

for the localization of three closely situated objects: a person, an apron, and a robe. The results demonstrate our fine-grained semantic alignment capability. In (c), the model successfully distinguishes among three players wearing similar clothing, based on subtle differences in their textual descriptions. In (d), we show that our method retrieves object positions based on the overall semantics of the phrase, rather than relying solely on the noun keyword *ramp*. It reflects our model's capacity for deeper grounding reasoning.

In addition, we compare our generated images and attention maps with those of SD. The results are visualized in Figure 6. Our generated images are more related to semantics of the textual phrases. For instance, our method correctly generates *neighborhood* and *Frisbee*. In contrast, SD tends to ignore both objects in the images. Furthermore, our attention maps could capture their positions, which are highlighted in red. Leveraging these positions, the objects mentioned in phrases, such as *dog* and *ice cream truck*, can be accurately identified during progressive learning.

6 Related Work

Weakly Supervised Phrase Localization (WSPL). WSPL models need to solely learn from image-phrase pairs during training. To address the challenge, detector-based works (Chen, Gao, and Nevatia 2018; Datta et al. 2019; Wang and Specia 2019; Gupta et al. 2020; Wang et al. 2021; Liu et al. 2021; Chen et al. 2022; Kuang et al. 2025) use object detectors and choose the correct proposals that are highly related to the corresponding phrases. Moreover, other methods (Zhang et al. 2018; Akbari et al. 2019; Arbelle et al. 2021) design auxiliary tasks for phrase localization, such as intra-modal classifications and inter-modal alignments. Recently, VLMs have been increasingly utilized for auxiliary-task-based WSPL. These methods involve either mixed-dataset pretrained VLMs (He et al. 2024a,b; Zeng et al. 2024; Huy et al. 2025; Luo et al. 2025) or fine-tuning task-specific networks supervised by VLMs (Shaharabany, Tewel, and Wolf 2022; Shaharabany and Wolf 2023; Gomel, Shaharbany, and Wolf 2023; Lin et al. 2024a,b). However, previous approaches remain limited in multi-object scenarios.

Progressive Learning for Grounding. Progressive learning refers to a learning paradigm where models are trained in a staged manner, starting with simpler tasks or data and progressively increasing the complexity. Recently, progressive learning has been introduced into vision-language grounding works (Lu et al. 2024; Yu and Li 2024; Wang et al. 2025; Xie et al. 2025), demonstrating its potential to improve models' capability. Xiao et al. (2023) iteratively select high-quality pseudo-labeled samples for training. The method updates model weights at each iteration to refine subsequent data selection. Garg, Kumar, and Rawat (2025) introduce a progressive learning framework with temporal and spatial curriculum modules to enhance spatio-temporal video grounding in weakly supervised settings. Le et al. (2025) leverage progressive multi-granular alignment to enhance compositional visual reasoning in VLMs. Unlike previous works, we leverage generated images to enable progressive learning from simple to complex, multi-object scenarios.

7 Conclusion and Future work

In this paper, we propose a novel DAPO framework for WSPL task. Specifically, we design a GAT method to fine-tune the grounding network with SD. It helps establish precise semantic alignment between entities and objects. To transfer learned alignment to multi-object scenarios, we incorporate training samples progressively, ordered by difficulty scores. Extensive experiments on classical and proposed benchmark datasets show the effectiveness of our framework, which outperforms several strong baselines.

In the future, we plan to leverage advanced text-to-image generation models for our framework, such as GPT-4o (OpenAI 2024) and Flux (Batifol et al. 2025). These generative models are expected to mitigate the domain shift between training and test distributions, thereby enhancing generalization in real-world settings.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2023YFC3305902, by the Special Project for Industrial Foundation Reconstruction and High-Quality Development of Manufacturing under Grant Agreement No. ZC25T320057/100, in part by the National Natural Science Foundation of China No. 62076032, in part by the China Computer Federation of Zhipu Foundation No. CCF-Zhipu202407, in part by Industry-University-Research Innovation Fund for Chinese Universities No. 2024MZ028, in part by BUPT Kunpeng and Ascend Center of Cultivation, and by BUPT innovation and entrepreneurship support program No. 2025-YC-A067.

References

Akbari, H.; Karaman, S.; Bhargava, S.; Chen, B.; Vondrick, C.; and Chang, S.-F. 2019. Multi-level multimodal common semantic space for image-phrase grounding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12476–12486. IEEE.

Arbelle, A.; Doveh, S.; Alfassy, A.; Shtok, J.; Lev, G.; Schwartz, E.; Kuehne, H.; Levi, H. B.; Sattigeri, P.; Panda, R.; et al. 2021. Detector-free weakly supervised grounding by separation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1801–1812.

Barthel, K. U.; Hezel, N.; Schall, K.; and Jung, K. 2019. Real-time visual navigation in huge image sets using similarity graphs. In *Proceedings of the 27th ACM International Conference on Multimedia*, 2202–2204.

Batifol, S.; Blattmann, A.; Boesel, F.; Consul, S.; Diagne, C.; Dockhorn, T.; English, J.; English, Z.; Esser, P.; Kulal, S.; et al. 2025. FLUX. 1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. *arXiv e-prints*, arXiv–2506.

Chefer, H.; Alaluf, Y.; Vinker, Y.; Wolf, L.; and Cohen-Or, D. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4): 1–10.

Chen, K.; Gao, J.; and Nevatia, R. 2018. Knowledge aided consistency for weakly supervised phrase grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4042–4050.

Chen, K.; Kovvuri, R.; and Nevatia, R. 2017. Query-guided regression network with context policy for phrase grounding. In *Proceedings of the IEEE International Conference on Computer Vision*, 824–832. IEEE.

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.

Cinbis, R. G.; Verbeek, J.; and Schmid, C. 2016. Weakly supervised object localization with multi-fold multiple instance learning. *IEEE transactions on pattern analysis and machine intelligence*, 39(1): 189–203.

Datta, S.; Sikka, K.; Roy, A.; Ahuja, K.; Parikh, D.; and Divakaran, A. 2019. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2601–2610.

Eftekhar, A.; Zeng, K.-H.; Duan, J.; Farhadi, A.; Kembhavi, A.; and Krishna, R. 2024. Selective Visual Representations Improve Convergence and Generalization for Embodied AI. In *The Twelfth International Conference on Learning Representations*.

Garg, A.; Kumar, A.; and Rawat, Y. S. 2025. Stpro: Spatial and temporal progressive learning for weakly supervised spatio-temporal grounding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 3384–3394.

Gomel, E.; Shaharbany, T.; and Wolf, L. 2023. Box-based Refinement for Weakly Supervised and Unsupervised Localization Tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16044–16054.

Grubinger, M.; Clough, P.; Müller, H.; and Deselaers, T. 2006. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *International workshop ontoImage*, volume 2, 13–23. Minori, Italy: IEEE.

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, R.; Cascante-Bonilla, P.; Yang, Z.; Berg, A. C.; and Ordonez, V. 2024a. Improved Visual Grounding through Self-Consistent Explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13095–13105.

He, R.; Yang, Z.; Cascante-Bonilla, P.; Berg, A. C.; and Ordonez, V. 2024b. Learning from Synthetic Data for Visual Grounding. *arXiv preprint arXiv:2403.13804*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Huy, T. D.; Huynh, D. A.; Xie, Y.; Qi, Y.; Chen, Q.; Nguyen, P. L.; Tran, S. K.; Phung, S. L.; van den Hengel, A.; Liao, Z.; et al. 2025. Seeing the Trees for the Forest: Rethinking Weakly-Supervised Medical Visual Grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

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.

Kuang, D.; Zhang, R.; Nie, Z.; Chen, J.; and Kim, J. 2025. Momentum Pseudo-Labeling for Weakly Supervised Phrase Grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 24348–24356.

Le, Q.-H.; Dang, L. H.; Le, N. H.; Tran, T.; and Le, T. M. 2025. Progressive multi-granular alignments for grounded reasoning in large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 4473–4481.

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.

Lin, P.; Li, R.; Ji, Y.; Yu, Z.; Feng, F.; Ma, Z.; and Wang, X. 2024a. Triple Alignment Strategies for Zero-shot Phrase Grounding under Weak Supervision. In *ACM Multimedia 2024*.

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

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference*, 740–755.

Liu, Y.; Wan, B.; Ma, L.; and He, X. 2021. Relation-aware instance refinement for weakly supervised visual grounding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5612–5621.

Lu, M.; Li, R.; Feng, F.; Ma, Z.; and Wang, X. 2024. Lgr-net: Language guided reasoning network for referring expression comprehension. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8): 7771–7784.

Luo, J.; Hossain, M. R. I.; Li, B.; and Sigal, L. 2025. Barking Up The Syntactic Tree: Enhancing VLM Training with Syntactic Losses. arXiv:2412.08110.

OpenAI. 2024. GPT-4o Technical Report. Technical report, OpenAI. Accessed: 2025-07-18.

Peng, Z.; Wang, W.; Dong, L.; Hao, Y.; Huang, S.; Ma, S.; Ye, Q.; and Wei, F. 2024. Grounding multimodal large language models to the world. In *The Twelfth International Conference on Learning Representations*.

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*, 2641–2649.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. Vienna, Austria: PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

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.

Wang, J.; and Specia, L. 2019. Phrase localization without paired training examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4663–4672.

Wang, J.; Wang, H.; Zhang, W.; Ji, K.; Huang, D.; and Zheng, Y. 2025. Progressive Language-guided Visual Learning for Multi-Task Visual Grounding. *arXiv preprint arXiv:2504.16145*.

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*, 14090–14100.

Wang, Z.; Yang, C.; Jiang, B.; and Yuan, J. 2024. A Dual Reinforcement Learning Framework for Weakly Supervised Phrase Grounding. *IEEE Transactions on Multimedia*, 26: 394–405.

Wu, S.; Fu, X.; Wu, F.; and Zha, Z.-J. 2022. Cross-modal semantic alignment pre-training for vision-and-language navigation. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4233–4241.

Xiao, J.; Yao, A.; Li, Y.; and Chua, T.-S. 2024. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13204–13214.

Xiao, L.; Yang, X.; Peng, F.; Yan, M.; Wang, Y.; and Xu, C. 2023. Clip-vg: Self-paced curriculum adapting of clip for visual grounding. *IEEE Transactions on Multimedia*, 26: 4334–4347.

Xie, M.; Wang, M.; Li, H.; Zhang, Y.; Tao, D.; and Yu, Z. 2025. Phrase decoupling cross-modal hierarchical matching and progressive position correction for visual grounding. *IEEE Transactions on Multimedia*.

Yang, L.; Shen, Y.; Mao, Y.; and Cai, L. 2022. Hybrid curriculum learning for emotion recognition in conversation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 11595–11603.

You, H.; Zhang, H.; Gan, Z.; Du, X.; Zhang, B.; Wang, Z.; Cao, L.; Chang, S.-F.; and Yang, Y. 2024. Ferret: Refer and Ground Anything Anywhere at Any Granularity. In *The 12th International Conference on Learning Representations*.

Yu, Z.; and Li, R. 2024. Revisiting Counterfactual Problems in Referring Expression Comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13438–13448.

Zeng, Y.; Huang, Y.; Zhang, J.; Jie, Z.; Chai, Z.; and Wang, L. 2024. Investigating Compositional Challenges in Vision-Language Models for Visual Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14141–14151.

Zhang, J.; Bargal, S. A.; Lin, Z.; Brandt, J.; Shen, X.; and Sclaroff, S. 2018. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10): 1084–1102.