Improved Visual Grounding through Self-Consistent Explanations

Ruozhen He[1] Paola Cascante-Bonilla[1] Ziyan Yang[1] Alexander C. Berg[2] Vicente Ordonez[1]
[1]Rice University [2]University of California, Irvine
catherine.he@rice.edu, pc51@rice.edu, zy47@rice.edu, bergac@uci.edu, vicenteor@rice.edu

2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 979-8-3503-5300-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/CVPR52733.2024.01244

Abstract

Vision-and-language models trained to match images with text can be combined with visual explanation methods to point to the locations of specific objects in an image. Our work shows that the localization –"grounding"– abilities of these models can be further improved by finetuning for self-consistent visual explanations. We propose a strategy for augmenting existing text-image datasets with paraphrases using a large language model, and SelfEQ, a weakly-supervised strategy on visual explanation maps for paraphrases that encourages self-consistency. Specifically, for an input textual phrase, we attempt to generate a paraphrase and finetune the model so that the phrase and paraphrase map to the same region in the image. We posit that this both expands the vocabulary that the model is able to handle, and improves the quality of the object locations highlighted by gradient-based visual explanation methods (e.g. GradCAM). We demonstrate that SelfEQ improves performance on Flickr30k, ReferIt, and RefCOCO+ over a strong baseline method and several prior works. Particularly, comparing to other methods that do not use any type of box annotations, we obtain 84.07% on Flickr30k (an absolute improvement of 4.69%), 67.40% on ReferIt (an absolute improvement of 7.68%), and 75.10%, 55.49% on RefCOCO+ test sets A and B respectively (an absolute improvement of 3.74% on average).

1. Introduction

Vision-and-language models that are trained to associate images with text have shown to be effective for many tasks and benchmarks [21, 27, 31, 41], including object detection [18, 54] and image segmentation [15, 36, 50]. Since these models are typically trained with in-the-wild data from the web, they can handle a wide range of vocabulary for objects as long as they are well represented in the training data. These models are often remarkably accurate [16, 22, 28, 47] even without tuning them to perform well in any particular downstream task [14, 32, 52]. The ALBEF model [27] was particularly capable of visual

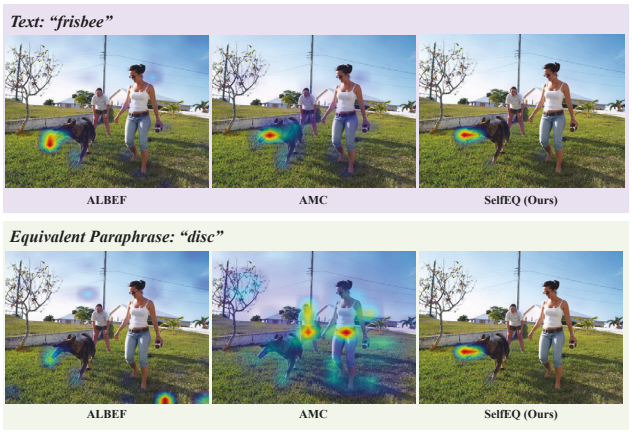

Figure 1. Previous models can localize the word *frisbee*, but struggle with equivalent but more uncommon referents such as *disc*. A model tuned with our proposed SelfEQ objective encourages consistent visual explanations for paraphrased prompts and performs well on both examples. SelfEQ not only enables a larger working vocabulary but also improves overall localization performance.

"grounding" – or in other words – the ability to localize objects in images by simply using it in conjunction with a visual explanation method such as GradCAM [43]. This capability is particularly remarkable given that this model was only supervised with images and text but no object location annotations of any type.

In order to improve the ability of vision-and-language models to perform localization, many methods have incorporated further finetuning with either box or segment annotations, or rely on pretrained object detectors or box proposal networks [5, 12, 19, 23, 30, 51]. Our work instead aims to improve the localization capabilities of models trained only on image-text pairs through weak supervision. But, how can we improve the ability of a model to localize objects without access to object location annotations? Consider the example in Figure 1 where a model is tasked with pointing to the location of the object *frisbee* in this image. The baseline model succeeds at finding the object but is unsuccessful at locating the object when prompted with the

2575-7075/24/$31.00 ©2024 IEEE
DOI 10.1109/CVPR52733.2024.01244

equivalent but more generic name *disc*. Regardless of the ability for the base model to find either of these, the visual explanations for these two prompts should be the same since the query refers to the very same object in both cases. Our work exploits this property by first generating paraphrases using a large language model and then proposing a weakly-supervised **Self**-consistency **EQ**uivalence Tuning (SelfEQ) objective that encourages consistent visual explanations between paraphrased input text pairs that refer to the same object or region in a given image.

Given a base pre-trained vision-and-language model purely trained on image-text pairs such as ALBEF [27], SelfEQ tunes the model so that for a given input image and text pair, the visual attention map extracted using GradCAM [43] produces a similar visual attention map when provided with the same image and a text paraphrase. Figure 2 provides an overview of our method. Another contribution of our work consists in exploiting a large language model (LLM) to automatically generate paraphrases for existing datasets such as Visual Genome [26] that contains textual descriptions of individual objects and regions, or MS-COCO [33] and CC3M [46] that contain global image descriptions. We find that SelfEQ not only expands the vocabulary of objects that the base model is able to localize but more importantly, improves the visual grounding capabilities of the model on standard benchmarks such as referring expression comprehension on the ReferIt benchmark [24] and region-phrase grounding in the Flickr30K Entities benchmark [40]. In summary, our key contributions are as follows:

- We design a novel objective, SelfEQ, to encourage vision-and-language models to generate self-consistent visual explanations for equivalent text phrases, thereby improving grounding capabilities while expanding the working vocabulary of the model.
- We propose to prompt large language models for generating paraphrased image descriptions of individual objects or regions. Particularly, we adopt Vicuna-13B [6] and design text prompts to obtain high quality paraphrases.
- We demonstrate the effectiveness of our method by outperforming previous methods, leading to 4.69% improvement on Flickr30k, 7.68% improvement on ReferIt, and 3.74% improvement on RefCOCO+.

Finally, we plan to release our code, generated paraphrases and model checkpoints upon publication.

2. Related Work

Our work is related to previous methods on visual grounding, especially those that are trained under weak supervision understood as without the use of bounding box or segment annotations and relying only on image-text pairs. From a technical perspective our work is related to methods that optimize visual explanations to improve the underlying model.

Visual grounding consists of localizing an input textual description in an image. Supervised methods are provided with text-image pairs and corresponding bounding boxes [8, 9, 12, 23, 51]. Other supervised methods leverage pretrained object detectors to obtain a region of interest and then identify the region that aligns most closely under their textual representations [5, 7, 17, 19, 35, 48]. In both cases, these methods use some form of box supervision during pre-training or at test time by relying on a pre-trained object detector. In contrast, our work focuses exclusively in the scenario where no bounding boxes or segment annotations are available at any stage.

Weakly-Supervised Grounding. Our setup is similar to that of Arbelle et al [3] where no box annotations or object detectors are used for grounding. This work proposes Grounding by Separation (GbS) where a model is trained on randomly alpha-blended pairs of images and the goal is to separate them conditioned on text prompts. Our method instead relies on data augmentation on the text side and while our method shows favorable results, our contribution is orthogonal. Shaharabany et al [45] builds a weakly-supervised phrase grounding model by creating a large amount of data by combining region boxes with a BLIP captioning model. Later work by Shaharabany and Wolf [44] employs layer-wise relevance propagation [38] to integrate relevancy and gradient information with the scores computed from each attention head in the transformer layers [4], or residual connections [1]. Our work compares favorably or on par with these methods but since we rely on gradient-based explanations, our work does not require making any modifications to the base network.

Visual Explanation Tuning. Related to our method is the strategy used by ALBEF [27] where the model is only supervised on image-text pairs and performs grounding through GradCAM [43], a gradient-based visual explanation method that outputs a heatmap indicating spatial relevance. Earlier, Xiao et al [49] used a similar strategy but further optimized gradient-based explanations using structural constraints derived from text. Recently, AMC [51] follows this strategy but further adds box supervision on the output maps using a margin-based loss. We adopt ALBEF [27] as our base model and also adopt a gradient-based explanation strategy but unlike AMC [51] we do not rely on box annotations for tuning this model and use our proposed SelfEQ objective instead. Javed et al [20] proposed an objective function that encourages consistent representations in embedding space for the same input prompt on different images. In contrast SelfEQ encourages consistent visual explanations for different prompts on the same image by relying on automatically generated paraphrases. Akbari et al [2] also optimizes attention maps but in their formulation the model backbone is modified to explicitly incorporate attention instead of relying on gradient-based attention maps.

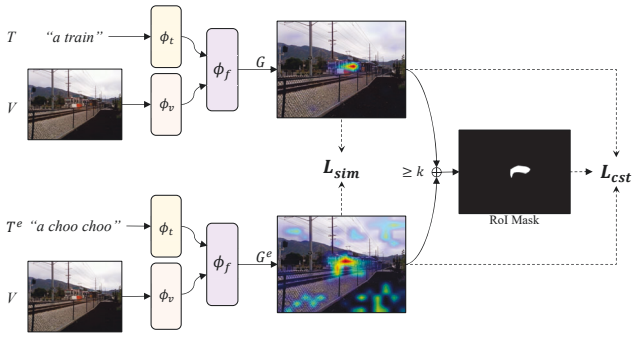

Figure 2. Overview of our proposed weakly-supervised **Self**-consistency **EQ**uivalence tuning objective. We input image-text and image-paraphrase pairs $\langle V, T\rangle$ and $\langle V, T^e\rangle$ to our base pretrained vision-and-language model. We then obtain gradient-based visual explanations $\langle G, G^e\rangle$ and compute a similarity loss between them. We also define an overlapping region of interest mask and encourage the model to predict consistently high saliency scores within this mask for each input pair.

3. Method

We start from a base vision-language model composed of a text encoder ϕ_t , an image encoder ϕ_v, and a multimodal fusion encoder ϕ_f, along with a dataset D to finetune this model consisting of image-text pairs $\langle T, V\rangle$. Section 3.1, introduces the training objectives for the base model which we also adopt for finetuning our baseline and are also used in conjunction with SelfEQ to finetune our final model. Section 3.2 introduces in detail SelfEQ, our self-consistency equivalence tuning objective that assumes the existence of paraphrases T^e for each input image-text pair $\langle T, V\rangle$, and Section 3.3 describes our approach for automatically generating paraphrases T^e using LLM-prompting. Figure 2 presents an overview of our proposed method.

3.1. Base Model: Preliminaries

Our base vision-and-language model is ALBEF [27] which relies on three widely used objectives for visual and textual representation learning: image-text matching, masked language modeling and a contrastive loss. We describe them here briefly as they are also re-used during fine-tuning.

Image-Text Matching Loss (ITM). This loss is calculated using the output of the `[CLS]` token to predict if the input image and the input text are matching or not and is defined as follows:

$$\mathcal{L}_{\text{itm}} = \mathbb{E}_{(V,T)\sim D}\, \mathcal{H}\left(\vec{y}, \phi_f^{\text{cls}}\left(\phi_v\left(V\right), \phi_t\left(T\right)\right)\right), \quad (1)$$

where $\vec{y}$ denotes a two-dimensional one-hot vector, indicating whether the sample $\langle V, T\rangle$ constitutes a match, ϕ_f^{cls} represents a linear layer followed by a softmax function, and $\mathcal{H}$ is the cross entropy loss function.

Masking Language Modeling Loss (MLM). This loss has been applied for various vision-language pretraining models [5, 27, 29, 34]. It integrates the contextual text and the input image to infer masked words in the input text. After utilizing a linear layer and a softmax activation function ϕ_f^{m} to individual output embeddings, the objective is expressed as:

$$\mathcal{L}_{\text{mlm}} = \mathbb{E}_{(V,T^{-\text{m}})\sim D}\, \mathcal{H}\left(\vec{t}^{\,\text{m}}, \phi_f^{\text{m}}\left(\phi_v\left(V\right), \phi_t\left(T^{-\text{m}}\right)\right)\right), \quad (2)$$

where the one-hot vector $\vec{t}^{\,\text{m}}$ denotes the masked token, and $T^{-\text{m}}$ represents the input masked text.

Image-Text Contrastive Loss (ITC). It improves the alignment between visual and textual representations by bringing closer the representations for corresponding text-image pairs relative to text-image pairs that do not correspond. This objective can be defined as follows:

$$\begin{aligned}\mathcal{L}_{\text{itc}} = \mathbb{E}_{(V,T)\sim D}\frac{1}{2}\Bigg[&\mathcal{H}\left(\vec{y}, \frac{\exp\left(\phi_v(V)\cdot\phi_t(T)\right)/\tau}{\sum_{b=1}^{B}\text{s}\left(V, T_b\right)}\right)\\ &+\mathcal{H}\left(\vec{y}, \frac{\exp\left(\phi_t(T)\cdot\phi_v(V)\right)/\tau}{\sum_{b=1}^{B}\text{s}\left(T, V_b\right)}\right)\Bigg],\end{aligned} \quad (3)$$

where B is the number of negative sample pairs, and τ is a temperature parameter for the softmax function.

The training objective for the base model is a combination of the previous three loss functions:

$$\mathcal{L}_{\text{vl}} = \mathcal{L}_{\text{itm}} + \mathcal{L}_{\text{mlm}} + \mathcal{L}_{\text{itc}}. \quad (4)$$

This loss $\mathcal{L}_{\text{vl}}$ will also be used to tune our baseline model.

3.2. Self-Consistency Equivalence Tuning

SelfEQ assumes that the model has access to paraphrases T^e for each input image-text pair $\langle V, T\rangle$ or, in practice, for a subset of those samples. Therefore we assume a finetuning dataset D' with triplets $\langle V, T, T^e\rangle$ such that T^e exists for a corresponding input text T. The first step to define our SelfEQ objective is to generate the explanation heatmaps (*i.e.,* attention maps) through GradCAM [43] conditioned on the input text. We extract intermediate feature maps from the multimodal interactive encoder ϕ_f for input pairs $\langle V, T\rangle$ and $\langle V, T^e\rangle$ as follows:

$$F = \phi\left(\phi_v(V), \phi_t(T)\right), F^e = \phi\left(\phi_v(V), \phi_t(T^e)\right), \quad (5)$$

where ϕ denotes the feature map extraction operation. We then proceed to calculate the gradient of F and F^e related to the image-text matching score $\mathcal{L}_{\text{itm}}$. This computation yields the attention maps for the original text and paraphrased text, referred to as G and G^e, respectively:

$$\begin{aligned}G &= \text{ReLU}\left(F \odot \nabla\mathcal{H}\left(\vec{y}, \phi_f^{cls}\left(\phi_v(V), \phi_t(T)\right)\right)\right),\\ G^e &= \text{ReLU}\left(F^e \odot \nabla\mathcal{H}\left(\vec{y}, \phi_f^{cls}\left(\phi_v(V), \phi_t(T^e)\right)\right)\right).\end{aligned} \quad (6)$$

Our SelfEQ tuning is based on the premise that if a vision-language model is identified as self-consistent, the attention maps produced for both the text and its equivalent paraphrase should yield nearly identical results. To achieve this, we first apply a simple mean squared error loss over the produced heatmaps so that their ℓ_2 distance is minimized and thus become more similar.

$$\mathcal{L}_{\text{sim}} = \mathbb{E}_{(V,T,T^e)\sim D'}\left[\frac{1}{N}\sum_{i,j}(G_{i,j} - G^e_{i,j})^2\right]. \tag{7}$$

Nevertheless, while minimizing a sum of pixel-wise distances contributes to self-consistency, without a regularization term this loss can easily fall into a trivial solution. For instance, it could lead to attention maps with uniformly negative or positive predictions, or just really small values. To address this limitation, we propose to integrate these heatmaps by defining a Region of Interest (RoI) mask. This mask is designed to preserve regions within the attention maps that possibly contain correct predictions. Our approach hinges on the observation that, despite the predictions of equivalent textual inputs being inconsistent, sometimes regions with large values or regions that overlap between the two heatmaps tend to be correct. As such, we assume that if the sum of attention scores at a given position (i, j) exceeds a certain threshold k, it is likely indicative of a correct prediction. We formalize the condition as follows:

$$M_{i,j} = \begin{cases} 1, (G_{i,j} + G^e_{i,j}) \geq k \\ 0, (G_{i,j} + G^e_{i,j}) < k \end{cases}. \tag{8}$$

The attention maps within RoI masks are obtained by element-wise multiplication as follows:

$$R = G \odot M, \quad R^e = G^e \odot M. \tag{9}$$

The integration of RoI masks allows us to use equivalent texts for mutual supervision, refining and improving the accuracy and providing regularization for the previously defined distance-based loss. Moreover, it could potentially address errors owing to unknown or less common words by working vocabulary expansion. Presuming one of the textual expressions is known and correctly understood, the model could extrapolate the meaning of the other equivalent expression via weak supervision. To implement it, we first compute the mean μ_{RoI}, μ^e_{RoI} and the standard deviation $\sigma_{RoI}, \sigma^e_{RoI}$ within the RoI as follows:

$$\mu_{RoI} = \frac{\sum_{i,j} R_{i,j}}{\sum_{i,j} M_{i,j}}, \mu^e_{RoI} = \frac{\sum_{i,j} R^e_{i,j}}{\sum_{i,j} M_{i,j}}, \tag{10}$$

$$\begin{aligned} \sigma_{RoI} &= \sqrt{\frac{\sum_{i,j} M_{i,j} \cdot (R_{i,j} - \mu_{RoI})^2}{\sum_{i,j} M_{i,j}}}, \\ \sigma^e_{RoI} &= \sqrt{\frac{\sum_{i,j} M_{i,j} \cdot (R^e_{i,j} - \mu^e_{RoI})^2}{\sum_{i,j} M_{i,j}}}. \end{aligned} \tag{11}$$

We propose a consistency loss ($\mathcal{L}_{\text{cst}}$), expecting the RoI regions of attention maps to achieve consistently high scores, further reinforcing self-consistency, accuracy, and potential working vocabulary expansion. This objective is formulated as follows:

$$\begin{aligned} \mathcal{L}_{\text{cst}} =& \mathbb{E}_{(V,T,T^e)\sim D'}\Big[\sigma_{RoI} + \sigma^e_{RoI} + \\ & \max(0, k/2 - \mu_{RoI}) + \max(0, k/2 - \mu^e_{RoI})\Big]. \end{aligned} \tag{12}$$

Finally, the objective of our self-consistency equivalence tuning is expressed as:

$$\mathcal{L}_{\text{SelfEQ}} = \mathcal{L}_{\text{sim}} + \lambda \cdot \mathcal{L}_{\text{cst}}, \tag{13}$$

where λ is a hyperparameter to control the relative influence of each loss.

3.3. Self-Consistency Data Augmentation

In this section we define a function $\mathcal{F}$ that can automatically map input textual phrases T as paraphrases T^e without the need to rely on human annotations such that $T^e \sim \mathcal{F}(T)$. We achieve this goal through a two-level prompting approach using a large language model which we describe in detail below.

Phrase Chunking: Using our first-level prompts, we aim to augment the original text using phrase chunking to encourage global captions to concentrate on more specific regions. Visual grounding seeks to localize objects in images based on textual inputs. In contrast, global captions usually describe the entire image, typically describing several objects. While training on global captions could be beneficial for learning cross-modal information, it may lead the model to predict a broader region (*i.e.*, global context) rather than a specific region. Phrase chunking (*i.e.*, shallow parsing [55]) aims to identify continuous sequences of tokens representing syntactic units, enabling the extraction of phrases from unstructured text. We leverage an LLM to segment global captions into object-centric short phrases. During training, we use these image-chunk pairs instead of global captions, effectively guiding the model attention toward localized regions of interest. We *refer the reader to the supplementary material for prompting details and generated examples.*

Paraphrase Generation: Our SelfEQ approach involves feeding the model with pairs of textual descriptions that refer to the same underlying concept, with the expectation that the model can make similar predictions for these equivalent description pairs $\langle T, T^e \rangle$. We augment our dataset by transforming the region-based captions (*i.e.*, text that only refers to a region in the image) and the object-centric short phrases we obtained from phrase chunking into equivalent paraphrases referring to the same concept through our second-level LLM-prompts.

```
Q: "a smiling, holding something and standing man"
A: {"group": "man",
    "synonym": ["adult male","gentleman"],
    "antonym": ["woman"],
    "hypernym": ["adult","grownup","male","male person","human being","human"],
    "meronym": ["man body","man head","man arms","man legs"]}
Q: "a desk"
A: {"group": "desk",
    "synonym": ["table","writing desk"],
    "antonym": [],
    "hypernym": ["writer tool","furniture"],
    "meronym": ["desktop","desk leg"]}
```

Figure 3. Two samples from our LLM-prompt for paraphrase generation. The first set showcases an example of a region-based caption, and the second set shows a non-sentence phrase. **Q** is the query text and **A** is the expected answer.

There are many ways to paraphrase, including substituting words, altering sentence structures, and rewriting sentences based on semantics. However, considering that self-consistency in vision and language is relatively underexplored, we adopt a straightforward strategy: Replacing the primary object in the sentence while retaining all other attributes. This strategy yields several benefits. First, it provides a consistent context, which serves as a reference for the model to identify equivalent descriptions. This enables the equivalent relationships of paraphrases to be learned intuitively. Second, it simplifies prompt designing and post-processing by detecting the primary object and generating its synonym.

To generate paraphrases for the dataset consisting of region-based captions, we select four textual descriptions in which the primary noun plays different syntactic roles. We further select two non-sentence phrases as examples of query texts in our prompts. We show an example of a region-based caption and a non-sentence phrase in Figure 3. To design our prompt, we identify the primary object in the query text **Q**. Then we use WordNet [37] to obtain synonyms automatically and further remove inaccurate or invalid words. We add **A** to indicate the expected answer and include other relationships such as *antonym, hypernym* and *meronym* to provide richer contexts for LLM in-context learning. *Additional prompting details and paraphrase samples are provided in the supplementary material.*

This two-level prompt-based LLM augmentation approach ensures that our model is exposed to textual inputs that share the same concept while varying in linguistic representation, thereby promoting self-consistency and working vocabulary expansion.

4. Experimental Settings

Training. We use ALBEF [27] as our base model in all our experiments, given its reported off-the-shelf visual grounding performance via GradCAM [43]. ALBEF combines a ViT-B [11] model for encoding images and a BERT-base [10] model for encoding text. It is pre-trained on a range of datasets, including ImageNet-1K [42], Conceptual Captions [46], SBU Captions [39], MS-COCO [33], and Visual Genome (VG) [26] excluding box annotations. We finetune ALBEF with image-text pairs from VG and MS-COCO without any type of box supervision (i.e., no bounding boxes or object detectors), following prior work [2]. We further leverage Vicuna-13B [6] as our LLM-prompting model to generate the object-centric short phrases (via shallow parsing or chunking) and the equivalent paraphrases for our self-consistency data augmentation. Additionally, we validate the effectiveness of our SelfEQ tuning and self-consistency data augmentation method by training on a pre-processed subset of the Conceptual Captions 3M (CC3M) dataset [46], which contains many noisy or unaligned web-crawled AltText-image pairs. With this subset, we achieve an absolute improvement of 2.15% on Flickr30k, 3.32% on ReferIt, and 1.33% on RefCOCO+; *refer to the supplementary material for detailed CC3M experiments.*

	Method	Training	Flickr30k	ReferIt
Box Supervision	Align2Ground [7]	VG-boxes	71.00	-
	12-in-1 [35]	VG-boxes	76.40	-
	InfoGround [19]	VG-boxes	76.74	-
	VMRM [12]	VG-boxes	81.11	-
	AMC [51]	VG-boxes	**86.59**	**73.17**
Without Box Supervision	TD [56]	VG	42.40	31.97
	SSS [20]	VG	49.10	39.98
	MG-BiLSTM [2]	VG	57.91	62.76
	MG-ELMo [2]	VG	60.08	60.01
	GbS [3]	VG	73.39	62.24
	g [45]	VG	75.63	65.95
	g++ [44]	VG	79.95	**70.25**
	SelfEQ (ours)	VG	**81.90**	67.40
	FCVC [13]	MS-COCO	29.03	33.52
	MG-BiLSTM [2]	MS-COCO	53.29	47.89
	MG-ELMo [2]	MS-COCO	61.66	47.52
	GbS [3]	MS-COCO	74.50	49.26
	g [45]	MS-COCO	75.43	61.03
	g++ [44]	MS-COCO	78.10	61.53
	SelfEQ (ours)	MS-COCO	**84.07**	**62.75**

Table 1. Visual Grounding results on two benchmarks using pointing game accuracy with two training datasets. SelfEQ yields generally the best overall performance among weakly-supervised methods, and comes second to g++ on the ReferIt benchmark when trained using VG. We also show at the top the results of methods using additional box supervision from Visual Genome (VG) either directly or through an object detector.

Evaluation. We conduct evaluations using Flickr30k [40] and ReferIt [24] under pointing game accuracy following previous weakly-supervised visual grounding works [2, 3]. To underscore the competitive edge of our method, we also

Method	Box Supervision	RefCOCO+	
		Test A	Test B
InfoGround [19]	Yes	39.80	41.11
VMRM [12]	Yes	58.87	50.32
AMC [51]	Yes	**80.34**	**64.55**
ALBEF [27]	No	69.35	53.77
SelfEQ (ours)	No	**75.10**	**55.49**

Table 2. Results on RefCOCO+ pointing game accuracy. SelfEQ shows significant improvements over off-the-shelf ALBEF and competitive results compared to box-supervised methods.

present its performance on RefCOCO+ [53], a challenging benchmark more commonly used for testing box-supervised methods [7, 12, 19, 35, 51].

4.1. Implementation Details

Our experiments are conducted on a single computing node with 8 NVIDIA A40 GPUs. During the training phase, input images are resized to 256×256 and augmented with horizontal flipping, color jittering, and random grayscale conversion. We set up an Adam optimizer [25] with a learning rate of 1e-5 and a batch size of 448 across all experiments. We empirically set the RoI threshold k to 0.8 and the hyperparameter λ to 1.0. For training with raw image-text pairs from the datasets, we employ the vision-language objective $\mathcal{L}_{\text{vl}}$ (Sec. 3.1), while for the subset with equivalent paraphrases, we use our self-consistency equivalence tuning objective $\mathcal{L}_{\text{SelfEQ}}$ (Sec. 3.2) and corresponding vision-language objective $\mathcal{L}^{e}_{\text{vl}}$. The composite objective function is given by $\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{vl}} + (1-\alpha)\cdot(\mathcal{L}_{\text{SelfEQ}} + \mathcal{L}^{e}_{\text{vl}})$, where α is initially set to 0 and increments to 1, remaining constant after the second epoch. Our hyperparameter values and schedules were determined empirically on a small validation subset.

5. Experimental Results

Our resulting model obtains the best performance on the task of weakly-supervised visual grounding compared to most methods under this setting and is comparable to several prior works that rely on some of box supervision. Moreover, our qualitative results show that our method can handle paraphrases and a larger working vocabulary without the needed to increment the training dataset significantly.

Flickr30k and ReferIt. We evaluate the effectiveness of our proposed SelfEQ method in Table 1, demonstrating its substantial lead over GradCAM-based weakly-supervised approaches. Our self-consistency equivalent tuning adapts well for both region-based (*i.e.,*, VG) and global-based (*i.e.,*, COCO) image-text pairs, yielding a performance gain of 4.69% on Flickr30k and 7.68% on ReferIt, compared to

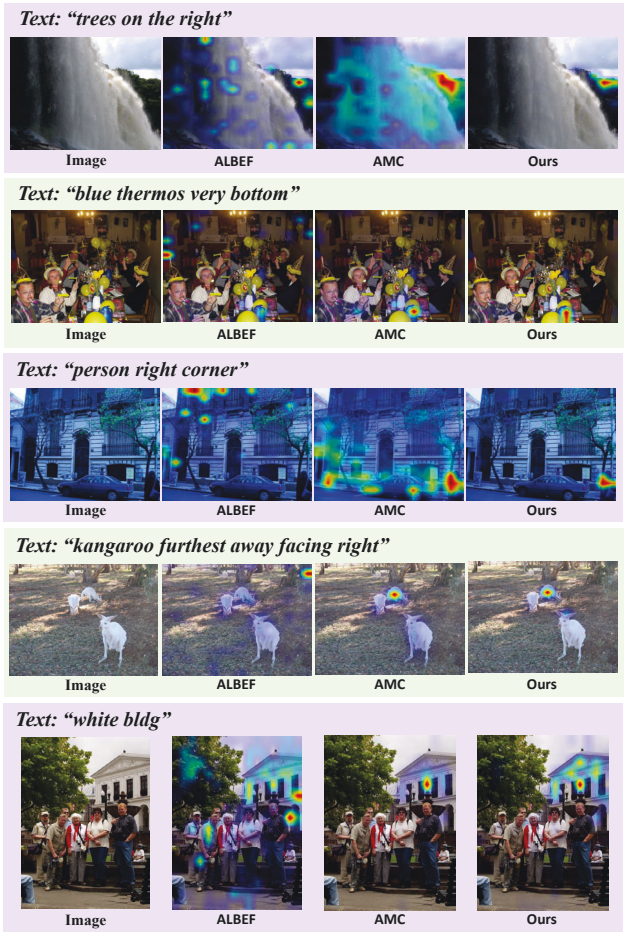

Figure 4. Qualitative results of our method in challenging visual grounding scenarios compared to prior works. On top of each row we show the reference text, the first column shows the image, then we show our base model ALBEF, the SotA box-supervised method AMC, and finally we show results with our method SelfEQ.

our base model ALBEF (see first row in Table 3). Notably, our method outperforms almost all box-supervised methods on Flickr30k [7, 12, 19, 35]. In the weakly-supervised setting, our method only comes second on ReferIt compared to g++[44] when trained on Visual Genome image-text pairs. This method leverages a custom architecture to produce a mask and uses heatmap supervision from a CLIP [41] model as pseudo-labels during training. We posit that our contribution is orthogonal and our approach would likely also benefit from similar supervision, since CLIP is trained in a much larger image-text pair dataset. Despite differences, our method still obtains higher performance when trained on MS-COCO and the best performance compared to all weakly-supervised methods on Flickr30K region-phrase grounding.

RefCOCO+. RefCOCO+ [53] serves as a rigorous benchmark for visual grounding, typically used to evaluate box-supervised techniques. In Table 2, we present the per-

Data	Objective	RefCOCO+		Flickr30k	ReferIt
		Test A	Test B		
-	$\mathcal{L}_{\text{vl}}$	69.35	53.77	79.38	59.72
T	$\mathcal{L}_{\text{vl}}$	72.30	54.22	78.75	65.86
$T + T^e$	$\mathcal{L}_{\text{vl}}$	71.55	53.51	78.05	64.57
$T + T^e$	$\mathcal{L}_{\text{SelfEQ}}$	**75.10**	**55.49**	**81.90**	**67.40**

Table 3. Ablation studies on different ways to utilize extra equivalent paraphrased data. The first row is off-the-shelf ALBEF performance before tuning. T denotes the textual captions from the dataset, and T^e corresponds to the associated equivalent paraphrases. $\mathcal{L}_{\text{vl}}$ is the vision-language objective, and $\mathcal{L}_{\text{SelfEQ}}$ is our self-consistency equivalence tuning objective.

formance of our weakly-supervised method (VG trained) against box-supervised methods. Our results indicate that our approach is competitive without reliance on any form of box annotations and significantly improves over the base ALBEF model.

Visual Grounding Analysis. Figure 4 provides qualitative results of our method in challenging scenarios, including occluded objects (row 1), small objects within complex scenes (row 2), objects partially shown in the corner of the image (row 3), multiple similar objects (row 4), and abbreviated text inputs (row 5). Our self-consistency equivalency tuning approach exhibits substantial improvements in the grounding capability of the base ALBEF [27] model. Remarkably, our approach even outperforms the state-of-the-art box-supervised method AMC [51] in multiple scenarios.

Self-Consistency Analysis. Figure 5 demonstrates qualitative results for the self-consistency capability across different equivalent paraphrases, encompassing terminology (row 1), synonym substitutions (row 2), and regional slang combining with different sentence structures (row 3). Although other methods succeed in localizing certain phrases, they demonstrate inconsistencies for the equivalent paraphrases. In contrast, our model finetuned with SelfEQ effectively establishes connections between semantically equivalent paraphrases, thereby enhancing the model self-consistency ability and potentially expanding its working vocabulary.

5.1. Ablation Studies

Data Quantity. We assess the impact of the data quantity of our generated paraphrases and compare our tuning strategy against standard vision-language objectives. We randomly sample portions of data from VG by 10% associated with our augmented equivalent paraphrases three times. To compare, we use the vision-language objective with VG text-image pairs as baselines. In Figure 6, we show the mean and standard deviation pointing game accuracy. The performance of the base vision-language objective does not exhibit a steady improvement with more text-image pairs. Although ReferIt performance increases, the performance on RefCOCO+ Test A remains mostly unchanged. Additionally, the performance on Flickr30k notably decreases, and there is a mixed effect on RefCOCO+ Test B, with half of the accuracy falling below the off-the-shelf ALBEF performance of 53.77%.

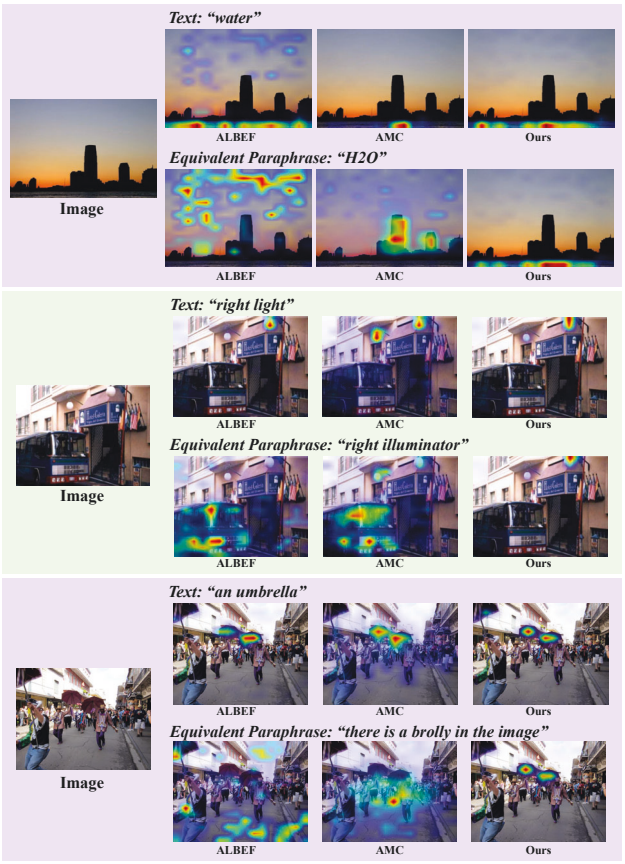

Figure 5. Qualitative results across equivalent paraphrases among different methods. For each image, we show a caption referring to an object in the first row and an equivalent paraphrase in the second row. Each column shows the results of ALBEF, the SotA box-supervised method AMC, and our SelfEQ method.

In contrast, SelfEQ consistently leads to performance enhancements with more equivalent paraphrases. Clear upward trends are observed on Flickr30k, ReferIt, and RefCOCO+ Test A as more data with corresponding equivalent paraphrases are added, meanwhile the gaps between the baselines generally widen. Notably, SelfEQ tuning maintains performance gains on Flickr30k, whereas the baselines performances drop. Although the trend on RefCOCO+ Test B is not consistently increasing, it is essential to emphasize that RefCOCO+ Test B is only a subset, and SelfEQ illustrates more stable and effective tuning performance on it, compared to the base vision-language objective.

These observations indicate that more equivalent paraphrases connecting with associated text phrases enable the

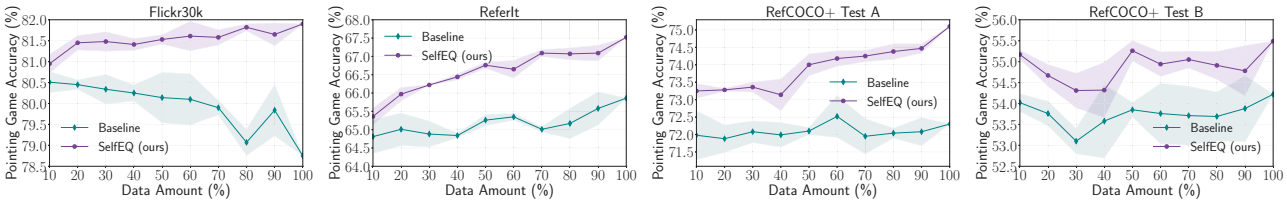

Figure 6. Tuning performance with different data quantities on Flickr30k, ReferIt, RefCOCO+ Test A and Test B. The purple and cyan lines represent SelfEQ (ours) and ALBEF baseline losses (vision-language objective), respectively. We show the impact of progressively augmenting captions via LLM-prompting for generating equivalent paraphrases tuned with our SelfEQ objective. Best viewed in color.

Format	Objective	Flickr30k	ReferIt
-	$\mathcal{L}_{\text{vl}}$	79.38	59.72
C	$\mathcal{L}_{\text{vl}}$	79.90	60.64
C	$\mathcal{L}_{\text{SelfEQ}}$	81.28	62.04
P	$\mathcal{L}_{\text{vl}}$	81.18	61.18
P	$\mathcal{L}_{\text{SelfEQ}}$	**84.07**	**62.75**

Table 4. Comparisons on data augmentation strategy for global-based captions in MS-COCO with or without the paraphrases. C is the global-based captions from MS-COCO, and P is our Vicuna-13B processed object-centric phrases separately. The first row is the off-the-shelf ALBEF performance before tuning.

model to acquire more valuable information during tuning. SelfEQ proves to be an effective and robust strategy for consistently improving performance with our generated paraphrases. With increased self-consistency augmented data, SelfEQ guides the model toward better grounding performance by enhancing its self-consistency capabilities.

Our method generates equivalent paraphrases for self-consistency enhancement, but it also contributes additional data for training. To ascertain the specific impact of our SelfEQ tuning strategy, we run a control experiment on this variable. As shown in Table 3, we assess the model's performance when equivalent paraphrases are integrated as regular image-text pairs with vision-language objectives (Sec. 3.1). This comparison reveals that merely augmenting the dataset with extra image-paraphrase pairs, without forming explicit linkages between the original text and its paraphrases, does not yield performance improvements.

Data Augmentation. For global-based captions in MS-COCO, we preprocess the captions C to object-centric short phrases P via LLM-prompting. As shown in Table 4, tuning with phrases P leads to better performance, benefiting both the vision-language objective ($\mathcal{L}_{\text{vl}}$) and our self-consistency equivalence tuning objective ($\mathcal{L}_{\text{SelfEQ}}$). This improvement is probably attributed to short phrases allowing the model to focus on a specific region rather than the entire scene, aligning more closely with the objective of visual grounding. By utilizing equivalent paraphrases with our SelfEQ objective (row 3 and 5), phrase chunking helps SelfEQ even more, indicating the important role of equivalent paraphrases in promoting self-consistency and grounding ability.

$\mathcal{L}_{\text{sim}}$	$\mathcal{L}_{\text{cst}}$	RefCOCO+		Flickr30k	ReferIt
		Test A	Test B		
✓		66.42	47.21	68.26	55.96
	✓	73.33	55.88	80.94	66.57
✓	✓	**75.10**	**55.49**	**81.90**	**67.40**

Table 5. Ablation studies on objective components of self-consistency equivalence tuning objective $\mathcal{L}_{\text{SelfEQ}}$.

Objective. Table 5 evaluates each component within our self-consistency equivalence tuning objective. The $\mathcal{L}_{\text{sim}}$ loss targets pixel-wise similarity, ensuring that maps for a caption and its equivalent paraphrase are identical. However, focusing solely on pixel-level similarity may neglect the precise spatial positioning of objects. To address this, the $\mathcal{L}_{\text{cst}}$ loss is proposed to identify the most likely correct object positions within the two maps (*i.e.*, RoI). It then encourages the model to yield consistently high attention scores within the RoI. By integrating both $\mathcal{L}_{\text{sim}}$ and $\mathcal{L}_{\text{cst}}$, self-consistency equivalence tuning objective fosters the model to not only align global similarities but also to pinpoint accurate object locations through the mutual supervision provided by a caption and its paraphrase, thereby enhancing self-consistency and accuracy.

6. Conclusion

We propose a novel weakly-supervised tuning approach coupled with a data augmentation strategy to enhance the localization capabilities of an image-text supervised model through self-consistency. Using an open-source LLM, we expand a dataset with equivalent paraphrases tailored to be object-centric. The augmented data is used to finetune a base model employing our novel self-consistency equivalence tuning objective. Our approach has been rigorously validated across pretraining on diverse datasets, ranging from region-based captions (*i.e.*, VG) to global-based captions (*i.e.*, COCO). Our method achieves superior and self-consistent performance on three benchmarks and is even competitive with some box-supervised methods.

Acknowledgments. This work was partially funded by NSF Award #2201710 and a Ken Kennedy Institute's SLB PhD Fellowship.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020. 2

[2] Hassan Akbari, Svebor Karaman, Surabhi Bhargava, Brian Chen, Carl Vondrick, and Shih-Fu Chang. Multi-level multimodal common semantic space for image-phrase grounding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12476–12486, 2019. 2, 5

[3] Assaf Arbelle, Sivan Doveh, Amit Alfassy, Joseph Shtok, Guy Lev, Eli Schwartz, Hilde Kuehne, Hila Barak Levi, Prasanna Sattigeri, Rameswar Panda, et al. Detector-free weakly supervised grounding by separation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1801–1812, 2021. 2, 5

[4] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 782–791, 2021. 2

[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 *European conference on computer vision*, pages 104–120. Springer, 2020. 1, 2, 3

[6] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 2, 5

[7] Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2601–2610, 2019. 2, 5, 6

[8] Chaorui Deng, Qi Wu, Qingyao Wu, Fuyuan Hu, Fan Lyu, and Mingkui Tan. Visual grounding via accumulated attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7746–7755, 2018. 2

[9] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1769–1779, 2021. 2

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 5

[11] 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. 5

[12] Zi-Yi Dou and Nanyun Peng. Improving pre-trained vision-and-language embeddings for phrase grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6362–6371, 2021. 1, 2, 5, 6

[13] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482, 2015. 5

[14] Yuting Gao, Jinfeng Liu, Zihan Xu, Jun Zhang, Ke Li, and Chunhua Shen. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. *arXiv preprint arXiv:2204.14095*, 2022. 1

[15] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision*, pages 540–557. Springer, 2022. 1

[16] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan A Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining. *arXiv preprint arXiv:2205.14459*, 2022. 1

[17] Eyal Gomel, Tal Shaharbany, and Lior Wolf. Box-based refinement for weakly supervised and unsupervised localization tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16044–16054, 2023. 2

[18] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 1

[19] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In *European Conference on Computer Vision*, pages 752–768. Springer, 2020. 1, 2, 5, 6

[20] Syed Ashar Javed, Shreyas Saxena, and Vineet Gandhi. Learning unsupervised visual grounding through semantic self-supervision. *Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI)*, 2019. 2, 5

[21] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 1

[22] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 2021. 1

[23] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. 1, 2

[24] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in pho-

tographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 2, 5

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[26] 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*, 123:32–73, 2017. 2, 5

[27] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 2, 3, 5, 6, 7

[28] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 1

[29] 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 preprint arXiv:1908.03557*, 2019. 3

[30] 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

[31] 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

[32] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021. 1

[33] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2, 5

[34] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 3

[35] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10437–10446, 2020. 2, 5, 6

[36] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7086–7096, 2022. 1

[37] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 5

[38] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern recognition*, 65:211–222, 2017. 2

[39] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011. 5

[40] 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. 2, 5

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 6

[42] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 5

[43] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 1, 2, 3, 5

[44] Tal Shaharabany and Lior Wolf. Similarity maps for self-training weakly-supervised phrase grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6925–6934, 2023. 2, 5, 6

[45] Tal Shaharabany, Yoad Tewel, and Lior Wolf. What is where by looking: Weakly-supervised open-world phrase-grounding without text inputs. *Advances in Neural Information Processing Systems*, 35:28222–28237, 2022. 2, 5

[46] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 2, 5

[47] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022. 1

[48] Josiah Wang and Lucia Specia. Phrase localization without paired training examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4663–4672, 2019. 2

[49] Fanyi Xiao, Leonid Sigal, and Yong Jae Lee. Weakly-supervised visual grounding of phrases with linguistic structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5945–5954, 2017. 2

[50] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022. 1

[51] 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, 2, 5, 6, 7

[52] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021. 1

[53] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016. 6

[54] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021. 1

[55] Feifei Zhai, Saloni Potdar, Bing Xiang, and Bowen Zhou. Neural models for sequence chunking. In *Proceedings of the AAAI conference on artificial intelligence*, 2017. 4

[56] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018. 5