Investigating Compositional Challenges in Vision-Language Models for Visual Grounding

*Yunan Zeng [1,2,3] Yan Huang [1,2] Jinjin Zhang [3] Zequn Jie [3] Zhenhua Chai [3] †Liang Wang [1,2]
[1]Center for Research on Intelligent Perception and Computing (CRIPAC)
[2]Institute of Automation, Chinese Academy of Sciences (CASIA)
[3]Meituan
yunan.zeng@cripac.ia.ac.cn {yhuang, wangliang}@nlpr.ia.ac.cn
{zhangjinjin05, chaizhenhua}@meituan.com zequn.nus@gmail.com

Abstract

Pre-trained vision-language models (VLMs) have achieved high performance on various downstream tasks, which have been widely used for visual grounding tasks in a weakly supervised manner. However, despite the performance gains contributed by large vision and language pre-training, we find that state-of-the-art VLMs struggle with compositional reasoning on grounding tasks. To demonstrate this, we propose Attribute, Relation, and Priority grounding (ARPGrounding) benchmark to test VLMs' compositional reasoning ability on visual grounding tasks. ARPGrounding contains 11,425 samples and evaluates the compositional understanding of VLMs in three dimensions: 1) attribute, denoting comprehension of objects' properties; 2) relation, indicating an understanding of relation between objects; 3) priority, reflecting an awareness of the part of speech associated with nouns. Using the ARPGrounding benchmark, we evaluate several mainstream VLMs. We empirically find that these models perform quite well on conventional visual grounding datasets, achieving performance comparable to or surpassing state-of-the-art methods but showing strong deficiencies in compositional reasoning. Furthermore, we propose a composition-aware fine-tuning pipeline, demonstrating the potential to leverage cost-effective image-text annotations for enhancing the compositional understanding of VLMs in grounding tasks. Code is available at link.

1. Introduction

Vision-language models (VLMs) have achieved high performance on various downstream tasks, including many zero-shot learning and text-guided vision tasks [2, 4, 19, 21, 34, 41]. Many endeavors leverage the intrinsic visual-linguistic alignment representation within these models and integrate VLMs with explainability methods [37, 52] for image localization tasks. This approach serves to alleviate the significant time and cost required for dense manual annotation, particularly in tasks such as weakly supervised visual grounding [13, 20, 39, 40], weakly supervised segmentation [23, 25] and open-vocabulary segmentation [24, 46]. Visual grounding is a pivotal task in vision-language. We adopt the widely used Grad-CAM [37] algorithm, integrating it into VLMs without introducing additional complexity. Our methodology unveils that VLMs can attain state-of-the-art results in weakly supervised visual grounding tasks through the application of explainability techniques.

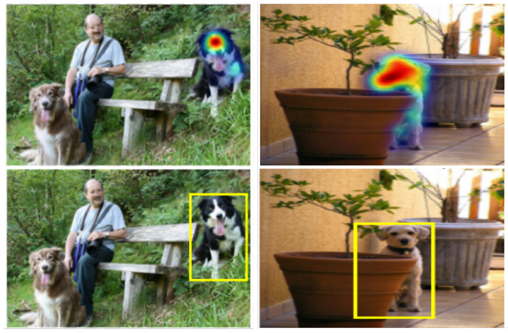

(a) brown dog (b) pot behind dog

Figure 1. Typical examples of testing compositional understanding of CLIP on visual grounding task. CLIP encounters challenges in discerning the authentic object from deceptive ones. (Left) CLIP is misled by a dog of a distinct color. (Right) CLIP is misled by an object within the phrase. Both instances suggest a deficiency in CLIP's grasp of compositional structure.

However, we find VLMs encounter challenges in grounding while compositional reasoning is involved. As shown in Figure 1, CLIP exhibits difficulty distinguishing

[*] Work was done during the author's internship at Meituan.
[†] Corresponding author.

between the “brown dog” and the “black dog”, and fails to ground the “pot behind dog” while falsely activating the dog area. To investigate this problem, we present ARPGrounding, a novel grounding benchmark designed to assess the fine-grained visio-linguistic compositionality across three dimensions: attribute, relation, and priority. We define two objects of compositional ambiguity in a single image, each corresponding to a semantic textual expression. We test models’ compositional understanding by compelling models to pick the correct object based on the provided text. Using ARPGrounding benchmark, we observe that while all of these VLMs can achieve performance levels comparable to state-of-the-art results on conventional visual grounding benchmarks, they struggle to perform beyond chance levels in relatively simple tasks that require compositional understanding. VLMs encounter difficulty differentiating between two objects with distinct attributes, falter in distinguishing relations, and fail to discern the primary and secondary targets within the text.

VLMs are pre-trained on extensive datasets featuring intricate scenes and detailed captions. These datasets are notable for their rich compositional structure. However, training on these datasets has been proven insufficient in addressing the dearth of compositional structure in VLMs. Recent research has revealed that the challenges faced by VLMs in compositional understanding stem from an artifact of their contrastive training objective, leading to shortcut learning while optimizing for the objectives [42, 49]. These shortcuts enable VLMs to get relatively good performance in downstream tasks but leave deficiencies in visio-linguistic tasks that demand a more nuanced understanding of objects. In response, we propose a novel composition-aware fine-tuning pipeline to fine-tune VLMs without using dense annotations. We first generate a text pair for each image where each component in the pair describes distinct objects using dependency parsing. Then we introduce a pretext task that is aimed at augmenting diversity within the grounding heatmaps of these two texts. This approach yields substantial improvements in the compositional understanding of VLMs.

Our main contributions are three-fold:

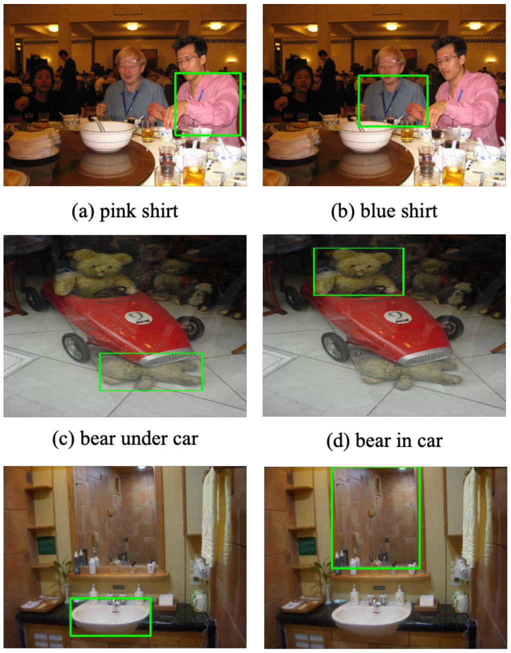
1. We employ explainability techniques to analyze the VLMs and attain state-of-the-art results in weakly supervised visual grounding tasks.
2. We present ARPGrounding, a novel grounding dataset designed to assess the fine-grained visio-linguistic compositionality. Using ARPGrounding, we discover VLMs’ deficiencies in compositional reasoning.
3. We propose a composition-aware fine-tuning pipeline. Empirical results show that this training pipeline effectively aids in enhancing compositional understanding of VLMs.

2. Related Work

Evaluating vision-language compositionality. In recent studies, researchers have introduced evaluation benchmarks for assessing the compositional abilities of vision-language models. These studies have revealed that current models exhibit a limited understanding of compositionality [28, 36, 43, 49, 51]. Specifically, VLMs exhibit limitations in accurately counting objects within an image [31], lack proficiency in commonsense knowledge and reasoning abilities [47, 48], falter in comprehending verbs [10, 30], encounter difficulties in integrating objects with their attributes [36], face challenges in understanding spatial relations [14, 35], and demonstrate insensitivity to word order [42]. Our objective is also to evaluate vision-language compositionality. However, we employ a more fine-grained grounding task instead of a text-image matching task to explicitly assess whether the model correctly associates attributes and relations with the objects.

Weakly supervised visual grounding. Many prior methods [5, 9, 44] for weakly supervised visual grounding rely on pre-trained object detectors for Region of Interest (ROI) localization. These methods typically project text and ROIs into a joint visual-textual embedding space, transforming the grounding task into a retrieval task. However, detector-based approaches neglect the holistic context of the entire image and encounter challenges related to transfer and generalization when dealing with objects in distinct domains or belonging to different classes. In contrast, detector-free methods address this issue by forgoing the use of object detectors. These methods perform dense localization for the given query phrases, thereby generating grounding heatmaps instead of ranking ROIs. Some earlier works [1, 3, 12, 45, 50] define and optimize auxiliary tasks using weakly supervised data. While these auxiliary tasks may not be identical to the grounding objective, optimizing them yields the desired results in visual grounding. More recent research [38, 39] has leveraged VLMs trained on large-scale visual-text alignment datasets. These approaches use VLMs’ explainability maps as pseudo-labels to train grounding models, resulting in state-of-the-art performance in the field.

Gradient-based localization. The technique of pinpointing the most distinctive regions within an image for a specific task serves as a prevalent approach to explain a model’s decision-making process visually. Class activation maps (CAM) [52] was introduced as a technique to yield weighted feature maps for various networks with minimal adjustments to the model. A subsequent enhancement, Gradient-weighted Class Activation Mapping (Grad-CAM) [37], further improves upon CAM by directly employing gradients to produce weighted feature maps without necessitating model modifications or retraining. The grounding heatmaps generated by these methods can be directly



(a) pink shirt

(b) blue shirt

(c) bear under car

(d) bear in car

(e) bathroom sink under bathroom mirror

(f) bathroom mirror above bathroom sink

Figure 2. Examples from our ARPGrounding dataset, wherein we primarily evaluate models' proficiency in comprehending three types of compositionality. Each sample in ARPGrounding consists of an image containing two intricately distinguishable objects. (a) and (b) exemplify a sample of attribute composition, (c) and (d) illustrate an example of relation composition, while (e) and (f) demonstrate an example of priority composition. We test whether the model can pick the correct object according to the text.

optimized to guide the model toward solutions that align more closely with human-based annotations. Our proposed method is also based on Grad-CAM heatmaps. We apply the Grad-CAM algorithm to VLMs, enabling them to localize objects in images based on textual semantics. We leverage Grad-CAM for fine-tuning VLMs to guide the generated heatmaps to be more distinguishable with regard to samples that have different compositionality.

3. ARPGrounding Benchmark

In this section, we describe how ARPGrounding benchmark is constructed and how we use it to evaluate compositional grounding performance of VLMs.

3.1. Dataset

In order to construct a compositional grounding dataset, we require images containing object bounding box annotations, along with corresponding attribute and relation annotations. Therefore we establish ARPGrounding benchmark upon Visual Genome (VG) [16], a large-scale dataset comprising over 100,000 images, annotated with objects, attributes, and relations. Additionally, we manually filter the dataset when necessary to ensure data quality.

As shown in Figure 2, ARPGrounding contains three types of compositionality: attributes, relations, and priorities. Each sample in ARPGrounding consists of an image containing two objects selected with bounding boxes. These paired objects serve as mutual compositional distractors. Each sample includes distinct text corresponding to these two objects, and these texts can be differentiated based on attributes, relations, or priorities.

Attribute. Our primary goal is to discern two objects within a single image. These objects belong to the same category but possess distinct attributes. Our objective is to assess the model's ability to accurately identify the correct object based on the provided textual information. To locate such object pairs, we initiate the process by traversing the scene graph annotations of VG for a given image. We identify objects with identical class labels, excluding those occupying less than 0.05% of the total image area, and then remove object pairs that overlap. This procedure results in two non-overlapping objects of the same class within the image, appropriately sized.

Subsequently, we aim to recognize attributes that can effectively differentiate between these two objects. A straightforward approach assumes that distinct attribute names lead to differentiation. However, we must exercise caution as captions and scene graphs may not fully encompass all attributes [28]. Therefore, we need to carefully ascertain these attributes. For instance, if both objects are ”fairy” and ”white” dogs, and one object is labeled solely as ”fairy” while the other one is only as ”white,” these attributes, although different, may not effectively distinguish between the objects. To address this concern, we leverage WordNet [29] to confirm that the attributes of both objects share the same grand hypernym while being distinct from one another. Finally, we concatenate the attribute and object name to form a concise and accurate text description for the respective object. For quantitative analysis, we only use one attribute for each object.

Relation. The text pair for relation compositionality probing has to meet the requirement: the two texts should be distinguishable from each other solely based on their relation. Firstly, the two texts need to differentiate from each other through their relation to test the model's understanding of relation compositionality. Additionally, for quantitative analysis, the text should be identical in all aspects except for the relation. In this context, we primarily explore the simplest case of text, specifically the object1-relation-object2 triplet scenario. To meet the requirements for quantitative testing, no attribute should be included in the text as it could be utilized to distinguish between the two objects. Moreover, object1 in both texts should possess the same class label but correspond to different objects in the image, thereby eliminating the influence of the object class. Furthermore, object2 should be identical in both texts.

To identify suitable triplet pairs in the image, we traverse all triplet pairs in the scene graph, ensuring that object1 in two texts belong to the same category and that their bounding boxes do not overlap. We also verify that the relations have distinct names and that object2 are the same. Simultaneously, we filter out bounding boxes with an area less than 0.05% of the total image area. However, we find that relations with distinct names can still have similar semantics, such as ”computer above table” and ”computer on top of table,” as well as ”girl wearing jacket” and ”girl in jacket.” Therefore, we manually filter the relation data, ensuring that objects in the pair can be distinguished based on text.

Priority. While testing models with the relation dataset, we observed that models were not only influenced by the relation but also affected by another object present in the text. When instructing the model to ground a triplet: object1-relation-object2, both instances of object1 are highlighted, demonstrating the model's inability to reason about the relation. Furthermore, object2 is also highlighted, indicating the model's deficiency in reasoning about the subject among nouns. To investigate this issue, we propose the priority dataset. By utilizing the scene graph of an image, we can easily obtain object1-relation-object2 triplet pairs that reverse the positions of object1 and object2. However, we have observed that the relations in these pairs may not necessarily align with each other. On one hand, enforcing strict relation consistency could lead to violations of visual semantics, as exemplified by the contradicting meanings in texts like ”slow cooker on top of microwave” and ”microwave on top of slow cooker.” On the other hand, the relationship between objects plays a crucial role in distinguishing the subject among nouns within a given text.

We utilize VG to construct our ARPGrounding datasets. Within the 108,249 images from VG, we meticulously gathered a total of 6,632 samples for attribute, 370 samples for relation, and 4,423 samples for priority. Each sample consists of two object-text pairs, which pertain to distinct object instances within an image, yet exhibit ambiguity in terms of compositionality. The relatively lower number of relation samples compared to the other two categories can be attributed to the fact that in cases involving two ambiguous objects within an image, it is considerably more likely for them to possess distinguishable attributes but less likely for them to exhibit a distinguishable relation with another object. More detailed statistics of ARPGrounding can be found in the Appendix.

3.2. Metric

The performance on ARPGrounding is evaluated based on the mean activation values inside the bounding boxes of the heatmaps. Each sample in ARPGrounding comprises two objects, each represented by a bounding box and the associated text. Correct association between the texts and their respective bounding boxes defines a sample's correctness. Given M_0, M_1 represent binary masks derived from object bounding boxes and H_0, H_1 denote the generated heatmaps, where masks and heatmaps match the image size, the mean activation of M_i and H_j is defined as:

$$act(M_i, H_j) = \frac{1}{\sum M_i} \sum M_i \odot H_j \tag{1}$$

The $\odot$ is element-wise multiplication. The score of a sample is computed based on mean activation combinations:

$$f(M_0, H_0, M_1, H_1) = \begin{cases} 1 \text{ if } act(M_0, H_0) > act(M_1, H_0) \\ \quad \text{and } act(M_0, H_1) < act(M_1, H_1) \\ 0 \text{ otherwise} \end{cases} \tag{2}$$

This metric tests whether the ground truth object for a given text is scored higher than the alternative object in the image and whether this holds for the other object-text pair in the sample as well.

Finally, we compute the mean accuracy across the entire dataset:

$$acc = \frac{1}{N} \sum f(M_0, H_0, M_1, H_1) \tag{3}$$

3.3. Gradient-based Localization

To enable VLMs to discern objects in the image, we employ explainability techniques to analyze the VLMs. Our approach adopts the widely used Grad-CAM algorithm [37], seamlessly integrating it into VLMs without introducing additional complexity. While Grad-CAM was originally tailored for single-modal Convolutional Neural Networks [17], we guide the model to focus on locating the targets described by semantically intricate textual expressions by utilizing the gradient flow from texts to the intermediate layer of vision transformer.

To compute the grounding heatmap, we first extract an intermediate attention map A_z in the multimodal transformer ϕ_f and denote this function as f_z, the input image and text are symbolized as $\boldsymbol{v}$ and $\boldsymbol{t}$ respectively:

$$A_z = f_z(\phi_f(\boldsymbol{v}, \boldsymbol{t})) \tag{4}$$

Then, we calculate the gradient of A_z with respect to the score of image-text pair. The score denotes the image-text similarity in the case of being trained with image-text contrastive loss or signifies the image-text matching score when trained with image-text matching loss, depending on the specific model. We represent the score uniformly as y and the gradient is calculated as:

$$G_z = \nabla A_z = \frac{\partial y}{\partial A_z} \tag{5}$$

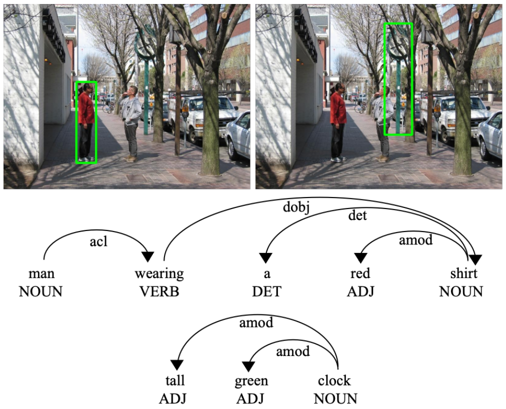


Figure 3. Examples of dependency parsing using spaCy [11] (acl: clausal modifier of a noun, dobj: direct object of a verb, det: determiner, amod: adjectival modifier of a noun). Each arc connects a parent node to a child node. We utilize the root of the dependency tree, namely, ”man” and ”clock,” to ascertain each element in the pair describes different objects. Bounding boxes serve solely for visualization purposes.

Next, we calculate the heatmap H using A_z and G_z as follows:

$$H = \text{ReLU}(A_z \odot G_z) \tag{6}$$

where $\odot$ is an element-wise multiplicaiton. 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.

4. Composition-aware Fine-tuning

Regarding the identified challenges in VLMs when it comes to the task of compositional grounding, we suggest a direct solution involving weak supervision. Our composition-aware fine-tuning pipeline first generates a text pair for each image where each component in the pair describes distinct objects. Then we introduce a pretext task that is aimed at augmenting diversity within the grounding heatmaps of these two texts. This pipeline significantly aids in enhancing the VLMs’ understanding of compositional semantics. In the following section, we delineate the methodologies for text pair generation and elaborate on the loss functions utilized during the training process.

Text pair generation. Existing image-text datasets often feature multiple descriptions for the same image. We generate text pairs by sampling different texts for each image. Incorporating dependency parsing into the text sampling phase, we analyze the grammatical structure of a sentence to identify associated words and ascertain their relationships. An illustration of dependency parsing with spaCy [11] is depicted in Figure 3. For a given image, all the texts of the image are represented in a tree structure, with directed arcs representing the connections between each word in the sentence. We leverage the root of the dependency tree to formulate text pairs that describe distinct objects. Those pairs contain different objects that are associated with various attributes, relations, and priorities, and provide rich compositional information for the subsequent training phase.

Loss. To harness the rich compositional information in the text pair, we propose a pretext task to induce variability in the generated heatmaps. Considering a training batch crafted through the aforementioned procedure $\mathcal{B} = \{(\boldsymbol{v}^i, \boldsymbol{t}_0^i, \boldsymbol{t}_1^i)\}_{i=1}^n$, where $\boldsymbol{v}^i$ represents the image and $\boldsymbol{t}_0^i, \boldsymbol{t}_1^i$ denote the text pair, we proceed with fine-tuning VLMs utilizing the following loss functions:

$$\mathcal{L} = E_{(\boldsymbol{v},\boldsymbol{t}_0,\boldsymbol{t}_1)\sim\mathcal{B}}[\frac{1}{w \cdot h}\sum H_0 \odot H_1] \tag{7}$$

Where H_0, H_1 are the heatmaps generated corresponding to $\boldsymbol{t}_0$ and $\boldsymbol{t}_1$ respectively. w and h stand for the width and height of H_0, H_1. This novel loss leverages coarse-grained text descriptions instead of dense bounding box annotations, thereby applying only weakly supervised guidance to the heatmap. This loss function directs the model to generate distinct heatmaps for diverse textual inputs, encompassing various objects, attributes, relations, and priorities. Consequently, it encourages the model’s outputs to be more discriminative and alleviates noise arising from compositional ambiguity within the generated heatmaps.

5. Experiments

In this section, we report the results of four state-of-the-art VLMs on conventional visual grounding datasets and ARPGrounding dataset. Our findings indicate that VLMs demonstrate proficiency on conventional visual grounding datasets but reveal limitations on ARPGrounding. We also explore our fine-tuning strategy on both CLIP and ALBEF to demonstrate its effectiveness.

5.1. Datasets and Models

Datasets. In addition to ARPGrounding, we employ three datasets: VG, Flickr30k entities [33], and ReferIt [15] following Akbari et al. [1]. We follow Akbari et al. [1] to report the *pointing game* [50] accuracy. These datasets serve as the foundation for demonstrating VLMs’ performance on weakly supervised visual grounding tasks. As for composition-aware fine-tuning, we use region descriptions in VG to construct the text pair. We apply a filtering process to the images within VG to ensure that there is no overlap between the training data and the test splits of any other datasets. Ultimately, we collect a total of 83,517 images for fine-tuning.

Models. We evaluate four VLMs. Specifically, we assess OpenAI’s CLIP [34]—a dual-stream model pre-trained on a dataset comprising 400 million image-text pairs with contrastive objectives. Additionally, we examine ALBEF

Table 1. Results on visual grounding and ARPGrounding datasets.

Method	Visual Grounding				ARPGrounding			
	VG	Flickr	ReferIt	Mean	Attribute	Relation	Priority	Mean
random	11.15	27.24	24.30	20.90	25.00	25.00	25.00	25.00
WWbL [39]	62.31	75.63	65.95	67.96	-	-	-	-
WWbL^{++} [38]	66.63	79.95	70.25	72.28	-	-	-	-
CLIP	57.46	75.26	56.77	63.16	42.78	9.19	11.24	21.07
ALBEF	75.04	84.49	69.26	76.26	61.25	29.19	14.94	35.13
METER	60.52	75.28	66.41	67.40	43.70	26.49	38.39	36.19
BLIP2	69.50	84.96	68.71	74.39	23.46	31.35	15.28	23.36

[18], which introduces contextualized multimodal fusion with a co-attention mechanism and is additionally trained with masked language modeling and image-text matching losses. Furthermore, we evaluate METER [6], characterized by stacked transformer encoding layers on both visual and text encoders. Lastly, our analysis includes BLIP2 [21], where a Q-Former is connected to a frozen image encoder, aiming to achieve improved image-text alignment while preserving robust visual representation.

5.2. Evaluation of VLMs

Models perform well on conventional visual grounding datasets. In Table 1, we present models' performance on VG, Flickr, and ReferIt to demonstrate their performance on weakly supervised visual grounding datasets. We also calculate the mean score for convenient comparison. We adopt the same *pointing game* accuracy metric and compare our results against previous state-of-the-art methods WWbL^{++} [38]. We find that all four VLMs can achieve performance on par with or surpass WWbL^{++} without any training. Both ALBEF and BLIP2 outperform WWbL^{++}, setting new state-of-the-art benchmarks. Additionally, CLIP and METER exhibit performance levels that are comparable to WWbL^{++} according to the mean score.

Models exhibit deficiencies in compositional grounding datasets. Models struggle across the board on ARPGrounding, often performing close to or below random chance. As shown in Table 1, models exhibit suboptimal performance when assessed in terms of their average performance across three compositional categories. From a model-specific perspective, we observe a meager absolute improvement of approximately 10% compared to the random chance from the two top-performing models, namely ALBEF and METER, while the remaining models perform even worse than random chance. Models also demonstrate varying reasoning capabilities when confronted with distinct forms of compositionality. Notably, models perform relatively better in the attribute category with a mean score of 42.80, but their performance is comparatively poorer in relation and priority, scoring mean values of 24.06 and 19.96, respectively, indicating that detecting relation and priority is a more challenging problem than attribute recognition. Overall, models exhibit deficiencies in ARPGrounding.

5.3. Composition-aware Fine-tuning

Training details. We employ composition-aware fine-tuning on CLIP and ALBEF to showcase its efficacy. In particular, we initiate the fine-tuning of CLIP and ALBEF using identical pre-trained weights as used in section 5.2. Each image in the training dataset is associated with approximately 50 textual descriptions. We select two descriptions per image, each delineating distinct objects. We implement our framework on PyTorch [32] and train it for 10 epochs on an NVIDIA A100 GPU with 80GB of memory. For CLIP, we resize the input images to 224 × 224 and cap the maximum length of each text to 77 as suggested by Radford et al. [34]. We use a batch size of 256 and a learning rate of 1e-7. For ALBEF, we resize the input images to 384 × 384 and set the maximum length of each text to 30 as Li et al. [18]. We use a batch size of 54 and a learning rate of 2e-7. Both models are optimized using AdamW [27] optimizer with 50 steps of warmup and a cosine-annealing learning rate scheduler.

Table 2. Grounding results comparison of fine-tuning two state-of-the-art VLMs.

Method	Visual Grounding			ARPGrounding		
	VG	Flickr	ReferIt	Attribute	Relation	Priority
CLIP	57.46	75.26	56.77	42.78	9.19	11.24
TSVLC	57.57	75.11	56.78	42.85	9.46	11.15
DAC	58.50	76.91	58.33	42.63	11.35	8.70
CLIP caft. (Ours)	**60.43**	**78.07**	**63.75**	**44.56**	**13.24**	**21.30**
ALBEF	**75.04**	84.49	69.26	61.25	29.19	14.94
ALBEF caft. (Ours)	74.80	**85.93**	**74.67**	**66.34**	**38.65**	**24.21**

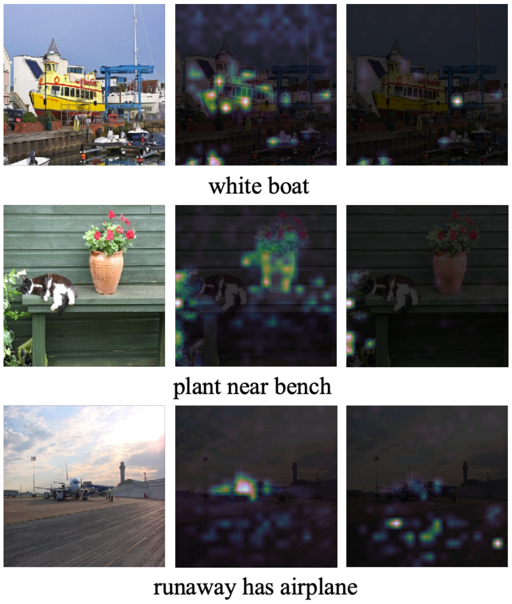
Results. In Table 2, we present a comparative analysis between CLIP and CLIP caft., as well as ALBEF and ALBEF caft., where "caft." denotes composition-aware fine-tuning. Additionally, we include the results of TSVLC [7] and DAC [8], both of which are based on CLIP and have demonstrated promising outcomes in addressing compositional challenges in image-text matching tasks. For CLIP, composition-aware fine-tuning enhances CLIP's performance on Attribute from 42.78 to 44.56, on Relation from 9.19 to 13.24, and on Priority from 11.24 to 21.30. Conversely, no significant improvement is observed for TSVLC and DAC. For ALBEF, this methodology elevates performance on Attribute from 61.25 to 66.34, on Relation from 29.19 to 38.65, and on Priority from 14.94 to 24.21. Notably, it substantially augments the efficacy of both models in conventional visual grounding datasets as well, only ALBEF slightly decreases in VG by 0.24%. Overall, ALBEF caft. becomes the best model, in comparison to all other models, and in ARPGrounding-Priority, it lags behind only METER. Figure 4 illustrates some visualization results, more visualization results are shown in Appendix.

In general, our proposed composition-aware fine-tuning yields significant improvements in ARPGrounding and also enhances performance in conventional visual grounding datasets. Our results underscore the efficacy of extracting rich information from diverse textual sources related to an image, even in the absence of dense annotations, leading to substantial advancements in compositional understanding. While diverse vision-language pre-training objectives contribute significantly to representation learning, Ma et al. [28] suggests that merely expanding the scale of pre-training datasets is ineffective in capturing compositional structures. Hence, we furnish evidence supporting the notion that pursuing algorithmic enhancements can augment model capacity through the utilization of cost-effective image-text annotations.

5.4. Analysis of Compositional Grounding

In this section, we report the fine-grained score of attribute and relation of ARPGrounding, we also conduct a case study on priority since categorizing priorities is not straightforward.

We categorize attributes into action, color, material, size, and state, and divide relations into action and spatial following Zhao et al. [51]. We use text that falls into these categories to distinguish two objects, therefore the random performance is 50%. As shown in Figure 3, for attributes, models get relatively high sores on color and material, low on action, size, and state, and perform worst on size. This might stem from the need for scene-based reasoning when comprehending size. Size-related terms (e.g., small, large, giant) may lack consistency in visual size due to influences such as camera angles and the distance between the object and the camera. For relations, CLIP and METER excel in spatial relations, while ALBEF and BLIP2 demonstrate superior performance in action relations. This suggests no uniform tendencies in relation to grounding performance. For priority, we plot the ten most frequent object pairs and count the frequency of each object being chosen, as shown in Figure 6. We observe that models exhibit a bias toward specific objects rather than adhering to priority considerations.


white boat

plant near bench

runaway has airplane

Figure 4. Visualization of grounding results of ALBEF and ALBEF caft. From left to right, the three columns depict the original image, the heatmap generated by ALBEF, and the heatmap generated by ALBEF caft. We demonstrate that our composition-aware fine-tuning pipeline effectively enhances compositional understanding.

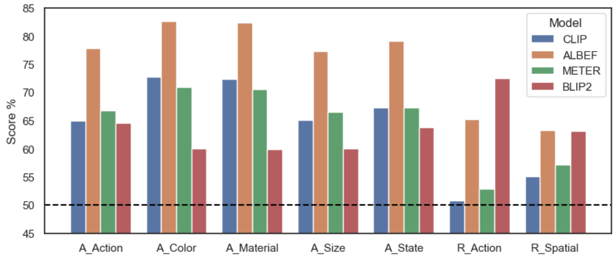


Figure 5. Analysis of compositional grounding on attribute and relation splits in ARPGrounding. We further categorize attributes into action, color, material, size, and state. These are represented in the figure with X-axis labels starting with 'A'. Relations are divided into action and spatial, denoted in the table with labels starting with 'R'. A horizontal dashed line signifies chance performance.

5.5. Ablation Studies

In this section, we present ablations to demonstrate the choices in the proposed composition-aware fine-tuning pipeline. Particularly we investigate how much does the text pair generation process and the new pretext task benefit fine-tuning.

As described in section 4, the generation of a text pair involves utilizing a dependency parsing tree to sample a pair

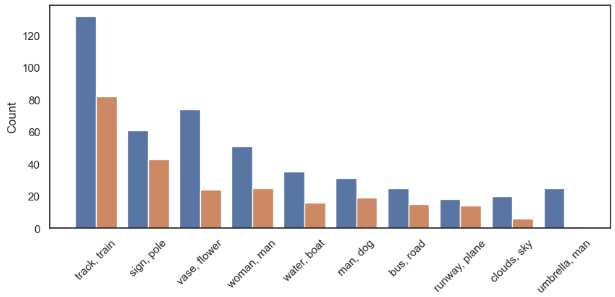


Figure 6. Analysis of compositional grounding on priority. We plot the ten most frequent object pairs and count the frequency of each object being chosen by CLIP. The more frequent object in each pair was manually selected for leftward plotting.

Table 3. Ablation studies of text pair generation process. †, ‡ represents fine-tuned with text pair (1), (2) respectively. We use the same loss function with caft. in these ablation studies.

Method	Visual Grounding			ARPGrounding		
	VG	Flickr	ReferIt	Attribute	Relation	Priority
CLIP	57.46	75.26	56.77	42.78	9.19	11.24
CLIP †	55.67	75.11	53.05	45.67	9.46	5.63
CLIP ‡	**60.52**	**78.47**	63.61	44.26	11.89	13.90
CLIP caft. (Ours)	60.43	78.07	**63.75**	**44.56**	**13.24**	**21.30**
ALBEF	**75.04**	84.49	69.26	61.25	29.19	14.94
ALBEF †	57.36	67.93	56.46	55.71	17.84	8.61
ALBEF ‡	72.71	83.46	73.32	65.65	35.95	18.80
ALBEF caft. (Ours)	74.80	**85.93**	**74.67**	**66.34**	**38.65**	**24.21**

Table 4. Ablation studies of the new pretext task. "ft." represents fine-tuning CLIP with contrastive loss, and fine-tuning ALBEF with image-text matching loss. "caft." represents trained with the new pretext task. We use the same training data in these ablation studies.

Method	Visual Grounding			ARPGrounding		
	VG	Flickr	ReferIt	Attribute	Relation	Priority
CLIP	57.46	75.26	56.77	42.78	9.19	11.24
CLIP ft.	57.53	76.35	60.20	43.15	11.35	6.67
CLIP caft. (Ours)	**60.43**	**78.07**	**63.75**	**44.56**	**13.24**	**21.30**
ALBEF	**75.04**	84.49	69.26	61.25	29.19	14.94
ALBEF ft.	65.77	73.58	63.89	58.46	30.54	8.82
ALBEF caft. (Ours)	74.80	**85.93**	**74.67**	**66.34**	**38.65**	**24.21**

of text that pertain to distinct objects. We design two text pair variants including (1) text pair that refers to the same object and (2) text pair that is randomly sampled. We follow the composition-aware fine-tuning pipeline but use different training data. Results of fine-tuning CLIP and ALBEF with these two variants are shown in Table 3. It can be observed that fine-tuning with text pair (1) adversely affects grounding performance. Conversely, fine-tuning with text pair (2) contributes to improved performance. This is attributed to the fact that, on average, images in the VG dataset contain approximately 35 objects, making random sampling likely to yield pairs describing different objects. However, the optimal results are achieved through sampling via the dependency parsing tree. Ablation studies of the proposed pretext task are shown in Table 4. We provide CLIP fine-tuned with contrastive learning loss and ALBEF fine-tuned with image-text matching loss on our generated text pairs. The results indicate that both contrastive learning loss and image-text matching loss diminish the grounding performance of both CLIP and ALBEF across nearly all test datasets. In contrast, when CLIP and ALBEF are fine-tuned on the novel pretext task, there is a noticeable enhancement in performance across all test datasets. This proves that the generated text pairs offer valuable compositional information. Moreover, the new pretext task effectively extracts this information compared to contrastive learning and image-text matching.

5.6. Fully Supervised Methods on ARPGrounding

Table 5. ARPGrounding results of GLIP and Grounding DINO.

Method	ARPGrounding		
	Attribute	Relation	Priority
GLIP-T	34.91	12.43	11.76
GLIP-L	37.15	9.46	12.34
Grounding-DINO-T	37.24	6.76	15.42
Grounding-DINO-B	47.03	13.78	25.68

To explore the performance of fully supervised methods on compositional grounding, we test GLIP [22] and Grounding DINO [26] on ARPGrounding as shown in Table 5. We use models to generate the highest score bounding box of the text and compare the Intersection over Union (IOU) between the positive bounding box and the negative bounding box with regard to the text. We find these two fully supervised visual grounding models do no better than weakly supervised VLMs. The best performing fully supervised method, Grounding-DINO-B, falls short in attribute against ALBEF, lacks competitiveness with ALBEF, METER, and BLIP2 in relation, and demonstrates inferior performance compared to METER in priority.

6. Conclusion

In this paper, we first employ explainability techniques to achieve state-of-the-art results in weakly supervised visual grounding tasks using VLMs. We then introduce ARPGrounding, a novel benchmark, and reveal the limitations of VLMs in fine-grained visio-linguistic compositionality. To address these challenges, we propose a composition-aware fine-tuning pipeline that significantly enhances VLMs' compositional understanding without relying on dense annotations.

Acknowledgment. This work was jointly supported by the National Key R&D Program of China (2022ZD0116309) and the National Natural Science Foundation of China (62236010 and 62276261)

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. 2, 5

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 1

[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

[4] Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. Vlp: A survey on vision-language pre-training. *Machine Intelligence Research*, 20(1):38–56, 2023. 1

[5] 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

[6] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18166–18176, 2022. 6

[7] Sivan Doveh, Assaf Arbelle, Sivan Harary, Eli Schwartz, Roei Herzig, Raja Giryes, Rogerio Feris, Rameswar Panda, Shimon Ullman, and Leonid Karlinsky. Teaching structured vision & language concepts to vision & language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2657–2668, 2023. 6

[8] Sivan Doveh, Assaf Arbelle, Sivan Harary, Roei Herzig, Donghyun Kim, Paola Cascante-Bonilla, Amit Alfassy, Rameswar Panda, Raja Giryes, Rogerio Feris, et al. Dense and aligned captions (dac) promote compositional reasoning in vl models. *Advances in Neural Information Processing Systems*, 36, 2024. 6

[9] 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

[10] Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3635–3644, 2021. 2

[11] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017. 5

[12] Syed Ashar Javed, Shreyas Saxena, and Vineet Gandhi. Learning unsupervised visual grounding through semantic self-supervision. *arXiv preprint arXiv:1803.06506*, 2018. 2

[13] Lei Jin, Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Guannan Jiang, Annan Shu, and Rongrong Ji. Refclip: A universal teacher for weakly supervised referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2681–2690, 2023. 1

[14] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's" up" with vision-language models? investigating their struggle with spatial reasoning. *arXiv preprint arXiv:2310.19785*, 2023. 2

[15] 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. 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*, 123:32–73, 2017. 3

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 2012. 4

[18] 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*, 34:9694–9705, 2021. 6

[19] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 1

[20] Jiahao Li, Greg Shakhnarovich, and Raymond A Yeh. Adapting clip for phrase localization without further training. *arXiv preprint arXiv:2204.03647*, 2022. 1

[21] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1, 6

[22] 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. 8

[23] Yi Li, Hualiang Wang, Yiqun Duan, and Xiaomeng Li. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv preprint arXiv:2304.05653*, 2023. 1

[24] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana

Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023. 1

[25] Yuqi Lin, Minghao Chen, Wenxiao Wang, Boxi Wu, Ke Li, Binbin Lin, Haifeng Liu, and Xiaofei He. Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15305–15314, 2023. 1

[26] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 8

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 6

[28] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10910–10921, 2023. 2, 3, 7

[29] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 3

[30] Mitja Nikolaus, Emmanuelle Salin, Stephane Ayache, Abdellah Fourtassi, and Benoit Favre. Do vision-and-language transformers learn grounded predicate-noun dependencies? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1538–1555, 2022. 2

[31] Letitia Parcalabescu, Albert Gatt, Anette Frank, and Iacer Calixto. Seeing past words: Testing the cross-modal capabilities of pretrained v&l models on counting tasks. *IWCS 2021*, page 32, 2021. 2

[32] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019. 6

[33] 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

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 5, 6

[35] Navid Rajabi and Jana Kosecka. Towards grounded visual spatial reasoning in multi-modal vision language models. *arXiv preprint arXiv:2308.09778*, 2023. 2

[36] Arijit Ray, Filip Radenovic, Abhimanyu Dubey, Bryan A Plummer, Ranjay Krishna, and Kate Saenko. Cola: How to adapt vision-language models to compose objects localized with attributes? *arXiv preprint arXiv:2305.03689*, 2023. 2

[37] 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

[38] 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, 6

[39] 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. 1, 2, 6

[40] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. *arXiv preprint arXiv:2304.06712*, 2023. 1

[41] 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

[42] Ajinkya Tejankar, Maziar Sanjabi, Bichen Wu, Saining Xie, Madian Khabsa, Hamed Pirsiavash, and Hamed Firooz. A fistful of words: Learning transferable visual models from bag-of-words supervision. *arXiv preprint arXiv:2112.13884*, 2021. 2

[43] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022. 2

[44] 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

[45] 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

[46] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2945–2954, 2023. 1

[47] Hsiu-Yu Yang and Carina Silberer. Are visual-linguistic models commonsense knowledge bases? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5542–5559, Gyeongju, Republic of Korea, 2022. International Committee on Computational Linguistics. 2

[48] Shuquan Ye, Yujia Xie, Dongdong Chen, Yichong Xu, Lu Yuan, Chenguang Zhu, and Jing Liao. Improving common-

sense in vision-language models via knowledge graph riddles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2634–2645, 2023. 2

[49] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2022. 2

[50] 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. 2, 5

[51] Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv preprint arXiv:2207.00221*, 2022. 2, 7

[52] 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. 1, 2