Improving Visual Grounding by Encouraging Consistent Gradient-based Explanations

Ziyan Yang
Rice University
zy47@rice.edu

Kushal Kafle
Adobe Research
kkafle@adobe.com

Franck Dernoncourt
Adobe Research
dernonco@adobe.com

Vicente Ordonez
Rice University
vicenteor@rice.edu

Abstract

We propose a margin-based loss for tuning joint vision-language models so that their gradient-based explanations are consistent with region-level annotations provided by humans for relatively smaller grounding datasets. We refer to this objective as Attention Mask Consistency (AMC) and demonstrate that it produces superior visual grounding results than previous methods that rely on using vision-language models to score the outputs of object detectors. Particularly, a model trained with AMC on top of standard vision-language modeling objectives obtains a state-of-the-art accuracy of 86.49% *in the Flickr30k visual grounding benchmark, an absolute improvement of* 5.38% *when compared to the best previous model trained under the same level of supervision. Our approach also performs exceedingly well on established benchmarks for referring expression comprehension where it obtains 80.34% accuracy in the easy test of RefCOCO+, and 64.55% in the difficult split. AMC is effective, easy to implement, and is general as it can be adopted by any vision-language model, and can use any type of region annotations.*

1. Introduction

Vision-language pretraining using images paired with captions has led to models that can transfer well to an array of tasks such as visual question answering, image-text retrieval and visual commonsense reasoning [6, 18, 22]. Remarkably, some of these models are also able to perform visual grounding by relying on gradient-based explanations. While Vision-Language Models (VLMs) take advantage of the vast amounts of images and text that can be found on the web, carefully curated data with grounding annotations in the form of boxes, regions, or segments is considerably more limited. Our work aims to improve the grounding or localization capabilities of vision-language models further by tuning them under a training objective that encourages their gradient-based explanations to be consistent with human-provided region-based annotations from visually grounded data when those are available.

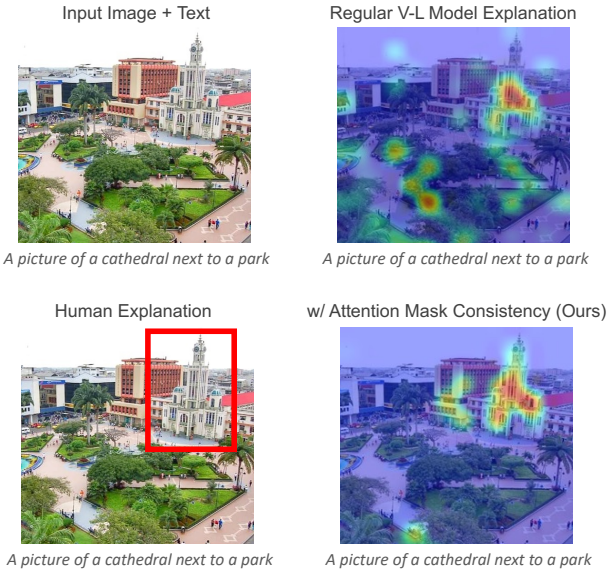

Figure 1. Gradient-based methods can generate heatmaps that explain the match between images and text for a Vision-language model (VLM). Our work aims to improve their ability to produce visual groundings by directly optimizing their gradient-based explanations so that they are consistent with human annotations provided for a reduced set of images.

Vision-language transformers extend the success of masked language modeling (MLM) to multi-modal problems. In vision-language transformers, objectives such as image-text matching (ITM), and image-text contrastive losses (ITC) are used in addition to MLM to exploit commonalities between images and text [6, 17, 18, 22]. We further extend these objectives to include our proposed Attention Mask Consistency (AMC) objective. Our formulation is based on the observation that gradient-based explanation maps obtained using methods such as GradCAM [30], can be used to explain the image-text matching of a VLM. Our AMC objective explicitly optimizes these explanations during training so that they are consistent with region annotations. Figure 1 illustrates an example input image and text

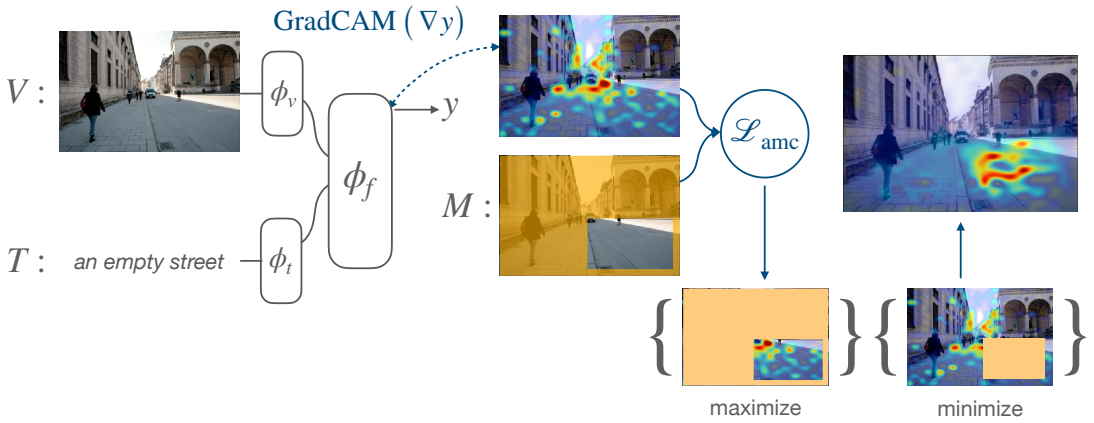

Figure 2. Overview of our method. Among other objectives, standard vision-language models are trained to produce a matching score y given an input image-text pair (V, T). For inputs containing an extra level of supervision in the form of region annotations (e.g. a triplet (V, T, M)), where M is a binary mask indicating the regions annotated by a human, we optimize the GradCAM [30] gradient-based explanations of the model so that the produced explanations are consistent with region annotations using $\mathcal{L}_{\text{amc}}$ by maximizing the energy in the heatmap that falls inside the region annotation and minimizing what falls outside. We accomplish this through soft margin losses as described in Sec. 3.2.

pair along with a gradient-based explanation obtained from a VLM model, a region annotation provided by a human, and an improved gradient-based explanation after the VLM model was tuned under our proposed objective.

Our work builds particularly upon the ALBEF model [17] which incorporates a vision-language model architecture based on transformers [36] and has already demonstrated off-the-shelf grounding capabilities using GradCAM. Gradient-based explanations in the form of heatmaps have been used extensively to explain the areas of the input images that most impact an output value of a model. In our formulation we actively leverage these heatmaps by designing a loss function that encourages most of the energy in the heatmaps to fall within the areas of the input image that most align with human provided region annotations. Figure 2 shows a detailed overview of our method and objective function. Given an input image and text pair, our goal is to maximize a soft margin between the energy of the heatmap inside the region annotation and the energy of the heatmap outside the region annotation. A soft-margin is important since typical human region annotations in the form of boxes do not exactly outline objects of different shapes, and in many cases models should still be able to ground an input text with multiple regions across the image.

We compare AMC extensively against other methods that use the same level of supervision but instead use an object detector such as Faster-RCNN [10, 11, 17, 23]. Our method obtains state-of-the-art *pointing game accuracy* on both Flickr30k and RefCOCO+. Our contributions can be summarized as follows: (1) We introduce a new training objective, AMC, which is effective, simple to implement and can handle multiple types of region annotations, (2) We show that AMC can improve the grounding capabilities of an existing vision-language model – ALBEF, and (3) the resulting model is state-of-the-art in two benchmarks for phrase grounding and referring expression comprehension.

2. Related Work

Vision-Language Representation Learning. Followed by the success of pretraining methods in NLP such as BERT [8], many transformer-based image-text models have been proposed to leverage benefits of pretraining on large-scale unlabeled image-text pairs [13, 18, 20]. While earlier pretraining methods rely on an object detector to divide an image into input tokens, some recent works, such as ALBEF [17], use an end-to-end vision transformer [9]. These pretrained models can then be finetuned to obtain impressive performance in a wide variety of vision-language tasks, such as image-text retrieval, visual question answering, and visual commonsense reasoning. While these models can perform some visual grounding by running them on the outputs of an object detector or using gradient-based explanations, they are not trained to take advantage of grounded data. Our AMC objective provides this additional capability by leveraging gradient-based explanations that can be easily obtained for a large variety deep learning models.

Gradient-based Localization. Localizing the most discriminative areas of an image for a given task has been widely used as a tool to provide visual explanation about a model. Class activation maps (CAM) [41] were proposed to provide weighted feature maps for any networks with minimal modifications to the model. Gradient-weighted Class Activation Mapping (GradCAM) [30] improves CAM by directly using gradients to obtain weighted feature maps

without the need for model modifications or retraining; The attention maps generated by these methods can be directly optimize to guide the model toward solutions that are more consistent with human-based annotations. Our proposed method is also based on GradCAM heatmaps. However, we use GradCAM during training to guide the generated heatmaps to achieve better consistency with known region and phrases that describe them. Recently, Pham et al [26] explored a similar idea by using segmentation masks to guide attention maps to focus on significant image regions for an attribute prediction task. Selvaraju et al [31] use saliency maps generated using DeepUSPS [24] at training time to guide attention maps in order to improve self-supervised representation learning. Similarly Pillai et al [27] rely on consistent explanations for generic representation learning using contrastive objectives. Our goal in using supervision on top of gradient-based heatmaps is to directly leverage these heatmaps to evaluate on visual grounding.

Visual Grounding Methods. Visual Grounding is a task that requires a model to select the region of an image described by a phrase. Several methods have been proposed to ground phrases to regions of an image, typically a bounding box [10, 11, 17, 23]. Visual grounding has also been used to improve performance on downstream tasks such as VQA [32]. These methods take advantage of object detectors which can provide high quality locations. The recently proposed GLIP model [19] incorporates an object detection model as part of its grounding objective, effectively combining vision-language pretraining with bounding box localization. Our work instead of outputting a box, optimizes its own gradient-based model explanations. Since our model does not output bounding boxes but heatmaps as an output it can generate more general groundings for phrases or objects that can not be mapped to a box such as stuff categories or references to multiple objects. Moreover, AMC can be used to improve an existing vision-language model such as ALBEF [17] without retraining from scratch. As vision-language models become larger and more robust, our proposed AMC objective can be readily applied.

3. Method

Vision-language pretraining consists of exploiting the structure of each input modality as well as their interactions. Our base model consists of three transformer encoders [8, 36]: An image encoder ϕ_v, a text encoder ϕ_t, and a multimodal fusion encoder ϕ_f. An input image V is encoded into a sequence of visual tokens $\{\mathbf{v}_{\text{cls}}, \mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_n\}$ and the text encoder encodes the input text T as a sequence of tokens $\{\mathbf{t}_{\text{cls}}, \mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_m\}$, where $\mathbf{v}_{\text{cls}}$ and $\mathbf{t}_{\text{cls}}$ are the embeddings of the `[CLS]` token for each transformer respectively. For each image-text pair drawn from a dataset $(V, T) \sim D$, a binary variable y represents whether the pair correspond with each other, i.e, whether the text actually describes the paired image. However, for some images a triplet $(V, T, M) \sim D$ might be available, where M additionally contains a region annotation, in the form of a binary mask, indicating the part of input image V that text T describes. In the following section we describe standard objectives used to capture intra-modality and inter-modality structure (Sec. 3.1), and then we describe our attention mask consistency objective (Sec. 3.2).

3.1. Standard Model Training Objectives

Masking Language Modeling (MLM) Originally introduced by BERT [8] in the context of language transformers, this objective has been adapted to multiple vision-language pretraining models such as [6, 18, 22] and is inspired by a long history in NLP of exploiting distributional semantics. The goal is to capture structure in the text by forcing the model to infer missing words from the input text. Each token in the input text is masked randomly with a small probability (usually 15%) and the model is then optimized to recover the masked tokens using information from both the remaining input text and the input image. Assume an input masked text is represented by T^{-m}, and the masked token is represented as a one-hot vector $\mathbf{t}^m$, the objective will be expressed as:

$$\mathcal{L}_{\text{mlm}} = \mathbb{E}_{(V,T^{-m})\sim D} \, \text{H}\left(\mathbf{t}^m, \phi_f^m\left(\phi_v(V), \phi_t(T^{-m})\right)\right), \tag{1}$$

where $\text{H}(\cdot,\cdot)$ is the cross-entropy between the missing token $\mathbf{t}^m$ and a probability distribution over tokens output by a function ϕ_f^m which augments ϕ_f with a linear projection layer and softmax function over a corresponding output embedding. This objective is optimized over a large sample of choices for masked tokens and image-text pairs.

Image Text Matching (ITM) Another common objective inspired by BERT's next sentence prediction objective, consists of image text matching. The purpose of this loss is to push the model to learn if a text and an image are matched. The output of the `[CLS]` token will be used to generate the output for this objective by adding a linear layer and a softmax activation function. We denote this entire operation as ϕ_f^{cls}. The objective is therefore defined as follows:

$$\mathcal{L}_{\text{itm}} = \mathbb{E}_{(V,T)\sim D} \, \text{H}\left(\mathbf{y}, \phi_f^{\text{cls}}\left(\phi_v\left(V\right), \phi_t\left(T\right)\right)\right), \tag{2}$$

where $\mathbf{y}$ is a one-hot vector with two entries $[y, 1-y]$ indicating whether the drawn sample (V, T) corresponds to a matching image-text pair or not.

Image-Text Contrastive Loss (ITC) This objective has been useful in weakly supervised grounding [11, 17]. We

follow ALBEF because it uses momentum distillation to potentially leverage a larger amount of negative image-text pairs. Assuming that each image-text pair is considered within a sample batch of K image-text pairs, this loss is defined as follows:

$$\mathcal{L}_{\text{itc}} = \mathbb{E}_{(V,T)\sim D} \frac{1}{2}\Bigg[\mathrm{H}\Big(\mathbf{y}, s(V,T)/\sum_{k=1}^{K} s(V,T_k)\Big) + \mathrm{H}\Big(\mathbf{y}, s(T,V)/\sum_{k=1}^{K} s(T,V_k)\Big)\Bigg], \tag{3}$$

where $s(V,T) = \exp(\phi_v(V)\cdot\phi_t(T)/\tau)$ computes a score between the output `[CLS]` token representations for the encoder transformer of each modality and τ is a temperature parameter, and $s(T,V)$ is defined similarly. The goal of this loss is to push for matching image-text pairs to have a closer representation than any non-matching image-text pair.

3.2. Attention Map Consistency (AMC)

In this section we explain in detail our proposed attention map consistency loss. Our proposed loss relies on first producing explanation heatmaps or "attention maps" using the GradCAM method [30]. In the context of vision-language transformers, this method can be used to highlight regions in the image that contribute to an image matching to an arbitrary input text, e.g., given an input image such as the one in Fig. 2, and an input text such as *an empty street*, we can generate a GradCAM visualization of areas in the input image that contribute to their matching score using $\mathcal{L}_{\text{itm}}$.

We assume that for a subset of images in our dataset we can obtain a triplet (V,T,M) where $M \in \{0,1\}^2$ is a binary mask such that $M_{i,j}$ is 1 if the location i,j is inside region or 0 otherwise, V is the input image, and T is an input text describing region M. This assumption is generally fair in comparison to previous works that instead leverage images annotated with labels and bounding boxes to train an object detector. In our case, we can easily support this setup by turning a label annotation, e.g., `dog` into a region textual caption by prompt engineering, e.g., *an image of a dog*. However, our binary masks are not restricted to being boxes.

In order to compute a GradCAM heatmap, we first extract an intermediate feature map F_z in the multimodal fusion transformer ϕ_f and denote this function as ϕ_z:

$$F_z = \phi_z\left(\phi_v(V),\ \phi_t(T)\right). \tag{4}$$

Then, we calculate the gradient of F_z with respect to the matching loss $\mathcal{L}_{\text{itm}}$ of this individual sample:

$$G_z = \nabla \mathrm{H}\left(\mathbf{y}, \phi_f^{\text{cls}}\left(\phi_v\left(V\right), \phi_t\left(T\right)\right)\right). \tag{5}$$

Next, we calculate a GradCAM attention heatmap A using F_z and G_z as follows:

$$A = \mathrm{ReLU}(F_z \odot G_z), \tag{6}$$

where $\odot$ is an element-wise multiplication. This heatmap is resized to the resolution of input images, and identifies which area in the image explains the model decision for its matching score.

The next step is to leverage the region annotations M so that the model focuses its heatmap scores in A inside the region of interest indicated by M. We first propose $\mathcal{L}_{\text{mean}}$ where we optimize a max margin loss so that the mean value of the heatmap inside of the region of interest is larger than the mean value of the heatmap outside as follows:

$$\mathcal{L}_{\text{mean}} = \mathbb{E}_{(V,T,M)\sim D}\Bigg[\max(0,\ \frac{1}{N^c}\sum_{i,j}(1-M_{i,j})\,A_{i,j} - \frac{1}{N}\sum_{i,j} M_{i,j}A_{i,j} + \Delta_1)\Bigg], \tag{7}$$

where Δ_1 is a margin term, and $N = \sum_{i,j} M_{i,j}$ is the number of locations inside the region of interest and N^c is the number of locations outside i.e. $\sum_{i,j}(1-M_{i,j})$. This loss aims to ensure that the attention map A contains most of the scores inside the region M subject to this margin. We also propose to jointly maximize the margin between the largest score inside the region of interest M and the largest score outside the region of interest by a margin Δ_2 as follows:

$$\mathcal{L}_{\text{max}} = \mathbb{E}_{(V,T,M)\sim D}\Bigg[\max(0,\ \max_{i,j}\left((1-M_{i,j})\,A_{i,j}\right) - \max_{i,j}\left(M_{i,j}A_{i,j}\right) + \Delta_2)\Bigg]. \tag{8}$$

Finally, we combine these two objectives:

$$\mathcal{L}_{\text{amc}} = \lambda_1\cdot\mathcal{L}_{\text{mean}} + \lambda_2\cdot\mathcal{L}_{\text{max}}, \tag{9}$$

where λ_1 and λ_2 are empirically determined weighting coefficients. We demonstrate in our experimental section that this objective effectively encourages model explanations that provide better grounding support for tasks such as referring expression comprehension and visual grounding.

4. Experiments

In this section, we describe our training setup and experimental evaluations. Our evaluations revolve around tasks that require pointing to the location in an image that is referred by an input text.

4.1. Training Details

Our model follows the architecture and training objectives of the ALBEF model [17] which uses the ALBEF-14M dataset as source of pretraining. ALBEF-14M is

Method	Detector	Flickr30k	RefCOCO+	
			test A	test B
Align2Ground [7]	Faster-RCNN (VG)	71.00	-	-
12-in-1 [23]	Faster-RCNN (VG)	76.40	-	-
InfoGround [11]	Faster-RCNN (VG)	76.74	39.80	41.11
VMRM [10]	Faster-RCNN (VG)	81.11	58.87	50.32
AMC∗	–	86.49	78.89	61.16
AMC (ours)	–	**86.59**	**80.34**	**64.55**

Table 1. Visual Grounding results using *pointing game* accuracy against the state-of-the-art for methods. For fairer comparisons, AMC∗ indicates a version of our model restricted to using only the box and label annotations from Visual Genome (VG) that were used to train the Faster-RCNN network used in the other methods.

a large image-text data collection including the following datasets: COCO [21], Visual Genome (VG) [16] (excluding box annotations), SBU [25], CC3M [34] and CC12M [4]. In this data collection, each image is paired with one or several image descriptions so that we can sample pairs $(V, T) \sim D$. Additionally, several vision-language transformer models such as UNITER [6] or VisualBERT [18] further leverage box annotations from the Visual Genome dataset. We use this dataset as an additional source of triplets $(V, T, M) \sim D$. We start with the ALBEF model and further finetune it for 20 more epochs on Visual Genome with boxes using our proposed AMC loss. Next, we describe in detail how we leverage the Visual Genome dataset to produce triplets (V, T, M) in more detail.

First, we provide a more detailed description of the Visual Genome dataset. This dataset consists of $108,077$ images and annotations in multiple formats such as boxes + region descriptions, boxes + object labels, and boxes + object attributes. At a first level, annotators of this dataset were asked to provide text that describes a region of the image and to provide a bounding box that covers the region. For instance, *a brown dog playing with a ball*. Then, the region descriptions were shown to other annotators that were asked to select objects from these regions and provide tight bounding boxes for selected objects and attributes e.g. *brown dog* and *ball*. Region bounding boxes and object+attribute bounding boxes are different for this dataset. The object detector trained by Anderson et al [2] on the object bounding boxes and object attributes of Visual Genome has been used by several previous visual grounding models. In order to compare fairly to these methods, we develop a model using the same training split as [2] and conduct experiments without the use of region descriptions. For completeness, we also conduct experiments using both boxes with attributes and boxes with region descriptions.

We construct textual descriptions for object bounding boxes using prompt engineering templates. For example, if an image contains an object *dog* with an attribute *brown*, we construct the description as *a brown dog*. We filter out bounding boxes smaller than 8% of the whole image. To further increase the localization capabilities of our method, we generate prompts with spatial references. For images with objects that correspond to more than one box, we select the leftmost/rightmost, top/bottom boxes and assign more detailed prompts such as *[obj] on the left*, *[obj] on the right*, *top [obj]* and *bottom [obj]*. Moreover, if the box falls into a corner of the image, we further assign them another level of spatial information such as *top left*, *top right*, *bottom left* and *bottom right*.

We conduct experiments on single node with 8 NVIDIA A40 GPUs. All experiments use a batch size of 512 and a learning rate of 1e-5 with an Adam optimizer [15]. We determine empirically based on a small validation set two margin losses: $\Delta_1 = 0.1$ and $\Delta_2 = 0.5$ and determine our weighting coefficient for our losses as $\lambda_1 = 0.2$ and $\lambda_2 = 0.8$, respectively. For data augmentation, we resize images into a resolution of 256×256 and apply horizontal flips, color jittering and random grayscale conversions. Our code and data are publicly available[1].

4.2. Visual Grounding

Visual grounding consists in automatically associating an area of an image with an arbitrary piece of input text. A popular benchmark for this task is Flickr30k [28]. We only use the validation and testing splits for this dataset and do not use it for training. Each split includes a thousand images and is used for all of our model selections and evaluations. In Flickr30k Entities, each object phrase may pair with multiple ground truth boxes in the image. Our model will take the phrase and whole image as inputs, and find the most related regions corresponding to the phrase.

We report experimental results for Flickr30k Entities [28] with both detector-based and detector-free methods. *Pointing game* accuracy is a widely used metric in previous works for this task [1, 3, 11, 37], and we follow the same setting as in [1] to calculate this measure: After obtaining a heatmap given an input phrase and an image, we extract the position of the maximal point of this heatmap, and if this point falls in the target box, we count this result as positive. For detector-based methods, we follow [11] to calculate the *pointing game* accuracy by first ranking proposals generated by an object detector and then retaining one box proposal with highest score as the result. If the center point of the selected box proposal falls within the target box, this result is counted as positive.

For Align2Ground [7] and 12-in-1 [23], we directly show the results reported in [3]. For InfoGround [11] we use

[1] https://github.com/uvavision/AMC-grounding

Method	VG-Boxes	Backbone	Flickr30k
gALBEF [17]	no	ALBEF	79.14
GbS [3]	no	PNASNet	73.39
MG [1]	no	ELMo + PNASNet	67.60
GAE [5]	no	CLIP	72.47
WWbL [33]	no	CLIP + VGG	75.63
GbS+IG [3]	yes	PNASNet	83.40
GbS+12-in-1 [3]	yes	PNASNet	85.90
AMC (ours)	yes	ALBEF	**86.59**

Table 2. Visual Grounding results using *pointing game* accuracy against methods that do not use object detectors or Visual Genome box supervision, showing that box supervision still makes a significant difference on this benchmark despite the fact that CLIP uses hundreds of millions of extra images for training compared to the ALBEF backbone.

their provided trained models. For VMRM [10], since they do not provide their trained model, we re-train it using the official code and their used features and boxes from MMF [35] and MAF [38]. Align2Ground, 12-in-1 and VMRM all use image features generated by object detectors trained on VG boxes and attributes [2]. InfoGround uses image features extracted from an object detector trained on VG boxes. We also compare to methods that do not use any form of box supervision including our backbone ALBEF as a baseline. We refer as gALBEF to our baseline which only uses GradCAM in combination with ALBEF as described in [17]. We additionally compare to MG [1], GbS [3] which report results on Flickr30k. Results for GAE [5] as well as WWbL are taken directly from [33]. In addition to fairly compare with GbS, we additionally report their results when ensembled with detector-based methods InfoGround and 12-in-1.

Our main comparison results for methods relying on Visual Genome boxes are summarized in Table 1 and results comparing against methods that are weakly supervised and hence do not use box information are shown in Table 2.

4.3. Referring Expression Resolution

Referring expressions are textual descriptions that refer unambiguously to an object or region of an image. Users are explicitly prompted to write a textual description to refer to a specific object. However the setup is similar to the visual grounding setup and as such, many previous methods compare their results across both benchmarks. We adopt the same *pointing game* accuracy metric and compare our results against previous methods in two benchmark datasets: RefCOCO+ [14, 40] and ReferIt [14]. We compare against the same set of methods as in the visual grounding task except for Align2Ground [7] and 12-in-1 [23] which do not provide results for RefCOCO+. Additionally, InfoGround [11] does not report results for RefCOCO+, therefore, we use their provided bounding boxes for COCO images [21] to perform this evaluation.

We describe in more detail each benchmark. RefCOCO+ [40] is a widely used referring expression dataset including 20K images from the COCO dataset [21]. The expressions in RefCOCO+ were collected so that they do not allow words such as *left* or *right*, making it slightly more challenging. From this dataset, we only use its validation and testing splits. The testing split of this dataset is further divided into two subsets: `test A` and `test B`, in which the former only includes people as the target objects and the latter includes all objects. The total number of testing images is 1.5K. Results for referring expression comprehension on RefCOCO+ are included in Table 1.

4.4. Box Recall Evaluation

Pointing game accuracy has been previously used for both detector-based [11] and detector-free [1, 3] methods. However, another metric that can be considered is *Recall@k* from detector-based methods [10, 11]. For *Recall@k*, a model will rank all the box proposals generated by an object detector, and select the top-k boxes as results. If a selected box and the ground truth box have an intersection over union (IoU) ≥ 0.5, then the selected box will be counted as positive. Table 3 shows results when we evaluate our method by using it to choose boxes from different bounding box proposals methods by selecting the boxes with high attention heatmap scores. We use boxes generated by the FasterRCNN [29] from Gupta et al [11] and the MaskRCNN [12] from Yu et al [39]. Using the MaskRCNN proposals, our method obtains consistently better results than VMRM, which is the current stats-of-the-art. However, we find this metric is influenced by the quality of boxes. For example, using the MaskRCNN proposals will get much better results than using the FasterRCNN proposals for VMRM [10].

Method	Boxes	RefCOCO+	
		test A	test B
VMRM [10]	FasterRCNN	30.04	30.78
VMRM [10]	MaskRCNN	46.63	40.52
gALBEF [17]	MaskRCNN	61.70	42.83
AMC	MaskRCNN	**68.04**	**46.55**

Table 3. We show recall@1 results on the RefCOCO+ validation and testing sets to complement our results using pointing game accuracy.

Method	Overall	People	Animals	Vehicles	Instrum.	Bodyparts	Clothing	Scene	Other
MG [1]	69.2	75.6	87.6	83.8	57.5	44.9	58.3	68.2	59.8
GbS [3]	74.5	83.6	89.3	92.1	83.3	53.2	50.1	71.3	66.7
gALBEF [17]	79.1	80.1	89.8	89.8	83.3	63.3	85.5	83.8	70.2
Align2Ground [7]	71.0	-	-	-	-	-	-	-	-
12-in-1 [23]	76.4	85.7	82.7	**95.5**	77.4	33.3	54.6	80.7	70.6
InfoGround [11]	76.7	83.2	89.7	87.0	69.7	45.1	74.5	80.6	67.3
VMRM [10]	81.1	88.0	92.3	94.3	66.7	55.1	79.8	85.1	69.9
AMC	**86.6**	**89.7**	**95.2**	93.8	**86.4**	**69.8**	**89.0**	**91.4**	**77.7**

Table 4. Breakdown of results by category for *pointing game accuracy* on Flickr30K entities visual grounding.

4.5. Discussion of Results

Table 1 contains the main results of our paper when compared to several other methods that rely on vision-language models coupled with box-level supervision from Visual Genome through an object detector – FasterRCNN trained on Visual Genome. Our results show a large advantage on all benchmarks but especially on RefCOCO+. We report more fine-grained results in Table 4 for Flickr30K Entities. There are eight categories in this dataset. We evaluate on each category and report the pointing game accuracy for them separately. For MG and Gbs, we report results when trained on COCO because their models achieved their best performances on Flickr30k Entities under this setting. In general, our method obtains better results for almost all the categories. For the category `vehicle` our method obtains 93.8%, which is only 1.7% lower than the best result from the best method (12-in-1).

In Table 2 we observe that methods that use box supervision still exhibit considerable better performance on visual grounding on Flickr30k Entities. Our method obtains 86.59%, which is 10.6% higher than WWbL, which is the best method that does not use box supervision from Visual Genome – however in terms of number of training images it relies on CLIP which was trained on 400M images with text compared to our method which uses 14M images with text plus 100k images with boxes. In addition we compare to GbS [3] when ensembled with detector-based methods InfoGround and 12-in-1. Our method is still superior when compared to these two strong baselines.

In Figure 3 we show and compare visual explanations obtained by our model against those obtained by VMRM [10] and GradCAM heatmaps generated by gALBEF. The text input for VMRM is the whole caption and a phrase and it produces a bounding box prediction. The model locates the positions of the phrase in the caption and selects boxes corresponding to the phrase with context information. For gALBEF and our method, the text input is only a phrase. We found our method can get more accurate and more complete objects from phrases. For example, in the last row, our method can provide a more precise heatmap for the referred object for *traditional asian clothing* than gALBEF, and in this case, VMRM is confused by the clothing from the woman instead of the boy. Additionally, in the second row, when the model is asked to find "the guitar", our method can accurately cover the guitar, but gALBEF covers several unrelated regions that probably contribute to the detection of `guitar` but do not provide an explanation that aligns with what a human would annotate for this image. We provide more qualitative results in our supplementary material.

Data	Flickr30k	ReferIt	RefCOCO+	
			test A	test B
object boxes	86.49	62.65	78.89	61.16
region boxes	85.14	59.16	77.89	61.26
both	86.59	64.27	80.34	64.55

Table 5. We conduct an ablation study to evaluate the effect of using "box" annotations corresponding to box + object labels + object attributes from Visual Genome, and "region" annotations corresponding to region boxes + region descriptions from Visual Genome.

Loss	Flickr30k	ReferIt	RefCOCO+	
			test A	test B
$\mathcal{L}_{\text{cosine}}$	84.85	61.21	76.41	60.81
$\mathcal{L}_{\text{mean}}$	82.83	57.63	75.34	56.90
$\mathcal{L}_{\text{max}}$	86.56	62.79	80.34	64.47
$\mathcal{L}_{\text{amc}}$	86.59	64.27	80.34	64.55

Table 6. We conduct an ablation study to evaluate the contribution of $\mathcal{L}_{\text{max}}$ and $\mathcal{L}_{\text{mean}}$ to our final accuracy and an alternative loss based on cosine similarities $\mathcal{L}_{\text{cosine}}$.

Figure 3. Qualitative comparison of the generated explanations for various images and input phrases. First column: original images from Flickr30k Entities; in each colored area from left to right: bounding boxes selected by VMRM; heatmaps generated by gALBEF; heatmaps generated by our method. On the top of each group of images, we show the caption and target phrases.

4.6. Ablation Studies

In this section, we present ablations against several choices in our model and contributing factors. Particularly we investigate how large is the effect from box supervision from Visual Genome both from object boxes, and region boxes.

Box Supervision. As described in section 4.1, VG [16] includes regions with descriptions and objects with attributes. We evaluate our method on each separately. For Flickr30k Entities and the ReferIt dataset, boxes with generated descriptions using attributes lead to better results than regions with descriptions. We believe this is caused by accurate localization information provided by boxes and spatial information from box descriptions. For RefCOCO+ test B, regions with descriptions perform slightly better than boxes with attributes. By combining boxes, regions and two kinds of descriptions, we obtain better alignment between phrases and image subareas. Our full set of results from this experiment are in Table 5.

Loss Choices. Instead of calculating our margin loss as in Eq. 9, we calculate and minimize the cosine distance between M and A. Therefore, the generated heatmap will be closer to the box mask. Results for all of our choices that we considered in our objective function are presented in Table 6. For all datasets, our method outperforms this cosine distance loss $\mathcal{L}_{\text{cosine}}$, proving our method is a better way to use box information than a perhaps more straightforward dot product optimization. Furthermore, we evaluate two components in Eq. 9: $\mathcal{L}_{\text{mean}}$ and $\mathcal{L}_{\text{max}}$. We find $\mathcal{L}_{\text{max}}$ is very significant in AMC, but $\mathcal{L}_{\text{mean}}$ also provides complementary information, especially for the ReferIt dataset. In general, combining two terms leads to a more comprehensive grounding ability but using $\mathcal{L}_{\text{max}}$ alone is also very competitive.

5. Conclusion

In this paper, we proposed Attention Map Consistency (AMC). From the intuition that a model should focus on meaningful regions guided by location information, we design an objective function that optimizes gradient-based explanation maps. Our approach achieves superior results on visual grounding compared to other methods with a similar level of supervision. It particularly surpasses methods relying on an object detector pretrained on Visual Genome.

Acknowledgments This work was supported by gifts from Adobe Research and NSF Awards IIS-2221943 and IIS-2201710. We are also thankful for positive comments and suggestions from anonymous reviewers.

References

[1] 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. 5, 6, 7

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 5, 6

[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. 5, 6, 7

[4] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021. 5

[5] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 397–406, 2021. 6

[6] 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, 3, 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. 5, 6, 7

[8] 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. 2, 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. 2

[10] 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. 2, 3, 5, 6, 7

[11] 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. 2, 3, 5, 6, 7

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 6

[13] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020. 2

[14] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 6

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[16] 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, 8

[17] 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, 4, 6, 7

[18] 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. 1, 2, 3, 5

[19] 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. 3

[20] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 2

[21] 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, 6

[22] 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. 1, 3

[23] 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, 3, 5, 6, 7

[24] Tam Nguyen, Maximilian Dax, Chaithanya Kumar Mummadi, Nhung Ngo, Thi Hoai Phuong Nguyen, Zhongyu Lou, and Thomas Brox. Deepusps: Deep robust unsupervised saliency prediction via self-supervision. *Advances in Neural Information Processing Systems*, 32, 2019. 3

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

[26] Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava. Learning to predict visual attributes in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13018–13028, 2021. 3

[27] Vipin Pillai, Soroush Abbasi Koohpayegani, Ashley Ouligian, Dennis Fong, and Hamed Pirsiavash. Consistent explanations by contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10213–10222, 2022. 3

[28] 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. 5

[29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 6

[30] 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, 4

[31] Ramprasaath R Selvaraju, Karan Desai, Justin Johnson, and Nikhil Naik. Casting your model: Learning to localize improves self-supervised representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11058–11067, 2021. 3

[32] Ramprasaath R Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. Taking a hint: Leveraging explanations to make vision and language models more grounded. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2591–2600, 2019. 3

[33] Tal Shaharabany, Yoad Tewel, and Lior Wolf. What is where by looking: Weakly-supervised open-world phrase-grounding without text inputs. *arXiv preprint arXiv:2206.09358*, 2022. 6

[34] 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. 5

[35] Amanpreet Singh, Vedanuj Goswami, Vivek Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Mmf: A multimodal framework for vision and language research, 2020. 6

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 3

[37] Liwei Wang, Jing Huang, Yin Li, Kun Xu, Zhengyuan Yang, and Dong Yu. Improving weakly supervised visual grounding by contrastive knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14090–14100, 2021. 5

[38] Qinxin Wang, Hao Tan, Sheng Shen, Michael W Mahoney, and Zhewei Yao. Maf: Multimodal alignment framework for weakly-supervised phrase grounding. *arXiv preprint arXiv:2010.05379*, 2020. 6

[39] 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. 6

[40] 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. 6

[41] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 2