Learning Visual Grounding from Generative Vision and Language Model

Shijie Wang[1*] Dahun Kim[2] Ali Taalimi[3] Chen Sun[1] Weicheng Kuo[2]
[1] Brown University [2] Google DeepMind [3] Google Cloud

Abstract

Visual grounding tasks aim to localize image regions based on natural language references. In this work, we explore whether generative VLMs predominantly trained on image-text data could be leveraged to scale up the text annotation of visual grounding data. We find that grounding knowledge already exists in generative VLM and can be elicited by proper prompting. We thus prompt a VLM to generate object-level descriptions by feeding it object regions from existing object detection datasets. We further propose attribute modeling to explicitly capture the important object attributes, and spatial relation modeling to capture inter-object relationship, both of which are common linguistic pattern in referring expression. Our constructed dataset (500K images, 1M objects, 16M referring expressions) is one of the largest grounding datasets to date, and the first grounding dataset with purely model-generated queries and human-annotated objects. To verify the quality of this data, we conduct zero-shot transfer experiments to the popular RefCOCO benchmarks for both referring expression comprehension (REC) and segmentation (RES) tasks. On both tasks, our model significantly outperform the state-of-the-art approaches without using human annotated visual grounding data. Our results demonstrate the promise of generative VLM to scale up visual grounding in the real world.

1. Introduction

Visual grounding aims to identify image region described by a natural language query. It provides a strong foundation for tasks like visual reasoning and human-robot interaction. One of the most popular visual grounding task referring expression comprehension (REC) [38, 64] aims to localize an object by language references and generate the corresponding bounding box. Another similar task is referring expression segmentation (RES), which requires pixel-level grounding to generate the referred object mask. Visual grounding tasks require models to accurately understand the natural language descriptions of the referred objects' categories, visual appearance, and their relationships with other objects or the scene. An important challenge is the limited scale of grounding datasets: Traditional visual grounding datasets like [21, 38, 40, 64] require detailed human labeling and verification to generate object-level text annotations, making it expensive and unscalable in practice. For example, the popular RefCOCO/+/g REC datasets only contain $\mathcal{O}$(10K) images and objects. In contrast, detection datasets [12, 26, 30, 49] are much larger and more diverse with $\mathcal{O}$(1M) images and $\mathcal{O}$(1-10M) objects. Such 2-3 orders of magnitude difference show the promise of learning grounding from large-scale detection datasets.

Generative Vision-Language Models (VLMs) [2, 6, 7, 28] flexibly unify multiple vision-language tasks such as visual-question answering (VQA) and image captioning without task-specific design by formatting various tasks as image-and-text to text generation. Most current VLMs [2, 7, 28, 36] are primarily pre-trained on large image-text datasets [7, 48, 66] for natural language generation tasks. They lack the ability to directly perform localization tasks (*i.e.* generating boxes or masks) without downstream fine-tuning. This raises the question of whether these VLMs inherently lack object-level knowledge and grounding capability. Conceptually, both scene-level and object-level visual information can be encapsulated within an image at various scales (*e.g.* an image of a forest and a tree). Practically, generative VLMs are usually trained on $\mathcal{O}$(1B) datasets that contain feature diverse distribution covering complex scenes and object-centric images. We thus hypothesize that VLMs can learn object-centric knowledge solely from image-level pre-training tasks and produce object-level descriptions when "zooming-in" an image to the single object region, which can be used for grounding model training. Consequently, without expensive fine-tuning on grounding dataset, a generalist generative VLM may naturally serve as a teacher model, providing supervision for student visual grounding models without extra human annotations.

As motivated above, we utilize PaLI-3 [6], a state-of-the-arts generative VLM to construct large-scale referring expression dataset automatically for scalable visual grounding. To better model human linguistic patterns beyond single-object descriptions, we also employ

[*] Work done while at Google DeepMind.

spatial relationship heuristics and attributes priors to enrich the generated region-text dataset. We create VLM-VG, a large-scale dataset for fine-grained visual grounding created based on two object detection datasets: COCO 2017 [30] and Objects365 v1 [49], which are 1-2 orders of magnitude larger than existing grounding datasets. We focus on the "zero-shot" referring expression comprehension/segmentation setting, where no human-annotated text annotations are utilized during pre-training. We pretrain a simple light-weight grounding model on the proposed dataset and perform zero-shot evaluation on the RefCOCO/+/g datasets, conducting systematic comparisons and analyses. Benefiting from the large-scale grounding pre-training with high-quality referring expression annotations from generative VLMs and relation modeling, we significantly outperform prior zero-shot REC/RES methods. In summary, our work has three main contributions:

- We hypothesize and empirically observe that image-text pre-trained generative VLMs can naturally provide high-quality object-level captions for single-object regions. In this way, we leverage a generalist VLM to automatically generate referring expression annotations to provide supervision for specialist visual grounding models, thereby overcoming the limitations of small-scale grounding datasets.
- We further enrich the auto-labeled region-text pairs by leveraging heuristics of spatial relationships and cross-domain knowledge of object attributes, providing more comprehensive and diverse query annotations.
- We introduce VLM-VG, a visual grounding dataset for scalable referring expression comprehension/segmentation, without requiring manual grounding annotations. By pre-training on VLM-VG, we achieve SoTA zero-shot performance on RefCOCO/+/g benchmarks in both REC and RES tasks with a lightweight Faster R-CNN based model.

2. Related Work

Visual Grounding is a critical ability for AI systems. Specifically, it aims to identify and associate specific parts of visual data with corresponding textual descriptions, thus bridging the semantic gap between the high-dimensional visual representations and the symbolic language representations. Referring expression comprehension/segmentation [31, 32, 38, 64] (REC/RES), and phrase grounding [21, 24, 43, 59] (PG) are two common visual grounding tasks. REC/RES tasks require the model to detect or segment a single object from the image based on language queries referring to the object while PG aims to localize each entity mentioned by the phrase. In this work, we focus on the referring expression comprehension and segmentation task requiring fine-grained object-level understanding capabilities. Compared with other localization tasks such as detection dataset [26,30,49], grounding datasets [38,64] are much smaller due to the high annotation cost.

Generative Vision Language Models has made tremendous advances across various areas. Inspired by token masking [10] or next-token prediction [3] training objectives utilized by text modeling, generative vision-language pre-training [7, 8, 17, 35, 36, 55–58, 67, 69] extends similar ideas to various tasks that can be modeled by a general "image-text to text" interface. Optimized by a single language modeling loss, large-scaled generative vision-language models can be flexibly applied to multiple image-language downstream tasks including image captioning, visual question answering (VQA), or image classification. Recent works [5,27,42,60,63] show VLMs can ground objects after large-scale region-text fine-tuning, which could be expensive and resource-consuming. Motivated by this, we investigate whether image-level generalist VLMs can ground objects without heavy downstream fine-tuning. We observe that VLMs could provide informative object-level captions given a single-object image regions. The generated regional captions can provide robust and generalizable supervision for specialist grounding models, enabling us to scale up visual grounding datasets for free.

Unsupervised Visual Grounding can be categorized into two types. One type of works [13, 34, 41, 50, 52, 62, 65] leverage pre-trained visual-language models (VLMs) such as CLIP [44], FLAVA [51], or Stable Diffusion [47] with strong visual language alignment capability to retrieve box or mask from pre-extracted box/mask proposals generated by off-the-shelf detectors [46] or segmentation models [14, 23]. Another line of works [29, 33, 42, 68] take in images directly and pre-train the grounding model on detection [30, 49] and grounding [20, 24] datasets, then perform zero-shot evaluation on the down-stream tasks. Specifically, [29, 33, 68] combine base detection models with visual-language fusion modules while [27, 42] utilize large language models to formulate REC task as language modeling.

We adopt the second way to pre-train a simple grounding model following [25] on our constructed grounding dataset, VLM-VG. Different from [19] which constructs pseudo-queries by parsing images with object and attribute detectors and composing attributes into on phrases by rule-based templates, and [42], which isolates object-related nouns from image-level captions to construct referring expressions, we propose to leverage generative visual-language models to construct language annotations automatically for object-centric regions in the images by using box annotations from detection dataset. In this way, we leverage detection datasets that are generally 1-2 orders of magnitude larger in terms of images and objects and bypass the requirements of human annotations.

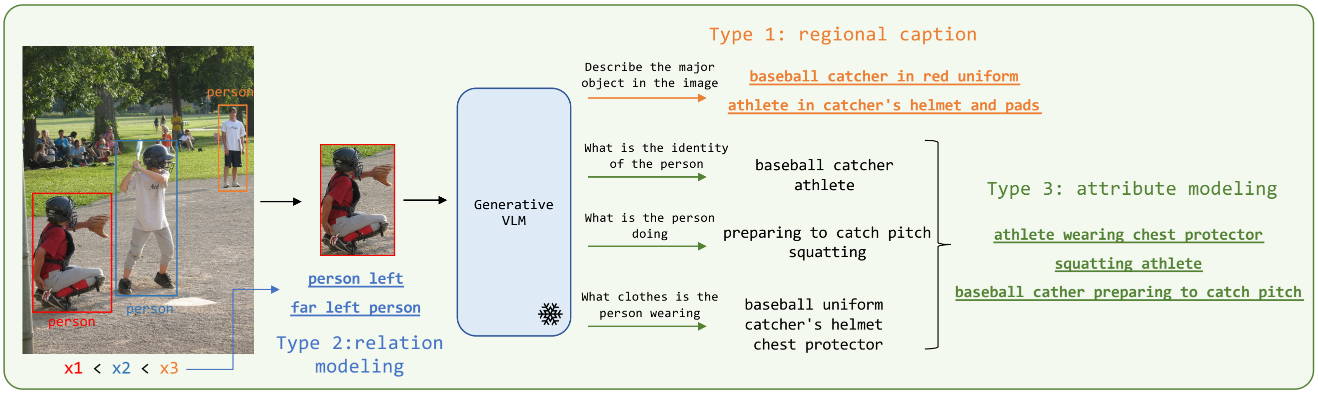

Figure 1. **Overview of referring expression generation.** We propose to generate grounding annotations automatically using generative vision language models. We construct three types of annotations to model the diversity of human linguistic norms. Type 1: regional caption directly generated by prompting the VLM with a generic instruction to describe the major object in the cropped image regions. Type 2: relational descriptions generated by rule-based methods utilizing spatial heuristics. Type 3: Attribute-rich descriptions that explicitly model attributes of the object by querying the VLM with "attribute" prompts.

3. Method

3.1. Generative VLM

Generative vision language models take image and text as input and generate text by autoregressive token prediction or masked token prediction. Since a variety of tasks can be converted into text generation tasks by designing different instructions and prompts, generative VLMs can thus handle multiple tasks without task-specific training objectives. Formally, a generative VLM takes in the image-text pair `<image, prompt>`, where different downstream tasks can be solved by different `prompt` designs. For instance, in the Visual Question Answering (VQA) task, the prompt can be constructed as a combination of the question and answering instructions. By prompting the generative generalist VLMs, it offers a unified interface to solve tasks directly and provide knowledge and information for other models.

3.2. Referring Expression Generation

In referring expression comprehension/segmentation tasks, the language queries generally focus on two key factors: accurate descriptions of objects' visual appearance and relationship with other objects in the images. Taking both into account, we propose to automatically generate large-scale referring expressions combining three types of queries. As demonstrated in Figure 1, we describe the referring expression generation process below:

Type 1: Regional Caption. Given an image containing multiple instances/objects represented by corresponding ground-truth bounding boxes, the goal is to generate descriptive language queries referring to the objects. For each image, we first filter out small objects whose area accounted for less than a threshold ratio K (*i.e.* K=0.05) of the entire image area. For the remaining objects, we crop the image using the associated bounding boxes to generate single object regions to isolate the objects from the influence of the context. Then we leverage a VLM to generate captions providing informative descriptions for the center object. In detail, we feed the object-centric regions to a state-of-the-arts generative VLM, PaLI-3 [6], and prompt it with language instructions following the image caption task. Since the cropped image may still include irrelevant object parts, we empirically design the prompt to instruct the model to focus on the center object and ignore the irrelevant background: "`Describe the major object in the image, ignore the background.`". Taking in the image-text pair of the cropped regions and language instructions, the VLM generates multiple predictions with corresponding confidence. We choose the top-5 answers with highest confidence as the expressions referring to the corresponding object without any further post-processing.

Type 2: Relation Modeling. The type 1 regional captions provide "local" descriptive references by isolating the center object from the whole image. However, the single object regions filter out global information such as positions in the scene and relationship with other objects, which is also a crucial perspective to refer to an object/instance. For example, in an image containing two cups with similar appearance positioned in different locations, it hard to successfully locate one cup solely relying on the regional descriptions on its visual appearance. We observe that spatial relationship is one of the most common relations in human cognition process to distinguish objects. Inspired by previous works [19,52], we design a rule-based method to model spatial relationship between objects utilizing bounding boxes as localization heuristics on three dimensions: horizontal, vertical and depth.

We model both relative and absolute spatial relationships which can be formulated by the tuple `(noun, rel, noun)` and `(noun, rel)` respectively. We use the object category label of detection dataset as the `noun` *e.g.* "person". To model the spatial relation, we treat the center coordinate of the bounding box as the position of an object. On the horizontal dimension, we compare one object with other objects by the x-coordinates for relative position. For absolute position, we normalize an object's x-coordinate against the image width, categorizing locations as "left" (`[0, 0.25)`), "center/middle" (`[0.25, 0.75]`), and "right" (`(0.75, 1]`). Vertically, we solely focus on absolute positioning through y-coordinate comparison. Similarly, we normalize the y-coordinates and categorize horizontal location as "top"(`[0, 0.25)`) and "bottom" (`(0.75, 1]`). For depth, following [19] that assumes objects with larger area is more likely to be closer to the camera (*i.e.* "front") and we only model the absolute position. We first calculate the ratio of the area of the smallest object to the largest object, we model depth relation for this image only if the ratio is smaller than a threshold (*i.e.* 0.4), which denotes there's significant scale differences among the objects. Upon satisfying this condition, we categorize an object's depth position by the the ratio of its area to the largest object: "behind" (`[0, 0.4)`) and "front" (`(0.8, 1]`). Following the rule-based methods, the extracted relation tuples are then formulated into referring phrases: "`noun rel noun`" for relative relation, "`rel noun`" or "`noun rel`" for absolute relation to expand the diversity of referring expressions.

The rule-based spatial relation modeling provides spatial-aware annotations absorbing information from localization heuristics. But it still has some limitations. For example, to determine objects' location solely based on its center point may fail especially when two objects are close or have occlusion, and using box area to categorize the depth is just a rough approximation since different objects might have various intrinsic sizes.

Type 3: Explicit Attribute Modeling. Attributes are important ways humans refer to objects. However, the generic region captions generated by prompting the VLM with a "general prompt" from type 1 may not contain adequate attribute information or may dilute the attribute information with other co-occurring words. We further ask the question: Will explicitly modeling object's attributes further complement the potential missing information in the generic captions and match human referring expression better?

We then propose to leverage both Large Language Model and Vision and Language Model to construct attribute-rich descriptions. Specifically, we first prompt GPT-4 [1] to enumerate common attributes to identify an object with the prompt "`what are the common attributes to identify an object?`" and then manually pick 7 typical attributes: *[cloth, action, gender, identity, color, material, shape]*. For each object category in the dataset, we further utilize GPT-4 to pick applicable attributes from the 7 candidate attributes. In this way, we associate each object class with its related attributes. We design a prompt template for each attribute (*e.g.* for *color*, the prompt is formulated as "`What is the color of the [object class]`"). The detailed prompt templates are listed in the supplementary. We then query PaLI-3 with the cropped region and attribute prompt to get the attribute of the center object by choosing the top-3 answers. Interestingly, we observe sometimes PaLI-3 will generate "unkown" or "unsuitable" when the attributes is not applicable to the object class. In this case, we just ignore this attribute for the specific object. Finally, after tagging each object with attributes, the attribute-rich descriptions are then formulated by combining attribute tags in the format of "`noun adj` or "`adj noun`" randomly, where we define *gender, identity* and the class label as the nouns, and *cloth, action, color, material, shape* as the adjectives.

3.3. VLM-VG: From Detection to Grounding

Following the steps above and combing the three types of annotations, we construct our dataset, VLM-VG, based on the training set of two object detection datasets COCO 2017 [30] and Objects365 v1 [49]. Specifically, for COCO, we filter out images present in the RefCOCO/+/g validation and test splits to avoid contamination during evaluation. For the Objects365 dataset, we keep the object categories corresponding to the 80 COCO classes to reduce the computation cost. For the vision-language model, we leverage the 5B-parameter PaLI-3 model which was fine-tuned on image captioning and visual question answering tasks, as detailed in [6].

In Table 1, we compare VLM-VG with existing visual grounding datasets in terms of size, data sources, and annotation methods. Traditionally, most visual grounding datasets rely on object-level text queries labeled by human annotators [22, 24, 38, 43, 64], which is associated with high labor costs, limited flexibility and scalability. Two recent datasets [27, 42] adopt image-caption pairs from web-collected datasets [4, 48] and utilizes GLIP [29] trained on Objects365 [49] and GoldG [20] to generate object boxes and uses NLP tools [16] to isolate object-level word chunks from image-level captions and expand them into referring expressions using dependency relations of the caption sentences. Despite its scale, this method is constrained by the quality of the web image captions (*e.g.* alt-text) and struggles to provide detailed and diverse references for objects from various perspectives that align with human linguistic norms.

Contrastively, VLM-VG extends existing detection datasets for visual grounding by constructing referring expressions using generative vision language models, with-

Dataset	Images	Objects	Text	Data Source	Box Anno.	Text Anno.
ReferIt [22]	20K	97K	131K	ImageCLEF	Human	Human
RefCOCO [64]	32K	50K	142K	COCO	Human	Human
RefCOCO+ [64]	20K	50K	142K	COCO	Human	Human
RefCOCOg [38]	27K	55K	85K	COCO	Human	Human
Flickr Entities [43]	31K	275K	513K	Flickr30k	Human	Human
Visual Genome [24]	108K	4.1M	4.1M	COCO	Human	Human
GRIT [42]	91M	137M	115M	COYO-700M, LAION-2B	Model	Web + Model
COVLM-97M [42]	97M	-	-	BLIP-2	Model	Web + Model
VLM-VG-COCO	81K	143K	2M	COCO*	Human	Model
VLM-VG-O365	431K	958K	14.2M	O365	Human	Model
VLM-VG (Ours)	512K	1.1M	16.2M	COCO*, O365	Human	Model

Table 1. Compare VLM-VG with existing visual grounding datasets. COCO* denotes the subset of COCO that excludes images used by RefCOCO/+/g. Different from all the existing grounding datasets, our VLM-VG datasets uses pure model model-genereated text annotations without relying on manual text labeling.

out the need of human annotations. Notably, VLM-VG is the first dataset where the texts are purely annotated by model with high quality region-text alignment. Sourcing from COCO and Objects365, VLM-VG contains 512K images, 1.1M objects and 16.2M referring expressions (5.6 words on average), averaging 14.7 referring expressions per object. This abundance of comprehensive and diverse descriptions facilitates a more accurate modeling of real-world referring expression, thereby enhancing the effectiveness and generalizability of visual grounding models in complex practical scenarios. The detailed category distribution of VG-VLM is shown in Figure A1.

3.4. Visual Grounding: Task and Model

To evaluate the quality of the generated dataset, we utilize a simple and lightweight model proposed by [25] on two visual grounding tasks: referring expression comprehension (REC) and segmentation (RES). In the REC/RES task, the model takes in an image and a language query about a specific object in the image to generate **one** bounding box/mask of the referred object.

The model adopts a simple and extensible architecture design: it includes a ResNet [15] as the image encoder to produce multi-level features, T5 [45] as the text encoder, a multi-level multimodal fusion network [25], and box/mask prediction heads [14]. The fusion network uses cross-attention and feature product to fuse the multi-level image and text features. After the vision-language fusion, a standard region proposer [46] and a box/mask prediction heads are applied to decode the predicted box/mask [25]. During training, the model is optimized by the box regression loss and optionally mask loss [14]. Following the common practice, we initialize the ResNet [15] backbone with checkpoints pretrained on the COCO detection task and the T5 encoder with language modeling pretrained checkpoint. We use a batch size of 1024, image size 384×384, total training steps 200K, Momentum optimizer with initial learning rate 0.04 and step decay at 70% and 90% of total steps by a factor of 0.1 and linear warmup of 500 steps. Due to the initialization, we set both image and text encoder learning rate to 0.1 of the initial learning rate. We also apply image scale jittering with a factor S uniformly sampled from $S \sim Uniform(0.5, 2.0)$ at training time.

For the RES task, since the segmentation datasets are generally less scalable than detection datasets, we propose to treat REC as the pre-training task of RES and initialize the grounding with our pre-trained REC models. During RES training, only mask head is fine-tuned while the rest of the model is frozen. We use batch size 64, learning rate 0.01 and 50K training steps for the RES task.

4. Experiments

4.1. Experiment Setting

To quantitatively evaluate the effectiveness of our human-free grounding annotations, we train the grounding model with constructed dataset for REC and RES task respectively. For REC task, we train on the full-set of VLM-VG combing both COCO and Objects365 splits, for RES task, we only train on VLM-VG-COCO since Objects365 doesn't contain mask annotations. After training, we perform zero-shot REC and RES evaluation on the three popular visual grounding benchmarks: RefCOCO [64], RefCOCO+ [64] and RefCOCOg [38]. For REC task, we report the top-1 accuracy where the predicted box is correct if its IOU with the ground-truth bounding box is greater than 0.5. For the RES task we report the overall IOU (oIOU) and mean IOU (mIOU) following the norms of the community.

4.2. Referring Expression Comprehension

We evaluate and report zero-shot referring expression comprehension performance on RefCOCO/+/g datasets and compare with state-of-the-art zero-shot methods in Table 2. The simple grounding model based on ResNet-101 trained on our proposed VLM-VG dataset achieves best average performance among the three datasets and 8 splits comparing with current SoTA models including two LLM-based models [27, 42], methods trained on ground-

Method	Vison Backbone	Pretraining Dataset	RefCOCO			RefCOCO+			RefCOCOg		Avg
			val	testA	testB	val	testA	testB	val	test	
CPT [62]	ResNeXt-152	OpenImages, COCO, O365, VG	32.2	36.1	30.3	31.9	35.2	28.8	36.7	36.5	33.5
ReCLIP* [52]	ViT-B+R50	CLIP, COCO	45.8	46.7	45.2	45.3	48.5	42.7	57.0	56.2	48.4
Red Circle* [50]	ViT-L+R50	CLIP,COCO	49.8	58.6	39.9	**55.3**	**63.9**	45.4	59.4	58.9	53.9
RelVLA [13]	ViT-B	PMD, HICO, SWiG, VG, COCO	52.5	52.7	52.9	50.8	53.4	<u>47.6</u>	61.3	60.9	54.0
VGDiffZero [34]	VAE	LAION-5B	28.0	30.3	29.1	28.4	30.8	29.8	33.5	33.2	33.9
Pseudo-Q† [19]	R50	VG, RefC-pseudo	56.0	58.3	54.1	38.9	45.1	32.1	49.8	47.4	47.7
GLIP [29]	Swin-T	O365, GoldG, Cap4M	50.4	54.3	43.8	49.6	52.8	44.6	<u>66.1</u>	<u>66.9</u>	53.6
Grounding-Dino [33]	Swin-T	O365,GoldG	50.4	57.2	43.2	51.4	57.6	45.8	**67.5**	**67.1**	55.0
MM-G [68]	Swin-T	O365, GoldG, GRIT, V3Det	53.1	59.1	46.8	52.7	58.7	**48.4**	62.9	62.9	55.6
Kosmos-2 [42]	ViT-L	KOSMOS-1, GRIT	52.3	57.4	47.3	45.5	50.7	42.2	60.6	61.7	52.2
CoVLM [27]	ViT-L	Pythia, CoVLM-97M	48.2	53.2	43.2	47.6	50.9	44.2	60.9	61.9	51.3
Ours	R50	COCO, VLM-VG	<u>60.8</u>	<u>67.1</u>	<u>54.9</u>	52.5	59.6	43.3	61.4	61.2	<u>57.6</u>
Ours	R101	COCO, VLM-VG	**63.4**	**68.5**	**57.6**	<u>53.9</u>	<u>60.9</u>	44.9	63.3	63.2	**59.5**

Table 2. Comparison with state-of-the-art methods on zero-shot referring expression comprehension (REC) tasks on RefCOCO/+/g dataset. Methods shown in top columns take pre-extracted proposals from FRCNN-101 as input, bottom methods take in images. The best two results are **bold-faced** and <u>underlined</u>, respectively. * denotes methods using model ensembling, Pseudo-Q† is directly trained on RefCOCO/+/g images, blue indicates grounding/localization datasets.

ing datasets with manual text labeling [13, 29, 33, 62, 68], and weakly-supervised method [19] utilizing images from RefCOCO/+/g. In particular, on RefCOCO splits which require spatial relationship modeling, our model largely outperforms the SOTA results by up to 7.4 points. On the challenging RefCOCO+ splits, we achieve the second best performance while the SoTA model Red Circle [50] uses model ensembling during inference. It's also worth noting that after switching to the smaller ResNet-50 backbone, our method still maintains a highly-competitive performance that is better than all other baseline models including methods take in object proposals from pre-trained Faster R-CNN based on ResNet-101. The strong performance achieved by the light-weight grounding models on the REC task demonstrates the quality of VLM-VG grounding annotations and the effectiveness to scale grounding pre-training with our proposed automatic annotation pipeline.

4.3. Referring expression Segmentation

We also conduct zero-shot evaluation on the referring expression segmentation (RES) task and compare with state-of-the-arts models by both mIOU and oIOU metrics on RefCOCO/+/g benchmarks in Table 3. We initialize the RES model with the pre-trained REC model from Sec 4.2 and tune the model on VLM-VG-COCO. We almost achieve best performance on all three datasets and largely outperform the existing methods, which a 5.1% and 4.1% of average improvement on oIOU and mIOU respectively with compact model designs, comparing with the state-of-the-arts models [54] utilizing large-scale segmentation pretrained SAM [23]. Particularly, we observe the similarly strong improvement on RefCOCO as REC results shows, achieving up to 13% improvement comparing with previous SoTA, further demonstrating the effectiveness of our proposed referring expression annotation pipelines. Similar to REC results, we observe the compact ResNet-50 based model also significantly outperforms existing methods.

4.4. Scaling beyond Detection Datasets

To evaluate the robustness and adaptability of our approach in leveraging generative Vision and Language Models (VLMs) for scalable visual grounding annotation, we further explore the use of images from the WebLI dataset [7] filtered to include images with high-quality text descriptions according to [18]. In WebLI datast, each image is annotated with noisy bounding boxes generated by OWL-ViT [39]. We selected bounding boxes with an OWL-ViT confidence score exceeding 0.8 to ensure a reasonable quality of visual data. From this subset, a total of 100,000 bounding boxes were randomly sampled and subsequently annotated by the PaLI-3 [6] model, which generated captions for these visual regions (type 1 annotations). Table 5 shows the zero-shot referring expression comprehension (REC) accuracy of our model on validation sets of RefCOCO splits. The training settings follow the ablation setups of the main paper (512 batch size for 50K iterations). The results show significant performance gains with respect to the increase of the dataset size. This demonstrates the promising potential of using web-collected billion-scale data to further scale up visual grounding datasets and models.

4.5. Ablations and Analysis

Dataset Design. In order to achieve a fine-grained understanding of the contribution of each dataset component, we conduct ablation study to evaluate the generated dataset on REC task step by step. We discuss five dataset variants where each variant is a subset of the next variant from top to down as demonstrated in Table 4.

Initially, we construct dataset variant (a) exclusively using COCO, solely utilizing the Type 1 regional captions,

Metric	Method	Vison Backbone	Pre-training Dataset	RefCOCO			RefCOCO+			RefCOCOg		Avg
				val	testA	testB	val	testA	testB	val	test	
oIOU	Grad-CAM [65]	R50	COCO†, CLIP	14.0	15.1	13.5	14.5	15.0	14.0	12.5	12.8	13.9
	Score map [65]	R50	COCO†, CLIP	19.9	19.3	20.2	20.4	19.7	20.8	18.9	19.2	19.8
	Region Token [65]	ViT-B	COCO†, CLIP	21.7	20.3	22.6	22.6	20.9	23.5	25.5	25.4	22.8
	Cropping [65]	R50	COCO†, CLIP	22.4	20.5	22.7	24.0	22.0	23.5	28.2	27.6	23.9
	Global-Local [65]	R50	COCO†, CLIP	24.6	23.4	24.4	25.9	24.6	25.6	30.1	29.8	26.1
	SAM-CLIP [41]	ViT-B	SAM, CLIP	25.2	25.9	24.8	25.6	27.8	26.1	33.8	34.8	28.0
	Ref-Diff [41]	VAE	LAION-5B	35.2	37.4	34.50	35.6	38.7	**31.4**	38.6	37.5	36.1
	Tas [54]	ViT-B	SAM, BLIP‡, LAION, CLIP	29.5	30.3	28.2	33.2	38.8	28.0	35.8	36.2	32.5
	Ours	R50	COCO, VLM-VG	43.2	46.4	39.6	35.7	39.2	29.4	42.1	42.9	39.8
	Ours	R101	COCO, VLM-VG	**45.4**	**48.0**	**41.4**	**37.0**	**40.7**	30.5	**42.8**	**44.1**	**41.2**
mIOU	Grad-CAM [65]	R50	COCO†, CLIP	14.0	15.1	13.5	14.5	15.0	14.0	12.5	12.8	13.9
	Score map [65]	R50	COCO†, CLIP	14.2	15.9	13.2	14.8	15.9	13.8	12.5	13.2	14.2
	Region Token [65]	ViT-B	COCO†, CLIP	23.4	22.1	24.6	24.5	22.6	25.4	27.6	27.3	24.7
	Cropping [65]	R50	COCO†, CLIP	24.3	22.4	24.7	26.3	23.9	25.7	31.3	30.9	26.2
	Global-Local [65]	R50	COCO†, CLIP	26.7	25.0	26.5	28.2	26.5	27.9	33.0	33.1	28.4
	CaR [53]	ViT-B, ViT-L	CLIP	33.6	35.4	30.5	34.2	36.0	31.0	36.7	36.6	34.3
	SAM-CLIP [41]	ViT-B	SAM, CLIP	26.3	25.8	26.4	25.7	28.0	26.8	38.8	38.9	29.6
	Ref-Diff [41]	VAE	LAION-5B	37.2	38.4	37.2	37.3	40.5	33.0	44.0	44.5	39.0
	Tas [54]	ViT-B	SAM, BLIP‡, LAION, CLIP	39.8	41.1	36.2	**43.6**	**49.1**	**36.5**	46.6	46.8	42.5
	Ours	R50	COCO, VLM-VG	47.7	51.8	44.7	41.2	45.9	34.7	46.6	47.1	45.0
	Ours	R101	COCO, VLM-VG	**49.9**	**53.1**	**46.7**	42.7	47.3	36.2	**48.0**	**48.5**	**46.6**

Table 3. Comparison with state-of-the-art methods on zero-shot referring expression segmentation (RES) tasks on RefCOCO/+/g dataset. The best two results are **bold-faced** and underlined, respectively. COCO† means method only uses images from COCO without annotations, BLIP‡ [28] contains grounding/localization dataset including COCO and VG, Blue indicates grounding/localization datasets.

as elaborated in Sec.3.2. The grounding model pretrained on variant (a) achieves an average performance of 46.2%, which stands in comparison with [19], a semi-supervised model trained on RefCOCO/+/g images and pseudo queries generated via rule-based methods. For variant (b), we augment the data source by incorporating images from Objects365 v1 dataset in addition to COCO compared to variant (a). By combining COCO and Objects365, we observe notable improvement across all three datasets, achieving an average performance of 51.7% with an improvement of 5.5%. This shows the substantial benefits of scaling grounding pretraining even with out-of-domain image sources such as Objects365 to benefits for model's generalizability. Variant (c) builds upon variant (b) by incorporating horizontal relation expressions (Relation v1) from Type 2, resulting in a significant performance improvement of up to 13 points on RefCOCO splits, which necessitate spatial relationship modeling. Interestingly, we also observe a consistent performance enhancement on RefCOCO+/g splits that don't contain positional references. Build on variant (c), variant (d) adds spatial modeling across vertical and depth dimensions (Relation v2) and further enhance the average performance to 58.7% and outperforms all the existing SoTA methods. Finally, following Type 3 in Sec.3.2, we further investigate whether explicit attribute modeling via prompting the generative VLMs can fix the potential missing details from the "general region captions" (type 1). As variant (e) shows, adding attribute-rich descriptions that incorporate supplementary knowledge from human and LLMs only yields limited improvement compared to (d). This suggests that generative VLMs, such as PaLI-3, possess the capability to generate high-quality and detailed descriptions for object-centric image regions that cover typical attributes and benefit scalable and generalizable visual grounding in the wild.

Transfer from REC to RES. Compared with detection dataset, segmentation dataset usually are harder to scale due to the cost of mask annotations. We thus investigate if pre-training on the REC task can further benefit to the RES task. We validate the influence of initializing the RES model with pre-trained REC checkpoints. Results in Table 7 shows that after initializing the model from pre-trained REC checkpoint, the overall RES performance achieves solid improvement especially for the RefCOCO+ splits, this shows VLM-VG data can significantly benefit the pixel-level zero-shot RES task by grounding knowledge transfer.

Text Annotation Source. We conduct ablations on the individual text annotations sources to investigate the influence of referring expression construction methods on COCO dataset. We construct text annotations using 4 different methods: 1) category name directly using class label provided by COCO dataset. 2) prompted caption generated by Type 2 in Sec. 3.2. 3) No-prompt caption generated using same method as 2) without providing any language prompts to PaLI-3. 4) Attribute-rich descriptions that constructed by Type 3 in Sec. 3.2. We experiment on a small training scale setting (512 batch size for 50K iterations) for efficiency. Results on REC validation set are shown in Table 6. We can see directly using category label as the referring expression produce a relative low performance due to the absence of detailed descriptions. Using the regional captions generated by prompting the VLM, the grounding

Variant	Source Dataset	referring expression	RefCOCO val	RefCOCO testA	RefCOCO testB	RefCOCO+ val	RefCOCO+ testA	RefCOCO+ testB	RefCOCOg val	RefCOCOg test	Avg
(a)	COCO	Regional Cap.	43.1	48.4	37.1	46.0	50.4	39.8	52.3	52.3	46.2
(b)	COCO, O365	Regional Cap.	47.4	50.9	42.0	51.4	54.5	45.9	60.7	60.9	51.7 (+5.5)
(c)	COCO, O365	Regional Cap. + Rel. v1	59.0	63.1	52.5	51.9	56.7	44.1	62.6	62.8	56.6 (+4.9)
(d)	COCO, O365	Regional Cap. + Rel. v2	63.0	66.3	**58.9**	52.4	56.7	**46.0**	**63.6**	62.8	58.7 (+2.1)
(e)	COCO, O365	Regional Cap. + Rel. v2 + Attr.	**63.4**	**68.5**	57.6	**53.9**	**60.9**	44.9	63.3	**63.2**	**59.5** (+0.8)

Table 4. **Ablation on dataset designs**. We compare the influence of the dataset designs including source dataset and referring expression types. In reffering expression, Regional Cap. denotes the Type 1 annotations in Sec.3.2, Rel. v1 denotes the horizontal relations in Type 2 and Rel. v2 represents all the three spatial relations. Attr. means the Type 3 attribute-rich descriptions.

Data size	RefCOCO	RefCOCO+	RefCOCOg	Avg
10k	19.0	19.8	21.0	20.4
30k	23.6	24.1	26.1	25.1
100k	26.0	27.9	30.8	29.4

Table 5. **REC with WebLI dataset.**

Data Source	RefCOCO	RefCOCO+	RefCOCOg
Category name	24.1	24.2	26.4
Prompted Caption	32.2	34.8	41.7
No-prompt Caption	27.3	29.6	40.1
Attributes	34.0	36.2	39.0

Table 6. **Ablation on text annotation source.**

Metric	REC Pt.	RefCOCO val	RefCOCO testA	RefCOCO testB	RefCOCO+ val	RefCOCO+ testA	RefCOCO+ testB	RefCOCOg val	RefCOCOg test	Avg
oIOU		43.1	43.0	**42.8**	28.9	30.3	26.4	38.2	39.1	36.5
	✔	**45.4**	**48.0**	41.4	**37.0**	**40.7**	**30.5**	**42.8**	**44.1**	**41.2**
mIOU		48.5	48.9	**49.2**	35.0	37.1	33.1	43.5	44.6	42.5
	✔	**49.9**	**53.1**	46.7	**42.7**	**47.3**	**36.2**	**48.0**	**48.5**	**46.6**

Table 7. **Ablation on REC pre-training for RES.**

Mixture Ratio	RefCOCO	RefCOCO+	RefCOCOg	Avg
1:1:1:1	50.8	47.7	55.3	51.2
1:2:1:2	48.6	46.8	52.6	49.4
2:2:1:1	52.3	46.2	53.5	50.7
1:1:2:2	49.7	48.4	55.0	51.0
2:1:2:1	52.2	47.1	54.4	51.3
1:2:1:2	48.6	46.8	52.6	49.4

Table 8. **Ablation on data mixture ratio.**

model achieves a huge improvement compared with the category label, demonstrating the abundant additional knowledge gained from the large generative VLMs. Besides, we also compare the influence of prompts during the regional caption generation in line 2 and line 3. We show that designing a proper instruction as the prompt do benefit the quality of generated captions and further enhance the grounding model's performance. Finally, by comparing the second row and the last row, we show that the generic regional captions can achieve similar performance with attribute-rich descriptions involving cross-domain knowledge, further showing the generative VLM's capability to provide informative and comprehensive descriptions of objects by "generic" instruction prompts.

Data Mixture Ratio. We conduct ablations on the data mixture ratios among the constituent datasets of VLM-VG in Table 8. The mixture ratio is given by (COCO-common, COCO-attribute, Ojbects365-common, Objects365-attributes) where the common datasets denote the combination of datasets from Type 1 and Type 2 in Sec. 3.2 while attributes dataset represent the attribute-rich descriptions from step 3. We conduct experiment on the same small setting on REC validation set. Results show that a simple 1:1:1:1 ratio leads to very strong performance compared to other ratios, so we adopt it directly.

Inference Time. We measure the inference time of our model with R50 backbone on 384×384 image to be 60ms on a single 1080-Ti GPU, which is in the same ballpark of existing REC specialist models [9, 37, 61].

4.6. Visualization and Analysis

We conduct extensive visualization and qualitative analysis of the VLM-VG dataset with respect to its quality and diversity as well as the REC and RES prediction results in the supplementary materials. The visualization results further demonstrate the quality and effectiveness of our proposed method and dataset for zero-shot visual grounding.

5. Conclusion

We present VLM-VG, a large-scale visual grounding dataset built by prompting generative VLM to generate region captions for scaling up visual grounding data. Despite training on image-text data, we observe that existing generative VLMs can produce quality region captions when prompted with object-centric crops from detection datasets and appropriate text instructions. In addition, we propose spatial relation modeling and explicit attribute modeling to capture the linguistic patterns of referring expression. Our VLM-VG dataset include 500K images, 1M objects, and 16M text queries. It is one of the largest grounding datasets to date, and the first with purely machine-generated texts. We demonstrate the advantage of VLM-VG by zero-shot transfer to referring expression comprehension (REC) and segmentation (RES) tasks on RefCOCO benchmarks. Our simple, lightweight model significantly outperforms the state-of-the-art approaches trained on more human-annotated data on both tasks. We hope these findings encourage the community to explore generative VLM for scaling up visual grounding dataset in the real world.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 4

[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] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2

[4] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset`, 2022. 4

[5] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 2

[6] Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, et al. Pali-3 vision language models: Smaller, faster, stronger. *arXiv preprint arXiv:2310.09199*, 2023. 1, 3, 4, 6

[7] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme Ruiz, Andreas Peter Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. PaLI: A jointly-scaled multilingual language-image model. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2, 6

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

[9] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1769–1779, 2021. 8

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 2

[11] Yunhao Ge, Xiaohui Zeng, Jacob Samuel Huffman, Tsung-Yi Lin, Ming-Yu Liu, and Yin Cui. Visual fact checker: Enabling high-fidelity detailed caption generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14033–14042, 2024. 12

[12] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 1, 13

[13] Zeyu Han, Fangrui Zhu, Qianru Lao, and Huaizu Jiang. Zero-shot referring expression comprehension via structural similarity between images and captions. *arXiv preprint arXiv:2311.17048*, 2023. 2, 6

[14] 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. 2, 5

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[16] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python. 2020. 4

[17] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[18] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 6

[19] Haojun Jiang, Yuanze Lin, Dongchen Han, Shiji Song, and Gao Huang. Pseudo-q: Generating pseudo language queries for visual grounding. In *Proceedings of the IEEE/CVF CVPR*, 2022. 2, 3, 4, 6, 7

[20] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. 2, 4

[21] 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. 1, 2

[22] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Lan-*

guage Processing (EMNLP), pages 787–798, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. 4, 5

[23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2, 6

[24] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 2, 4, 5

[25] Weicheng Kuo, Fred Bertsch, Wei Li, AJ Piergiovanni, Mohammad Saffar, and Anelia Angelova. Findit: Generalized localization with natural language queries. In *European Conference on Computer Vision*, pages 502–520. Springer, 2022. 2, 5

[26] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 1, 2

[27] Junyan Li, Delin Chen, Yining Hong, Zhenfang Chen, Peihao Chen, Yikang Shen, and Chuang Gan. CoVLM: Composing visual entities and relationships in large language models via communicative decoding. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 4, 5, 6

[28] 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, 7

[29] Liunian Harold Li*, Pengchuan Zhang*, Haotian Zhang*, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, 2022. 2, 4, 6

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1, 2, 4

[31] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23592–23601, 2023. 2

[32] Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L Yuille. Clevr-ref+: Diagnosing visual reasoning with referring expressions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4185–4194, 2019. 2

[33] 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. 2, 6

[34] Xuyang Liu, Siteng Huang, Yachen Kang, Honggang Chen, and Donglin Wang. Vgdiffzero: Text-to-image diffusion models can be zero-shot visual grounders. *arXiv preprint arXiv:2309.01141*, 2023. 2, 6

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

[36] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. UNIFIED-IO: A unified model for vision, language, and multi-modal tasks. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2

[37] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 8

[38] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016. 1, 2, 4, 5

[39] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European Conference on Computer Vision*, pages 728–755. Springer, 2022. 6

[40] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *European Conference on Computer Vision*, pages 792–807. Springer, 2016. 1

[41] Minheng Ni, Yabo Zhang, Kailai Feng, Xiaoming Li, Yiwen Guo, and Wangmeng Zuo. Ref-diff: Zero-shot referring image segmentation with generative models. *arXiv preprint arXiv:2308.16777*, 2023. 2, 7

[42] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 2, 4, 5, 6

[43] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 2, 4, 5

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

[45] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with

a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 5

[46] 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. 2, 5

[47] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[48] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 1, 4

[49] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 1, 2, 4

[50] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11987–11997, 2023. 2, 6

[51] 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. 2

[52] Sanjay Subramanian, William Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5198–5215, 2022. 2, 3, 6

[53] Shuyang Sun, Runjia Li, Philip Torr, Xiuye Gu, and Siyang Li. Clip as rnn: Segment countless visual concepts without training endeavor. *arXiv preprint arXiv:2312.07661*, 2023. 7

[54] Yucheng Suo, Linchao Zhu, and Yi Yang. Text augmented spatial aware zero-shot referring image segmentation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1032–1043, Singapore, Dec. 2023. Association for Computational Linguistics. 6, 7

[55] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. 2

[56] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 2

[57] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 2

[58] Yue Cao Bin Li Lewei Lu Furu Wei Jifeng Dai Weijie Su, Xizhou Zhu. Vl-bert: Pre-training of generic visual-linguistic representations. In *ICLR*, 2020. 2

[59] Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. Phrasecut: Language-based image segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10216–10225, 2020. 2

[60] Jiarui Xu, Xingyi Zhou, Shen Yan, Xiuye Gu, Anurag Arnab, Chen Sun, Xiaolong Wang, and Cordelia Schmid. Pixel-aligned language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13030–13039, 2024. 2

[61] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4683–4693, 2019. 8

[62] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797*, 2021. 2, 6

[63] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[64] 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. 1, 2, 4, 5

[65] Seonghoon Yu, Paul Hongsuck Seo, and Jeany Son. Zero-shot referring image segmentation with global-local context features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19456–19465, 2023. 2, 7

[66] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12104–12113, 2022. 1

[67] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. *arXiv preprint arXiv:2101.00529*, 2021. 2

[68] Xiangyu Zhao, Yicheng Chen, Shilin Xu, Xiangtai Li, Xinjiang Wang, Yining Li, and Haian Huang. An open and comprehensive pipeline for unified object grounding and detection. *arXiv preprint arXiv:2401.02361*, 2024. 2, 6

[69] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. *AAAI*, 2020. 2