Augment the Pairs: Semantics-Preserving Image-Caption Pair Augmentation for Grounding-Based Vision and Language Models

Jingru Yi, Burak Uzkent, Oana Ignat, Zili Li, Amanmeet Garg, Xiang Yu, Linda Liu
Amazon Prime Video

Abstract

Grounding-based vision and language models have been successfully applied to low-level vision tasks, aiming to precisely locate objects referred in captions. The effectiveness of grounding representation learning heavily relies on the scale of the training dataset. Despite being a useful data enrichment strategy, data augmentation has received minimal attention in existing vision and language tasks as augmentation for image-caption pairs is non-trivial. In this study, we propose a robust phrase grounding model trained with text-conditioned and text-unconditioned data augmentations. Specifically, we apply text-conditioned color jittering and horizontal flipping to ensure semantic consistency between images and captions. To guarantee image-caption correspondence in the training samples, we modify the captions according to pre-defined keywords when applying horizontal flipping. Additionally, inspired by recent masked signal reconstruction, we propose to use pixel-level masking as a novel form of data augmentation. While we demonstrate our data augmentation method with MDETR framework, the proposed approach is applicable to common grounding-based vision and language tasks with other frameworks. Finally, we show that image encoder pretrained on large-scale image and language datasets (such as CLIP) can further improve the results. Through extensive experiments on three commonly applied datasets: Flickr30k, referring expressions and GQA, our method demonstrates advanced performance over the state-of-the-arts with various metrics. Code can be found in `https://github.com/amzn/augment-the-pairs-wacv2024`.

1. Introduction

Phrase grounding identifies objects in a scene based on the understanding of language. It requires the model to comprehend the visual context and relate object regions with sentences or phrases [36, 47, 56]. Compared to conventional object detection, phrase grounding alleviates the bottleneck of fixed vocabulary and is able to generalize to unseen categories and attributes based on learning of nuance concepts of the free-form text [19, 49, 50]. As a result, a dataset that involves rich region-phrase and image-language correspondences is important for generalizable phrase grounding. For instance, MDETR [19] surpasses previous works with an effective pretraining datasets of 1.3M image-text pairs. On the other hand, GLIP [25] demonstrates strong zero-shot and few-shot transferability by scaling up visual concepts with 27M grounding data. Aside from data scale, existing works have barely explored data augmentation in the phrase grounding task, despite the significant role that data augmentation plays in defining effective predictions across various tasks [2, 6, 11].

Data augmentation has been extensively studied and employed in object detection [1, 11, 14] to increase the density and variety of training samples [60]. For phrase grounding task, the sample shortage problem is more severe. For example, Flickr30k Entities [36] contains 44.5k object categories, with only an average of 6.2 objects per category. Data augmentation could be crucial to improve model's generalization ability. To enhance phrase grounding understanding, some works [48, 51] employ data augmentations such as horizontal flipping in their pipeline. However, applying data augmentation to phrase grounding can easily disrupt the image-language correspondence. For instance, as shown in Figure 1 (a), color jittering can alter the colors of objects, causing a mismatch between the object regions and the corresponding noun phrases in the caption. Merely removing color-related words may lead to errors in ground-truth bounding boxes, as objects in different color may not be explicitly mentioned in the caption. Likewise, horizontal flipping is associated with words that convey left or right. A simple flip of words containing "left" or "right" can introduce image-caption misalignment (see Figure 1(b)).

To address these limitations, this paper proposes a novel text-conditioned augmentation approach, wherein we apply color jittering and horizontal flipping transformations to image-caption pairs that do not contain color-related keywords (e.g., red, yellow) and contain position-relevant words respectively. Furthermore, we utilize caption-independent data augmentations such as pixel-level and block-level masking to further enhance the learning of rep-

(a) Color Jittering (b) Horizontal Flipping

Figure 1. Examples of challenges in applying data augmentation in phrase grounding task. We highlight two common augmentations: (a) color jittering and (b) horizontal flipping. The bounding boxes in the image correspond to the same color phrases as those in the caption.

resentations. In this paper, we pick one of the representative method MDETR, and apply augmentation strategies on top of it to demonstrate the effectiveness in improving the phrase grounding model on three datasets: Flickr30k [36], referring expressions [55] and GQA [18]. These augmentation strategies can be seamlessly integrated into other orthogonal phrase grounding models by enriching image-caption correspondences in the training datasets. The experiments show that leveraging image encoder pretrained on larger-scale image-language datasets (e.g., CLIP [37]) leads to additional performance gain. Our contributions can be summarized as:

- We propose text-conditioned and text-unconditioned data augmentations to effectively enrich the data diversity, which can orthorgonally improve the vision and language grounding frameworks, e.g., MDETR.
- We show that by utilizing an image encoder pretrained on larger-scale vision and language datasets, such as CLIP, the embedding power can be significantly enhanced and thus improve the grounding performance.
- Extensive experiments on pretraining and downstream tasks demonstrate the superiority of the proposed method. Specifically, in the pretraining phrase grounding task, our method improves MDETR by 0.5%, 3.3% and 1% AP on the validation set of Flickr30k, referring expressions and GQA datasets, respectively.

2. Related Work

Grounding-based Vision and Language Models We can categorize existing grounding-based image and language models into two categories: two-stage and single-stage. Two-stage methods [5, 32, 53] rely on off-the-shelf object detectors to get object proposals and then process the language query for the task of interest. On the other hand, single-stage methods [3, 7, 10, 19, 25, 43, 46, 58] avoid using a separate off-the-shelf object detector and perform end-to-end training for detecting the referred object, reducing the computational complexity of the two-stage methods. For example, MDETR [19] has trained an object detector (*i.e.* DETR [1]) on a concatenation of learned image and language representations. Lite-MDETR [31] further reduces MDETR model size by leveraging a light-weight backbone and employing quantization. The most recent vision and language models [7, 19, 25, 26, 45] utilize large-scale transformers to improve the accuracy of the previous models with CNN backbones [3,46]. Other recent works [27,37,40] including CLIP [37] have developed image-language models trained on large-scale data with high-level image-to-text contrastive learning. We demonstrate that the phrase grounding model can be further improved by incorporating such models.

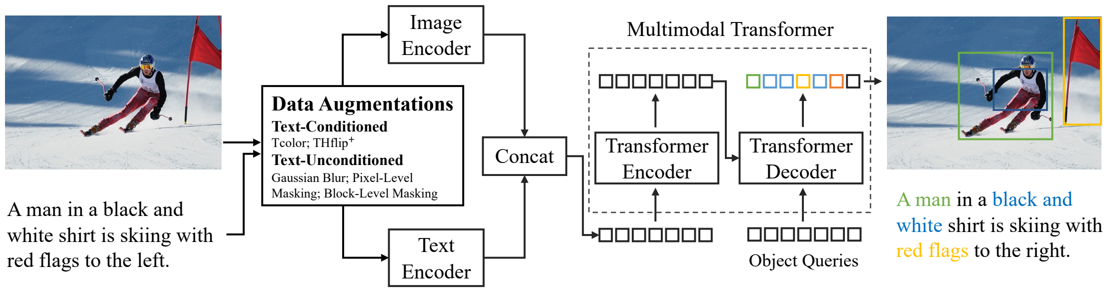

Figure 2. Illustration of phrase grounding framework, where input is an image-caption pair, and output is a set of grounded object regions mentioned in the caption. During training, we apply text-conditioned and text-unconditioned data augmentations to the input in order to increase sample's density and variety. The representations from image and text encoders are concatenated and fed into a multimodal transformer, which learns to associate the textual and visual modalities for vision and language tasks.

Figure 3. Illustration of text-conditioned and text-unconditioned data augmentations. Text-unconditioned data augmentations consist of Gaussian blur, pixel-level and block-level masking, while text-conditioned data augmentations involve text-conditioned color jitterring (TColor) and horizontal flipping (THflip$^+$).

Data Augmentation Data augmentation consistently leads to improved generalization [22, 41] in computer vision tasks. Elastic distortions with scale, translation and rotation, random cropping, image mirroring and color jittering are common data augmentations applied in classification models trained on natural images [22, 41, 57]. Compared to image classification, data augmentation is crucial for object detection as human annotations can be expensive and time-consuming [60]. Image mirror and multi-scale training are the most widely used augmentation strategies for object detection [1, 14]. Random erase [8, 59], additive noise [13], cut-and-paste [11], augmentation policy learning [60] are also utilized in object detection to improve generalization performances for detection models. In this paper, we employed random erase and additive noise, and tailored the commonly used color jittering and image mirroring augmentations specifically for the phrase grounding task.

3. Method

3.1. Preliminary From MDETR

MDETR [19] model consists of an image and text encoders, and a multimodal transformer (see Figure 2). Given an image encoder f_i and text encoder f_t parameterized by θ_i and θ_t, we denote the output representations $\mathbf{z}_i \in \mathbb{R}^{N\times D}$ and $\mathbf{z}_t \in \mathbb{R}^{M\times D}$ as

$$\mathbf{z}_i = f_i(x_i;\theta_i), \quad \mathbf{z}_t = f_t(x_t;\theta_t), \tag{1}$$

where x_i and x_t represent input image and text. N and M denote the number of image tokens and text tokens, respectively, D is the feature dimension. The image encoder is parameterized by a CNN (*e.g.*, ResNet [17]). The output image features are flattened as sequential image tokens, which are added with a sequence of position embeddings to preserve the spatial information. The text encoder learns text representations through a pre-trained transformer language model RoBERTa [29]. The image and text features are projected into a shared embedding space with a modality dependent linear projection. The modality-specific representations $\mathbf{z}_i$ and $\mathbf{z}_t$ are then concatenated and passed to a transformer encoder f_e, parameterized by θ_e as

$$\mathbf{z}_e = f_e([\mathbf{z}_i, \mathbf{z}_t];\theta_e) \tag{2}$$

where $\mathbf{z}_e$ is the output of transformer encoder. The output representation $\mathbf{z}_e$ as well as object queries $\mathbf{z}_q \in \mathbb{R}^{L\times D}$ are fed into a transformer decoder f_d, where L is the number of object queries. We denote the output object embeddings $\mathbf{z}_o \in \mathbb{R}^{L\times D}$ of MDETR as

$$\mathbf{z}_o = f_d(\mathbf{z}_e, \mathbf{z}_q; |\theta_d), \quad (3)$$

where θ_d is the transformer decoder parameters.

3.2. Data Augmentation

As described in Figure 1, directly applying general data augmentations to phrase grounding task can lead to misalignment between image and caption. To overcome this challenge, we apply text-unconditioned and text-conditioned data augmentation to the input of our model.

3.2.1 Text-Unconditioned Data Augmentation

The text-unconditioned image augmentations are applied regardless of the text query. We carefully select Gaussian blur, pixel-level and block-level masking (see Figure 3) in this direction.

Pixel-level and Block-level Masking. Inspired by the recent works in masked signal modeling works [15, 23], we add pixel-level masking as noise to input images without reconstruction. In particular, we randomly mask the input image pixels with a probability p. We show that a simple pixel-level masking augmentation improves the phrase grounding performance significantly. In addition, block-level masking [59] is adopted to randomly erase a block of pixels from input images (as shown in Figure 3). The masking augmentations add occlusion to input images and force the model to learn to generalize well [59] even though the present objects are not visually clear.

3.2.2 Text-Conditioned Data Augmentation

We further introduce text-conditioned data augmentations by modifying the input images in color and spatial space without breaking the image-caption correspondences of the phrase grounding task. Specifically, we introduce the novel text-conditioned color jittering and horizontal flipping.

Text-Conditioned Color Jittering. As shown in Figure 1(a), a general color jittering can bring errors when the caption contain color information. To address this limitation, we skip the color jittering when an input caption contains any color-related words, we term the method as TColor. In this way, we ensure the caption is still valid even though the image is altered in color space.

Text-Conditioned Horizontal Flipping. Due to the complex interplay of the word combinations (see Figure 1(b)), simply replacing word containing "left" with "right" or vice versa when applying horizontal flipping in phrase grounding task would introduce errors. One way to address this limitation is to skip the horizontal flipping if the caption contains "left" or "right". We term this method as THflip. Although is still valid, THflip misses opportunities that could help the model learn the non-trivial connections between positional word and flipped image. To address this issue, we create a keyword list that contains words "left" or "right", and their variants with suffix "-most, -side, -iest, -middle" or prefix such as "upper-, top-, bottom-, far-", etc. For those captions that are not appearing in the keyword list, we choose not to conduct the horizontal flipping data augmentation. We term the method as THflip$^+$. As depicted in the last column of Figure 3, THflip$^+$ preserves image-caption consistency after modifying the caption. Through ablation experiments, we demonstrate that THflip$^+$ improves the performance of phrase grounding task significantly especially on Referring expression comprehension dataset, when compared to THflip.

3.3. Training Losses

The training of our method involves three losses [19]: bounding box regression, soft token prediction and text-query contrastive alignment. We illustrate each loss below.

Object-Text Contrastive Alignment. To ensure the object representation is closer to the corresponding phrase text tokens in feature space compared to other objects, a contrastive loss (*i.e.*, InfoNCE [35]) is applied to the object embedding and text tokens. Given an object embedding $\mathbf{z}_{oi} \in \mathbb{R}^D$ and its aligned text token set $T_i^+ = \{\mathbf{t} \in \mathbb{R}^D\}$ where D is embedding length, the aligned contrastive loss for all objects can be formulated as:

$$\mathcal{L}_o = \sum_{i=1}^{L} \frac{1}{|T_i^+|} \sum_{j\in T_i^+} -\log\left(\frac{\exp(\mathbf{z}_{oi}^\top \mathbf{t}_j/\tau)}{\sum_{k=1}^{E}\exp(\mathbf{z}_{oi}^\top \mathbf{t}_k/\tau)}\right), \quad (4)$$

where $\tau = 0.07$ is a temperature parameter [44], E is the maximum number of text tokens, L is the number of object queries. Given a text token t_j and its aligned object embedding set O_j^+, the contrastive loss for all text tokens is:

$$\mathcal{L}_t = \sum_{j=1}^{E} \frac{1}{|O_j^+|} \sum_{i\in O_j^+} -\log\left(\frac{\exp(\mathbf{t}_j^\top \mathbf{z}_{oi}/\tau)}{\sum_{k=1}^{L}\exp(\mathbf{t}_j^\top \mathbf{z}_{ok}/\tau)}\right). \quad (5)$$

The object-text contrastive loss can be expressed as $\mathcal{L}_{align} = (\mathcal{L}_o + \mathcal{L}_t)/2$.

Soft Token Prediction. Following MDETR, rather than classifying the detected object, we utilize a soft token prediction method to identify the span of text tokens from input caption for each matched object. Given an object embedding $\mathbf{z}_{oi} \in \mathbb{R}^D$ where i indexes the predicted object, MDETR applies a linear layer to get the soft token predictions: $\mathbf{s}_i = f(\mathbf{z}_{oi})$, where $f : \mathbb{R}^D \to \mathbb{R}^E$ is a linear transformation function, E is the maximum number of text tokens. Cross entropy loss is utilized to train the soft token

Method	Image Encoder	Flickr30k AP	Flickr30k R@1	Referring Expressions AP	RefCOCO R@1	RefCOCO+ R@1	RefCOCOg R@1	GQA AP
MDETR [19]	RN101	32.2	71.7	27.3	60.7	48.5	47.5	19.6
Ours	RN101	35.6	75.2	32.6	66.4	54.9	51.8	23.1
Ours	RN101-CLIP	40.2	78.2	37.1	67.1	55.7	54.0	25.9
Ours	ViT-B-CLIP	41.1	78.5	39.5	71.1	57.6	54.1	27.8
MDETR [19]†	RN101	52.6	82.3	46.9	72.6	58.1	55.3	39.4
Ours†	RN101	53.1	83.3	**50.2**	**74.8**	61.0	57.1	40.4
MDETR [19]†	RN101-CLIP	54.0	83.5	45.9	71.2	57.1	54.4	40.9
Ours†	RN101-CLIP	**54.7**	**84.2**	49.2	72.7	**61.2**	**57.4**	**42.4**

Table 1. Phrase grounding evaluation results on validation sets of Flickr30k [36], referring expressions [55] and GQA [18]. Unless otherwise specified, models were trained on 256×256 pixel images, while the input resolution of ViT-B was 224×224. We denote CLIP for encoders pretrained from [37]. Models with † were trained on 800×1333 pixel images.

predictions:

$$\mathcal{L}_{stoken} = -\frac{1}{n^+}\sum_{i=1}^{L}\sum_{j=1}^{E} s^*_{ij}\log\frac{\exp(s_{ij})}{\sum_{k=1}^{E}\exp(s_{ik})}, \quad (6)$$

where $\mathbf{s}^*_i$ is an uniform distribution of all positive tokens [19], n^+ is the total number of matched objects, L is the number of object queries.

Bounding Box Regression. For bounding box coordinates regression, a multi-layer perceptron (MLP) module is applied to the object embedding $\mathbf{z}_{oi} \in \mathbb{R}^D, i = 1, \cdots, L$. Suppose the predicted box coordinates are $\mathbf{c}_i \in \mathbb{R}^4$, the bounding box regression loss is devised as:

$$\mathcal{L}_{box} = \frac{1}{n^+}\sum_{i}^{L}(L_1(\mathbf{c}_i, \mathbf{c}^*_i) + (1 - \text{GIoU}(\mathbf{c}_i, \mathbf{c}^*_i))), \quad (7)$$

where c^* is the ground-truth box coordinates for matched objects, n^+ indicates the total number of matched objects. The bounding box regression loss combines L_1 loss and generalized intersection over union (GIoU) [39] loss.

4. Experimental Details

Pretraining Datasets. We follow the same pretraining strategy as MDETR [19]. Specifically, the training images are created from a combination of MSCOCO [28], Flickr30k [36], and Visual Genome (VG) [21], where annotations are merged from Flickr entities [36], VG regions [21], referring expressions [34,55] and GQA train balanced set [18]. The dataset comprises bounding box annotations for objects mentioned in the language query, including 200k images and 1.3M aligned image-caption pairs. Among the annotations, Flickr30k Entities [36] contains 31.8k images with 5 sentences per image. It involves 44.5k object categories with 6.2 objects per category. MSCOCO [28] contains 37k images where annotations are collected from referring expressions (RefCOCO [55], RefCOCO+ [55] and RefCOCOg [34]). Visual Genome [21] contains 108k image with 18k object categories where each image has 50 descriptions. We present the phrase grounding performance results for the pretraining task on the validation sets of Flickr30k [36], referring expressions [55] and GQA [18].

Evaluation Metrics. We use average precision (AP) and Recall@K (R@K)) as evaluation metrics. AP [38] is a standard evaluation metric for object detection, it is the average areas under Precision-Recall curve at IoU threshold ranges from 0.5 to 1 with an interval of 0.05. R@K [36] measures the percentage of queries for which a correct match has rank of at most K.

Implementation Details The pretraining stage takes 40 epochs on V100 GPUs with an effective batch size of 128 for models with 256×256 pixel images, and a batch size of 64 for 800×1333 pixel images. We use the same implementation as MDETR. In particular, We use AdamW [30] as optimizer. Learning rate is initialized as 1e-4, which is divided by 10 at epoch 30. We apply the proposed augmentations randomly on input image-caption pairs.

5. Results And Discussions

In this section, we firstly show our method on pretraining phrase grounding task. In particular, we pick one of the state-of-the-art frameworks, *i.e.*, MDETR [19]), as our baseline to demonstrate the effectiveness of the proposed data augmentation. Next, we evaluate our model on two downstream tasks: phrase grounding and referring expression comprehension. Finally, a detailed ablation is presented to highlight different augmentation modules.

5.1. Phrase Grounding Pretraining

With proposed Data Augmentation, our method consistently outperforms MDETR [19]. In Table 1, we evaluate the pretraining performance of our model on validation sets of Flickr30k [36], referring expressions [55] and GQA [18]. In particular, on 256×256 pixel images, the AP of our method exceeds MDETR by 3.4%, 4.3% and 3.5% on Flickr30k, referring expressions and GQA datasets respec-

Figure 4. Visualization of phrase grounding prediction results with 800×1333 pixel images on Flickr30k validation datasets. The bounding boxes with different color correspond to the phrase with the same color in the caption. Underscore distinguishes overlapped phrases.

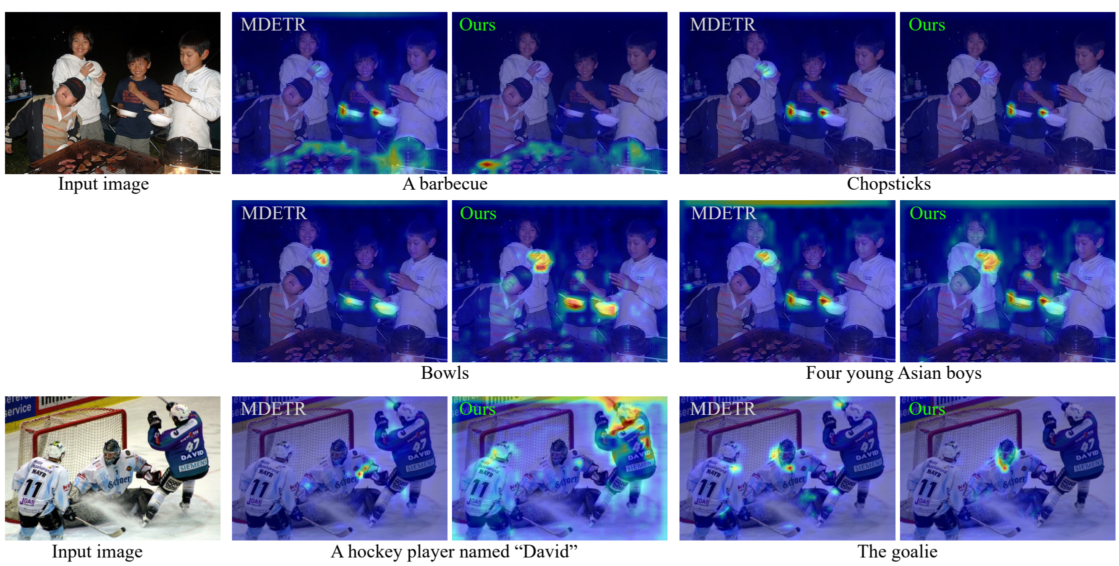

Figure 5. Visualization of attention maps queried by phrases.

tively. With ResNet101 pre-trained on CLIP, our method further improves 4.6%, 4.5% and 2.8% AP on the three datasets. By replacing ResNet101 with vision transformer (*i.e.,* ViT-B [9]) as image encoder, our method obtains further improvement by 0.9%, 2.4% and 1.9% AP. On the other hand, on 800×1333 pixel images, our model surpasses MDETR by 0.5%, 3.3% and 1% AP and by 0.7%, 3.3%, 1.5% AP on the three datasets with RN101 and RN101-CLIP image encoder, respectively. Since the keyword list for THflip^{+} is primarily derived from referring expressions, the significant improvement observed on this dataset suggests that introducing variability in image-caption associa-

Method	Pre-training data	RefCOCO			RefCOCO+			RefCOCOg	
		val	testA	testB	val	testA	testB	val	test
MAttNet [54]	None	76.65	81.14	69.99	65.33	71.62	56.02	66.58	67.27
ViLBERT [33]	CC (3.3M)	-	-	-	72.34	78.52	62.61	-	-
VL-BERT$_L$ [42]	CC (3.3M)	-	-	-	72.59	78.57	62.30	-	-
UNITER$_L$ [4]	CC, SBU, COCO, VG (4.6M)	81.41	87.04	74.17	75.90	81.45	66.70	74.86	75.77
VILLA$_L$ [12]	CC, SBU, COCO, VG (4.6M)	82.39	87.48	74.84	76.17	81.54	66.84	76.18	76.71
ERNIE-ViL$_L$ [52]	CC, SBU (4.3M)	-	-	-	75.95	82.07	66.88	-	-
MDETR-R101 [19]	COCO, VG, Flickr (200k)	86.75	89.58	81.41	79.52	84.09	70.62	81.64	80.89
Ours-R101	COCO, VG, Flickr (200k)	87.47	90.24	81.83	79.91	84.49	71.18	82.64	81.66
MDETR-R101-CLIP	COCO, VG, Flickr (200k)	87.35	90.46	81.93	79.73	84.22	70.94	82.35	81.59
Ours-R101-CLIP	COCO, VG, Flickr (200k)	**87.72**	**90.60**	**82.20**	**80.45**	**85.01**	**71.50**	**82.91**	**82.08**

Table 2. Results (R@1) on referring expression comprehension task.

Method	Val			Test		
	R@1	R@5	R@10	R@1	R@5	R@10
BAN [20]	-	-	-	69.7	84.2	86.4
VisualBert [24]	68.1	84.0	86.2	-	-	-
VisualBert† [24]	70.4	84.5	86.3	71.3	85.0	86.5
MDETR-RN101 [19]	78.9	88.8	90.8	-	-	-
MDETR-RN101†* [19]	82.5	92.9	94.9	83.4	93.5	95.3
Ours-RN101†*	83.3	93.0	95.1	83.7	93.6	95.4
MDETR-RN101-CLIP†*	83.5	93.5	95.3	84.0	94.0	95.9
Ours-RN101-CLIP†*	**84.2**	**93.7**	**95.5**	**84.7**	**94.3**	**95.9**

Table 3. Results of phrase grounding task on Flickr30k entities dataset (Any-Box protocol [19]). Models with † are pre-trained on COCO, models with $*$ are also pre-trained on VG and Flickr 30k.

tions enhances the model's learning ability for vision and language tasks.

With proposed Data Augmentation, qualitative results exhibit better semantic understanding. Figure 4 shows qualitative results of phrase grounding on the Flickr30k validation dataset, using RN101 as backbone and 800×1333 as input resolution. Our method shows increased robustness in suppressing redundant detection. For example, it effectively improves the redundant detection of "a flag" in the first image. In addition, it rectifies the erroneous detection of "barbecue" and "chopsticks" in the third column of images. More importantly, with the richer image-text correspondences introduced by data augmentation in the training dataset, the model shows better understanding of the context. As depicted in the last column of Figure 4, our method has correctly differentiated between "the goalie" and "a hockey player", whereas MDETR fails to recognize the "goalie". Moreover, by analyzing attention maps in Figure 5, we observe that our model mainly relies on helmet features of the goalie to make decisions, whereas MDETR is influenced by ambiguous features.

Failure Cases. We also visualize some failure cases in Figure 4. For example, in the third column, both models fail to identify the chopsticks held by the second boy. As the color of the chopsticks is similar to the boy's shirt, this may indicate that the framework has limitations in detection of small objects with a similar color as the background. In the last image, our model identifies two persons as "a hockey player named David", indicating that the model is incapable of recognizing text "David" on the shirt. In Figure 5, we also notice that MDETR does not consider text "David" even though it localizes the correct person. This may highlight the importance of text recognition from visual features for phrase grounding, which remains challenging due to limited training data. Nevertheless, the visualization results reveal that augmentations introduced in this work have significantly enhanced the model's robustness in the phrase grounding task.

5.2. Downstream Tasks

We finetune our model from pretraining stage for two downstream tasks: phrase grounding and referring expression comprehension, using the same training setting as MDETR [19]. Notice that there is no additional data augmentation applied for downstream tasks finetuning.

Phrase Grounding The phrase grounding downstream task is performed on Flickr30k entities dataset [36]. We follow the same setup as in MDETR [19] by taking the top 100 bounding box predictions and soft token predictions to align the bounding boxes to the caption. We compare our method to the prior works under the ANY-BOX protocol [19]. Similar to MDETR, the evaluation is conducted on models from pretraining stage, as additional fine-tuning does not yield further performance improvement. The results presented in Table 3 demonstrate that the proposed method consistently outperforms MDETR and other state-of-the-art methods. By further leveraging a backbone pretrained on large-scale vision and language datasets (*i.e.*, CLIP [37]), our model exhibits even more superior performance. This evidences that larger-scale foundation model enables better representation power benefited from the large language models.

Referring Expression Comprehension Referring expression comprehension is a task that localizes the objects being referred from input image. In this task, only one box will

Gaussian Blur	THflip	TColor	Pixel Mask	Block Mask	Flickr30k AP	Flickr30k R@1	Referring Expressions AP	RefCOCO R@1	RefCOCO+ R@1	RefCOCOg R@1	GQA AP
					32.2	59.0	27.3	60.7	48.5	47.5	19.6
✓					32.8	60.5	27.1	59.6	47.5	47.5	19.0
	✓				34.0	74.3	28.3	59.7	50.6	49.2	21.7
	✓+				34.8	74.2	**34.0**	**69.9**	54.0	**53.4**	21.4
		✓			32.4	72.1	29.6	63.9	49.8	48.4	20.2
			✓		34.5	73.1	29.8	63.9	50.3	48.9	20.6
				✓	33.8	73.6	29.7	65.4	50.7	49.7	21.1
	✓	✓	✓	✓	35.8	75.2	31.2	65.3	52.6	50.5	**23.5**
✓	✓	✓	✓	✓	**36.3**	74.5	31.4	64.3	51.5	52.3	23.1
✓	✓+	✓	✓	✓	35.6	**75.3**	32.6	66.4	**54.9**	51.8	23.1

Table 4. Ablation studies with different augmentation strategies on pretraining task. We evaluate the model performances on validation set of Flickr30k [36], referring expressions [55] and GQA [18]. THflip and TColor refer to text-conditioned horizontal flipping and color jittering. Symbol + indicates THflip+. Note that we use 224×224 pixel images for this experiments.

be returned for each input expression. We present the results on three commonly applied datasets, RefCOCO [55], RefCOCO+ [55] and RefCOCOg [34]. Following the same setting as MDETR, we finetune our model for 5 epochs as it is slightly different from pretraining phrase grounding task. As shown in Table 2, with the same backbone, our method consistently improves the performance over MDETR on both the validation and test splits across the three datasets, suggesting that the proposed data augmentations effectively enhance the learned vision and language representation.

5.3. Ablation Study

To study the effectiveness of the proposed data augmentations, we ablate our model on 256×256 pixel images and report the model performances of pretraining phrase grounding task in Table 4. We observe that the Gaussian blur alone is not effective on referring expressions and GQA datasets. On referring expressions, THflip+ outperforms THflip by 5.7% in terms of AP. It suggests that augmenting both images and captions can effectively enrich the data variance and thus improve the learned embedding representation power. TColor improves the baseline on three datasets, but it is not as effective as THflip+ and pixel and block-level masking. This is attributed to the fact that TColor skips augmentation when a caption contains color-related words ($\sim$ 49% in pretraining datasets), thus limiting the variance it could bring to the training samples. Pixel-level or block-level masking introduce further difficulties for the model to detect the occluded objects and therefore it is an effective method in improving the representations. Among all the augmentations, THFlip+ achieves the best on referring expressions. By combining all the augmentations, our approach further improves the performance of THflip+ on Flickr30k and GQA datasets without much sacrifice on referring expressions.

Masking Ratio. To study the influence of masking ratio of pixel-level masking augmentation on phrase grounding task, we visualize the results at masking ratio of 20%, 50%, 75% and 80% on three datasets in Figure 6. On Flickr30k, the masking ratio shows similar trend as MAE [16], where 75% ratio achieves the best performance. The higher masking ratio on referring expressions shows better performance, while the AP value is similar at 75% and 80% masking ratio. On GQA dataset, the AP value has minor difference across different masking ratios. Therefore, we empirically select 75% ratio for the pixel-level masking augmentation.

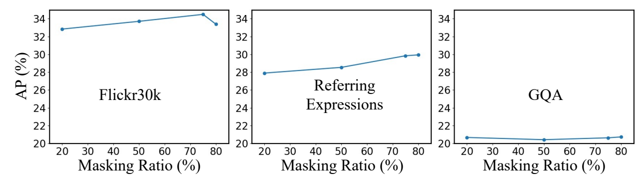

Figure 6. Phrase grounding performances with pixel-level masking augmentation only at different masking ratio on three datasets.

6. Conclusion

In this paper, we address the challenges that traditional data augmentations can hardly reserve the semantics in grounding-based vision and language tasks, as they can disrupt the image-caption correspondences. We propose text-conditioned (TColor and THflip+) and text-unconditioned (pixel-level and block-level masking) data augmentations to enrich the image-caption density and diversity while preserving semantic coherence between object regions and corresponding phrases. Achieving this is challenging due to the complex interplay of word combinations. With extensive experiments, we demonstrate that our method can effectively enhance the learned feature representations for grounding-based vision and language tasks. Further ablations show the effectiveness of our proposed augmentations against traditional Gaussian blur and masking operations. Future work will focus on how to generalize to an even broader range of the augmentations to further expand the variation of the input space.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 1, 2, 3

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 1

[3] Xinpeng Chen, Lin Ma, Jingyuan Chen, Zequn Jie, Wei Liu, and Jiebo Luo. Real-time referring expression comprehension by single-stage grounding network. *arXiv preprint arXiv:1812.03426*, 2018. 2

[4] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. *ArXiv*, abs/1909.11740, 2019. 7

[5] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020. 2

[6] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 113–123, 2019. 1

[7] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. *arXiv preprint arXiv:2104.08541*, 2021. 2

[8] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 3

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6

[10] Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. *arXiv preprint arXiv:2206.07643*, 2022. 2

[11] Debidatta Dwibedi, Ishan Misra, and Martial Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1301–1310, 2017. 1, 3

[12] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *ArXiv*, abs/2006.06195, 2020. 7

[13] Justin Gilmer, Nicolas Ford, Nicholas Carlini, and Ekin Cubuk. Adversarial examples are a natural consequence of test error in noise. In *International Conference on Machine Learning*, pages 2280–2289. PMLR, 2019. 3

[14] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 1, 3

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 4

[16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll'ar, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, 2021. 8

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[18] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 2, 5, 8

[19] Aishwarya Kamath, Mannat Singh, Yann LeCun, Ishan Misra, Gabriel Synnaeve, and Nicolas Carion. Mdetr - modulated detection for end-to-end multi-modal understanding. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1760–1770, 2021. 1, 2, 3, 4, 5, 7

[20] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Neural Information Processing Systems*, 2018. 7

[21] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 5

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 3

[23] Gukyeong Kwon, Zhaowei Cai, Avinash Ravichandran, Erhan Bas, Rahul Bhotika, and Stefano Soatto. Masked vision and language modeling for multi-modal representation learning. *arXiv preprint arXiv:2208.02131*, 2022. 4

[24] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *ArXiv*, abs/1908.03557, 2019. 7

[25] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 1, 2

[26] Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. *arXiv preprint arXiv:2106.03089*, 2021. 2

[27] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards

unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*, 2020. 2

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[29] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 3

[30] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[31] Qian Lou, Yen-Chang Hsu, Burak Uzkent, Ting Hua, Yilin Shen, and Hongxia Jin. Lite-mdetr: A lightweight multimodal detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12206–12215, June 2022. 2

[32] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019. 2

[33] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Neural Information Processing Systems*, 2019. 7

[34] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 5, 8

[35] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[36] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 1, 2, 5, 7, 8

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 5, 7

[38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 5

[39] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 5

[40] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021. 2

[41] Patrice Y Simard, Yann A LeCun, John S Denker, and Bernard Victorri. Transformation invariance in pattern recognition—tangent distance and tangent propagation. In *Neural networks: tricks of the trade*, pages 239–274. Springer, 2002. 3

[42] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *ArXiv*, abs/1908.08530, 2019. 7

[43] Burak Uzkent, Amanmeet Garg, Wentao Zhu, Keval Doshi, Jingru Yi, Xiaolong Wang, and Mohamed Omar. Dynamic inference with grounding based vision and language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2624–2633, 2023. 2

[44] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 4

[45] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Tubedetr: Spatio-temporal video grounding with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16442–16453, 2022. 2

[46] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving one-stage visual grounding by recursive subquery construction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 387–404. Springer, 2020. 2

[47] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4683–4693, 2019. 1

[48] Ziyan Yang, Kushal Kafle, Franck Dernoncourt, and Vicente Ordonez. Improving visual grounding by encouraging consistent gradient-based explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19165–19174, 2023. 1

[49] Lewei Yao, Jianhua Han, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, and Hang Xu. Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment. *arXiv preprint arXiv:2304.04514*, 2023. 1

[50] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *arXiv preprint arXiv:2209.09407*, 2022. 1

[51] Yuan Yao, Qianyu Chen, Ao Zhang, Wei Ji, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Pevl: Position-enhanced pre-training and prompt tuning for vision-language models. *arXiv preprint arXiv:2205.11169*, 2022. 1

[52] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-

language representations through scene graph. In *AAAI Conference on Artificial Intelligence*, 2020. 7

[53] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1307–1315, 2018. 2

[54] Licheng Yu, Zhe L. Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. Mattnet: Modular attention network for referring expression comprehension. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1307–1315, 2018. 7

[55] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016. 2, 5, 8

[56] Zhou Yu, Jun Yu, Chenchao Xiang, Zhou Zhao, Qi Tian, and Dacheng Tao. Rethinking diversified and discriminative proposal generation for visual grounding. *arXiv preprint arXiv:1805.03508*, 2018. 1

[57] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 3

[58] Haotian* Zhang, Pengchuan* Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *arXiv preprint arXiv:2206.05836*, 2022. 2

[59] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008, 2020. 3, 4

[60] Barret Zoph, Ekin D Cubuk, Golnaz Ghiasi, Tsung-Yi Lin, Jonathon Shlens, and Quoc V Le. Learning data augmentation strategies for object detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pages 566–583. Springer, 2020. 1, 3