SynGround: Learning from Synthetic Data for Visual Grounding

Ruozhen He[1]
catherine.he@rice.edu
Ziyan Yang[1]
zy47@rice.edu
Paola Cascante-Bonilla[2]
paola@cs.stonybrook.edu
Alexander C. Berg[3]
bergac@uci.edu
Vicente Ordonez[1]
vicenteor@rice.edu

[1] Rice University, Houston, Texas, USA
[2] Stony Brook University, Stony Brook, New York, USA
[3] University of California, Irvine, Irvine, California, USA

Abstract

We introduce SynGround, a novel approach to learn representations for visual grounding using a combination of real data along with synthetic images, synthetic referring expressions, and corresponding bounding boxes. We explore various strategies to best generate additional image-text pairs and image-text-box triplets using pretrained models under different settings and varying degrees of reliance on real data. Through comparative analyses with synthetic, real, and web-crawled data, we identify factors that contribute to performance differences. We find that SynGround can improve the localization capabilities of a vision-and-language model and offers the potential for arbitrarily large scale data generation. Particularly, data generated with SynGround improves the pointing game accuracy of a pretrained ALBEF model by 4.81% and improves BLIP by 17.11% absolute percentage points on average across RefCOCO+ and Flickr30k benchmarks.

1 Introduction

Vision-and-language models (VLMs) pretrained on image-text pairs have become exceedingly accurate across various tasks [6, 31, 38, 41, 42, 45, 47, 53, 58]. Leveraging web-sourced datasets, these models showcase a strong ability to comprehend and process an extensive vocabulary of objects and scenes. Our work focuses on the task of visual grounding –also known as *referring expression comprehension*– which consists of mapping arbitrary input text to image regions. To solve this task, recent methods finetune vision-and-language models pretrained on web-scale image-text pairs using a large but more modest number of images annotated with bounding boxes or other region annotations or directly use pretrained object detectors [9, 11, 20, 25, 32, 33, 42, 55, 57]. The resulting VLMs can then be used to perform visual grounding over an arbitrary vocabulary of objects.

Collecting annotations for tasks that require localizing objects is considerably more expensive than for other tasks. Region annotations in the form of bounding boxes or segments can not be easily obtained from the web in the same way that image-text pairs can be found, and require more cognitive effort to annotate manually than providing a single label. For instance, curating the widely used Visual Genome [36] dataset involved contributions from 33,000 unique workers over 6 months, following 15 months of experimentation and refinement of the data representation. Manual labeling restricts the scale of such region-annotated data compared to image-text datasets [69]. Recent work has championed the use of synthetic data, also referred to as *learning from models*, even for tasks that require only image-text pair supervision [74]. Our work takes this paradigm one step further by investigating whether synthetic data obtained from models is ready to make improvements for visual grounding.

This paper leverages recent advancements in text-to-image generation [52, 61, 67], large language modeling [12, 76] and models for other vision-and-language tasks [40, 42, 44] to design an effective pipeline to supervise vision-and-language models for visual grounding. We refer to this pipeline as *SynGround* and present a systematic analysis to justify each stage of our data generation process. While there have been several attempts to train visual recognition models with synthetic data by using automatically generated image-text pairs [9, 21, 27, 68, 74, 75], our work is the first to use generative models to synthesize grounded image data. Moreover, we assess the efficacy of synthetic data by comparing performance with the use of real and web-crawled data, identifying specific limiting factors. We also investigate whether synthetic data can augment real data and examine its scalability.

Our findings and contributions are summarized as follows: (1) For a text-to-image generative model, detailed prompts obtained from image captioners yield the most effective synthetic image-text pairs for visual grounding, surpassing those generated from concatenated region descriptions or LLM-generated text. (2) To obtain synthetic image-text-boxes, both layout-conditioned generative models and object detectors using synthetic image-text pairs show promise. However, layout-conditioned models are more limited. (3) We use our findings to propose SynGround, an effective pipeline to synthesize image-text-boxes for visual grounding. This method leverages exhaustive image descriptions for image synthesis, an LLM for text synthesis from phrase extraction, and an open-vocabulary object detector for bounding box generation. (4) Our results show that the use of our generated synthetic data outperforms using web-crawled data (*see supp.*). Additionally, our synthetic data can augment real data (Sec. 3.3) and shows an upward trend in terms of scalability (Sec. 4).

2 Related Work

Visual Grounding. Visual grounding generally refers to tasks that require associating text descriptions with relevant regions within images. Supervised methods are typically trained with images associated with text descriptions and corresponding box or region annotations [16, 17, 20, 33, 87], or integrate pretrained object detectors [26, 59] to identify the most relevant regions with respect to input text descriptions that refer to objects or regions [9, 14, 23, 25, 46, 80]. While weakly-supervised methods bypass the need for bounding boxes [3, 28, 71, 72], they rely on datasets such as Visual Genome [36], which provides multiple phrases describing various regions in each image. However, the process of manually annotating dense text descriptions and their corresponding boxes is labor intensive. Although some studies collect more data [83] or generate annotations for existing image-text datasets [56, 82, 89], our contribution is orthogonal as we aim to investigate the feasibility and limitations of gen-

erating and using synthetic data. Related to grounding methods incorporating tuning of visual explanations [28, 38, 84, 87], we explore visual grounding in a more general context, aiming to localize phrases using gradient-based model explanations (i.e. GradCAM [70]) rather than generating boxes [42]. Compared to boxes, explanation maps provide a more flexible representation that can be used to point to multiple objects or background regions. A recent grounding method in this category is AMC [87], which proposed an attention mask consistency objective to optimize the gradient-based explanations of ALBEF [38] to improve localization performance. We adopt ALBEF as one of our main base models and tune it with attention mask consistency on image-text-boxes as a testbed for our generated data.

Learning from Synthetic Data. The use of synthetic data has been widely explored across various computer vision tasks, including image classification [22, 50, 55], semantic segmentation [10, 60, 62], object detection [64, 65], human pose estimation [35, 78], action recognition [79], and many other domains [1, 13, 29, 37, 48, 49, 51, 63, 64, 77, 86, 88]. In contrast to works that generate synthetic data using 3D-rendering [24, 91] or physically realistic engines [6, 7, 15, 19, 72], our approach aligns more closely with research adopting diffusion models. He *et al.* [27] use GLIDE [52] for generating synthetic images to improve a pretrained CLIP model [58] in zero-shot and few-shot classification, while its performance is adversely affected when trained from scratch on synthetic data. Azizi *et al.* [4] fine-tune Imagen [67] on ImageNet [66] and subsequently leverage its synthetic data to augment the real ImageNet training set, resulting in initial improvement followed by degradation upon scaling up. Fan *et al.* [21] investigate the scaling laws of synthetic images and identify related factors. StableRep [75] proposes a self-supervised method with a multi-positive contrastive loss that learns representations from synthetic images generated for captions in large-scale datasets [8, 18], thereby boosting linear probing image classification performance. SynCLR [74] uses LLM-generated synthetic captions. Our research not only generates image-text pairs but also provides corresponding synthetic boxes, facilitating a comprehensive exploration of the efficacy of synthetic image-text-box triplets in visual grounding.

3 Methodology

We investigate effective strategies to generate image-text-boxes $\langle I, T, B \rangle$ to improve the visual grounding ability of VLMs. We assume our base VLM comprises a text encoder ϕ_t, a visual encoder ϕ_v, and a multimodal fusion encoder ϕ_f. We first introduce the objectives for tuning a base model on image-text pairs $\langle I, T \rangle$ and image-text-box triplets $\langle I, T, B \rangle$. Sec. 3.1 explores various image-text synthesis strategies with an image generator Ψ_g, while Sec. 3.2 delves into multiple approaches for box synthesis. In the following sections, we conduct extensive experiments and analyses with our proposed image-text-box synthesis, SynGround, which integrates an image caption generator Ψ_c, a text-to-image generator Ψ_g, a large language model Ψ_t, and an object detector Ψ_d. Sec. 3.3 shows evaluation of SynGround when combined with Real Data.

Preliminaries and Setup. We adopt ALBEF [38] as our base model to explain our methodology, as it serves as the main backbone for the state-of-the-art fully-supervised and weakly-supervised visual grounding models [28, 87] on benchmarks such as Flickr30k and RefCOCO+. However, we later also adapt and validate the effectiveness of SynGround using multiple VLMs including some that output boxes instead of heatmaps (Sec. 4 and *supp.*).

Image-Text Matching. Image-text matching objectives include a standard image-text matching loss ($\mathcal{L}_{\text{itm}}$), an image-text contrastive loss ($\mathcal{L}_{\text{itc}}$), and a masking language modeling loss

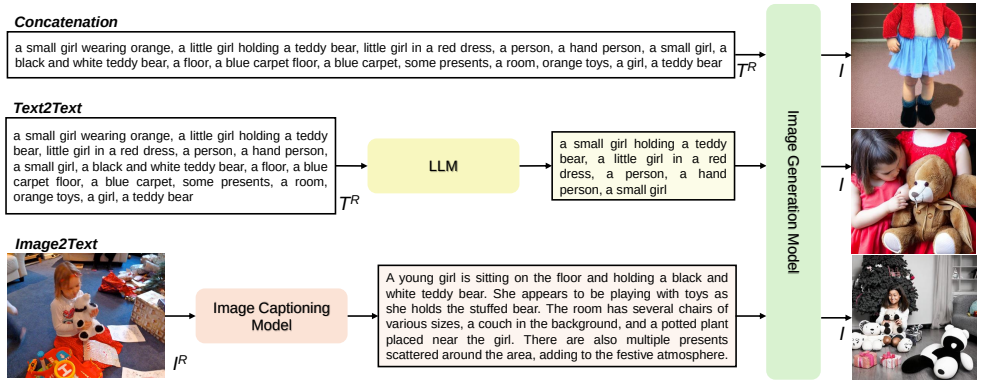

Figure 1: Various approaches for obtaining text descriptions to use as input prompts for an image generator, resulting in image-text pairs. 1) **Concatenation:** Descriptions are generated by concatenating real text, 2) **Text2Text:** LLM summary on real text, and 3) **Image2Text:** Descriptions are produced from an image captioner.

($\mathcal{L}_{\text{mlm}}$). The image-text matching loss $\mathcal{L}_{\text{itm}}$ evaluates the compatibility between an image-text pair $\langle I, T \rangle$ by tuning the output of a `[CLS]` token using a cross-entropy loss. The image-text contrastive loss $\mathcal{L}_{\text{itc}}$ aligns visual and textual representations using contrastive learning by sampling a set of negative image-text pairs and using a temperature scaling parameter to normalize the scores. The masking language modeling loss $\mathcal{L}_{\text{mlm}}$ uses both visual inputs and textual context to predict masked tokens from the input text. The overall objective to tune the base model on image-text pairs is $\mathcal{L}_{\text{vl}} = \mathcal{L}_{\text{itm}} + \mathcal{L}_{\text{itc}} + \mathcal{L}_{\text{mlm}}$.

Image-Text-Box Matching. We adopt an attention mask consistency objective $\mathcal{L}_{\text{amc}}$ to add box supervision on top of the ALBEF model [37]. It uses gradient-based explanation heatmaps G through GradCAM [70], and maximizes the consistency between this map and region annotations. This objective considers two terms. $\mathcal{L}_{\text{max}}$ encourages the maximum value of G inside a target box B to surpass the maximum value outside by a margin δ_1.

$$\mathcal{L}_{\text{max}} = \mathop{\mathbb{E}}_{(I,T,B)\sim D}\left[\max(0, \max_{i,j}\left((1-B_{i,j})\,G_{i,j}\right) - \max_{i,j}\left(B_{i,j}G_{i,j}\right) + \delta_1)\right],$$

where $B_{i,j}$ is 1 when pixel location i,j is inside the box, and 0 otherwise. $\mathcal{L}_{\text{mean}}$ encourages the heatmap G to have a higher mean inside the box than outside by a margin δ_2.

$$\mathcal{L}_{\text{mean}} = \mathop{\mathbb{E}}_{(I,T,B)\sim D}\left[\max(0, \frac{\sum_{i,j}(1-B_{i,j})\,G_{i,j}}{\sum_{i,j}(1-B_{i,j})} - \frac{\sum_{i,j}B_{i,j}G_{i,j}}{\sum_{i,j}(B_{i,j})} + \delta_2)\right].$$

The full $\mathcal{L}_{\text{amc}}$ objective is $\mathcal{L}_{\text{amc}} = \mathcal{L}_{\text{max}} + \lambda \cdot \mathcal{L}_{\text{mean}}$, where λ is a trade-off hyperparameter. The base model is tuned with both $\mathcal{L}_{\text{vl}}$ and $\mathcal{L}_{\text{amc}}$ on image-text-box triplets.

Visual Grounding Evaluation. Following prior work that predicts heatmaps, our evaluation uses pointing game accuracy, which measures the proportion of instances where the maximal activation point within generated heatmaps correctly falls within the annotated ground-truth box regions [2, 14, 20, 25, 28, 38, 46, 87]. We conduct evaluation across multiple benchmarks, including RefCOCO+ [90] and Flickr30k [57].

Table 1: Comparisons of image-text synthesis strategies. We assess the effectiveness of synthetic image-text pairs from text concatenation, Text2Text, and Image2Text pipelines, by evaluating the performance improvements over an ALBEF model. For reference we also include the performance that would be obtained by finetuning ALBEF on real image-text pairs from Visual Genome (VG).

Category	Row	Image	Text	$\langle I, T \rangle$	RefCOCO+		Flickr30k	Δ_{avg}
					Test A	Test B		
ALBEF	1	-	-	–	69.35	53.77	79.38	-
ALBEF + VG	2	VG	VG	1,649,546	71.41	54.06	79.90	+0.96
Concatenation	3	Syn-C	VG	1,649,546	67.57	53.14	76.99	-1.60
Text2Text	4	Syn-V	VG	1,649,546	67.41	52.14	77.80	-1.72
	5	Syn-V	LLM_C	530,233	70.28	52.08	78.97	-0.39
Image2Text	6	Syn-B	VG	1,649,546	56.88	48.48	73.93	-7.74
	7	Syn-B	BLIP-2_C	267,199	68.15	51.50	78.30	-1.52
	8	Syn-L	VG	1,649,546	65.35	50.28	76.85	-3.34
	9	Syn-L	LLaVA_P	384,455	70.22	52.30	78.34	-0.55
	10	Syn-L	LLaVA_C	716,198	69.94	53.26	78.83	-0.16
	11	Syn-L	LLaVA_L	680,093	69.84	**53.61**	79.44	+0.13
	12	Syn-L	LLaVA_S	1,031,521	**70.31**	52.55	**80.73**	**+0.36**

3.1 Generating Synthetic Image-Text Pairs

To generate image-text-box triplets for visual grounding, we first explore synthesizing effective image-text pairs for visual grounding. As shown in Fig. 1, we investigate three alternatives for conditioning a text-to-image generation model Ψ_g [61]. (1) *Concatenation*: Merging all captions of a real image I as a prompt for Ψ_g. (2) *Text2Text*: Using an LLM Ψ_t [12] to create a cohesive prompt given a set of text descriptions. (3) *Image2Text*: Employing an image captioning model Ψ_c to generate new captions for real images I^R as prompts for Ψ_g.

Table 1 presents a comparisons for each of these strategies under our base ALBEF model when tuned with only image-text objectives $\mathcal{L}_{\text{vl}}$. Although image-text matching objectives are not designed specifically for visual grounding, well-aligned region phrases from Visual Genome (VG) can improve visual grounding performance by 0.96% on average (Row 2). The *Concatenation* strategy (Syn-C) degrades the average performance by 1.60%, indicating that the text-to-image generation model Ψ_g is ineffective with long yet potentially redundant prompts. For *Text2Text*, the ALBEF model was trained with generated synthetic images from summaries generated with an LLM (Syn-V). For the text, we try two strategies, 1) Pairing the generated images with captions from the VG dataset, or using the LLM output summaries split with commas (LLM_C). Rows 4 and 5 show that this approach is also ineffective. For the *Image2Text* strategy, we experiment with two distinct styles of image captioning models: BLIP-2 [40] that yields condensed phrases, and LLaVA [44] that produces detailed paragraphs. Image generation with prompts obtained with BLIP-2 (Syn-B) and LLaVA (Syn-L) lead to images with insufficient alignment with VG captions to improve performance (Rows 6 and 8). Notably, an opposite influence is observed when Syn-B and Syn-L are paired with phrases extracted from their own prompts (Rows 7 and 9). BLIP-2 captions, usually short and object-centric (*e.g.,* "a dog, a cat"), are split into phrases by commas, showing improved performance over Syn-B and VG captions, possibly due to better cross-modal alignment, but still below the baseline. We find that LLaVA-synthesized images (Syn-L) paired with phrases extracted from LLaVA captions can enhance grounding performance (Table 1, Rows 11 and 12). **This indicates that detailed prompts suit the text-to-image synthesis model better.** We compare four ways to partition the LLaVA captions into phrases: LLaVA_P and LLaVA_C, segmented by periods and commas, respectively, LLaVA_L for longer LLM ex-

Table 2: Efficacy of synthetic image-text-boxes generated with either GLIP [42] or GLIGEN [43]. For reference, we also include the reported performance of finetuning ALBEF [38] with an AMC loss [87] on real image-text-box triplets from Visual Genome (VG).

Model	Row	Image	Text	Box	$\langle I,T,B\rangle$	RefCOCO+		Flickr30k	Δ_{avg}
						Test A	Test B		
ALBEF	1	-	-	-	–	69.35	53.77	79.38	-
AMC	2	VG	VG	VG	1,649,546	78.89	61.16	86.46	+8.00
ALBEF	3	Syn-A	VG	VG	1,649,546	68.79	56.88	84.57	+2.58
+	4	Syn-R	VG	VG	1,649,546	68.25	55.78	84.59	+2.04
GLIGEN	5	Syn-R	VG_R	VG_R	725,974	71.66	56.15	84.84	+3.38
	6	Syn-T	VG_T	VG_T	725,974	71.80	56.68	84.73	+3.57
	7	Syn-I	VG_I	VG_I	652,657	73.05	58.38	84.39	+4.44
ALBEF	8	Syn-L	$LLaVA_L$	GLIP	659,927	72.39	55.94	86.53	+4.12
+	9	Syn-L	$LLaVA_{S,L}$	GLIP	1,658,333	72.25	**57.05**	86.71	+4.50
GLIP	10	Syn-L	$LLaVA_S$	GLIP	998,406	**73.70**	56.35	**86.89**	**+4.81**

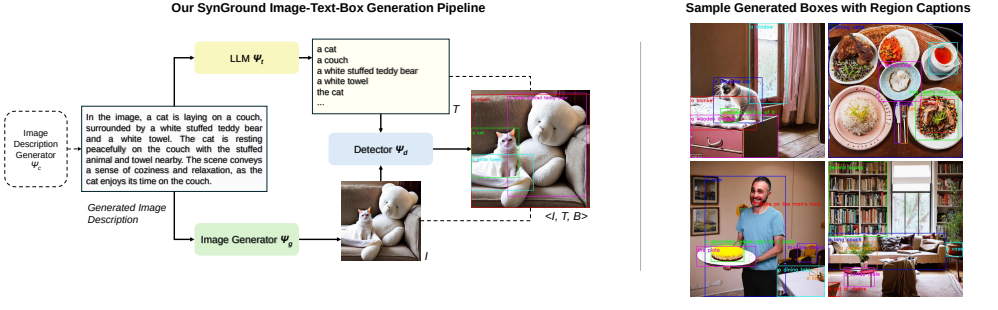

Figure 2: Left: An overview of our SynGround image-text-box synthesis pipeline. Right: Some sample generated image-text-box triplets. We use an image description generator Ψ_c to output a description that serves as a prompt to an image generator Ψ_g to obtain a synthetic image I. This description is also used to obtain text phrases T by prompting an LLM Ψ_t. Finally, the synthetic text and image are fed into an object detector Ψ_d for synthetic boxes B.

tracted phrases and $LLaVA_S$ for shorter phrases. **Our experiments demonstrate that the *Image2Text* strategy, particularly with LLaVA captioning and LLM phrase extraction, yields the most effective synthetic image-text pairs for visual grounding.** *More details in supplementary materials.*

3.2 Generating Synthetic Image-Text-Box Triplets

This section discusses two pipelines for image-text-box synthesis. The first pipeline builds on the success of *Image2Text* (Sec. 3.1) and additionally uses an open vocabulary object detector Ψ_d[42] to generate region annotations for each synthetic text phrase (See Fig. 2). The detector also serves as a filtering system for image-text pairs whose confidence scores are above a threshold. Table 2 compares pairing synthetic images with shorter phrases ($LLaVA_S$), longer phrases ($LLaVA_L$), and both ($LLaVA_{S,L}$). **The shorter phrases outperform others (Row 10)**, leading to an average performance gain of 4.81%. However, combining shorter and longer phrases ($LLaVA_{S,L}$) –despite increasing the amount of data–

Table 3: Training on synthetic and/or real data. We compare visual grounding improvements for the base model (Row 1), using the full amount of real data from VG (Row 2), a percentage of the real data from VG that could be plausibly annotated by 33,000 human workers in the same time that it takes to generate SynGround images on a single GPU (Row 3), synthetic data (Row 4), and both (Row 5).

Method	**Data**	**#Images**	$\langle I, T, B \rangle$	**RefCOCO+**		**Flickr30k**	Δ_{avg}
				Test A	Test B		
ALBEF [33]	Off-the-Shelf	–	–	69.35	53.77	79.38	-
AMC [57]	Real	94,893	1,649,546	78.89	61.16	86.46	+8.00
AMC′	Real	76,829	183,282	76.96	59.07	85.01	+6.18
SynGround$_S$	Synthetic	94,893	998,406	73.70	56.35	86.89	+4.81
SynGround	Real & Synthetic	189,786	2,627,952	**79.06**	**63.67**	**87.26**	**+9.16**

does not further improve performance, suggesting redundancy in the information conveyed by phrases with different lengths.

Alternatively, we investigate using the layout-conditioned generator GLIGEN [43], synthesizing images conditioned on the text and corresponding bounding boxes. Directly inputting all real VG texts and boxes (Row 3) results in a modest increase of 2.58% compared to the baseline (Row 1). **We observe the ineffectiveness of using regions with multiple textual descriptions, as this tends to generate unrealistic or implausible content.** To address this, we explore 3 strategies: Random selection of text-box inputs (VG$_R$), reduction based on average CLIP [58] text dissimilarity (VG$_T$), and selecting a maximum number of boxes with an IoU below 0.5 (VG$_I$). Random selection keeps at most 10 boxes per image, resulting in a reduction of about 50% of the data. Random text-box synthesized images Syn-R (Row 5) outperform the all-text-box conditioned variant (Syn-A, Row 3). Also, pairing Syn-R with all text-box data from Real VG (Row 4) does not match the effectiveness of either Syn-A with all text-boxes or Syn-R with selected text-boxes, underscoring the importance of image-text-box alignment. Sorting by CLIP text dissimilarity to select at most the top 10 inputs (Syn-T, VG$_T$) marginally improves random selection. **Yet, the most significant improvement stems from selecting as many boxes as possible with an IoU below 0.5.** The images (Syn-I) generated with this strategy match the best practice in the GLIP-based pipeline (Row 10).

Our results show the potential of using a layout-conditioned generative model for image-text-box synthesis. However, either generating non-overlapping and natural layouts or generating text for visually coherent layouts poses a substantial challenge, limiting the advancement to synthesis without real image-text-box data. Even with layout generation models [30, 34], strong constraints of natural composition and non-overlapping bounding boxes detract from their efficiency and effectiveness compared to the object detector approach.

We use our findings to define **SynGround**, a processing pipeline for generating synthetic image-text-boxes for visual grounding (Table 2, Row 10). Fig. 2 shows an overview of the SynGround pipeline along with representative examples of our generated image-text-boxes, including images with specific and recognizable entities (the first image shows “a Siamese cat"), complex scenarios with composite subjects (the second image shows “rice, beans and meat"). The third image shows a synthetic person with unrealistic features, observed in several generated results. This contrasts with improvements on RefCOCO+ Test A (a person-only subset), suggesting that realistic object details are not crucial for visual grounding. The fourth image showcases creative objects with unusual attributes such as a pink coffee table, which showcases diversity in our generated data. *More qualitative examples are in supp.*

Table 4: Factors causing the performance gap with the real data. We investigate how each model caused the ineffectiveness compared to the real data. 'R" for real and "S" for synthetic. I: Off-the-shelf base model. II: Learning from real data. III-V: Sequentially replacing real boxes, text, and images with synthetic variants.

	I	T	B	$\langle I,T,B \rangle$	RefCOCO+		Flickr30k	Δ_{avg}
					Test A	Test B		
I	-	-	-	–	69.35	53.77	79.38	-
II	R	R	R	1.65M	78.89	61.16	86.46	+8.00
III	R	R	S	1.60M	76.88	59.79	86.76	+6.98
IV	R	S	S	1.00M	73.11	57.35	87.49	+5.15
V	S	S	S	0.99M	73.70	56.35	86.89	+4.81

3.3 Using Real and/or Synthetic Data

SynGround can augment training with real data. Table 3 presents comparisons between training exclusively on real data from the Visual Genome (VG) dataset, synthetic data from SynGround, and a combination of both. The baseline performance (Row 1) is significantly enhanced by incorporating synthetic data, yielding an average improvement of 4.81% (Row 3). While it falls short of the gains achieved through training on real data (Row 2), SynGround offers an average improvement of 9.16% when combined with real data (Row 5), outperforming the state-of-the-art (Row 2) [87] on RefCOCO+ [90] Test A and B, and Flickr30k [57] benchmarks using Pointing Game accuracy. More importantly, the SynGround generation takes 501 GPU hours on a single NVIDIA A40, which is around $1/9$ of VG's data curation time from 33,000 unique workers [36]. The performance, obtained by training on **a percentage of the VG dataset that could plausibly be collected within an equivalent time span** using 33,000 human annotators as reported in the original study (Row 3), **is on par with SynGround$_S$.** *Computation details and comparisons are in the Supp.*

4 Discussion and Analysis

In this section we analyze the effectiveness of SynGround and what are the contributing factors to its performance with respect to real datafurther validation of SynGround under BLIP, and performance at various data scales.

Real-Synthetic Performance Gap Factors. Table 4 analyzes the factors contributing to the performance gap between synthetic and real data. Experiment I is the off-the-shelf ALBEF performance, serving as a baseline. Experiment II provides the results from training on real VG image-text-boxes, leading to an average improvement of 8%. Experiment III retains real images and texts from VG, but employs GLIP-generated boxes. The 1.02% decrease in performance compared to Experiment II suggests that the synthetic boxes, while effective, may lack the precision of manual-annotated equivalents. Experiment IV further replaces real VG captions with synthetic captions from SynGround (*i.e.,* LLaVA$_S$), resulting in an additional average reduction of 1.83%. This decline could stem from a reduction in the number of captions ($\sim$600K fewer) or discrepancies in image-text alignment, coverage, and diversity compared to manually curated captions (*details in the supplementary material.*). Interestingly, the performance on Flickr30k is enhanced by 1.03% over real data (II), showing a potential distribution shift from synthetic captions. In Experiment V, the setting consists en-

Table 5: Training on both synthetic and real data. We compare grounding improvements for BLIP (I) by using the real data (II, IV) and synthetic data (III, V) w/o the AMC box-supervised loss.

	Box Data	$\langle I, T, B \rangle$		RefCOCO+		Flickr30k	Δ_{avg}
				Test A	Test B		
I	×	–	–	58.56	38.00	64.54	–
II	×	R	1.65 M	68.86	52.85	64.08	+8.23
III	×	S	0.99 M	63.45	44.39	68.21	+4.98
IV	✓	R	1.65 M	78.47	61.96	85.35	+21.56
V	✓	S	0.99 M	71.78	54.82	85.83	+17.11

Table 6: Performance gap between real and synthetic data analyses with BLIP. "R" for real and "S" for synthetic. I: Off-the-shelf. II: Trained on real data. III-V: Sequentially replacing real boxes, text, and images with synthetic variants.

	I	T	B	$\langle I, T, B \rangle$	RefCOCO+		Δ_{avg}
					Test A	Test B	
I	–	–	–	–	58.56	38.00	–
II	R	R	R	1.65 M	78.47	61.96	+21.56
III	R	R	S	1.59 M	75.72	58.50	+19.74
IV	R	S	S	1.00 M	72.44	55.94	+18.00
V	S	S	S	0.99 M	71.78	54.82	+17.11

tirely of synthetic image-text-box data, eliminating real images from the dataset. Compared to Experiment IV, it modestly drops another 0.34%. This minor decrement, relative to the changes observed with synthetic texts and boxes, indicates that **synthetic images maintain a level of effectiveness for visual grounding tasks comparable to their real counterparts.**

Efficacy and Generalization on Other VLMs. This section experiments with an additional off-the-shelf VLM, BLIP [39], to further examine the effectiveness of our synthetic data and verify the generalizability of our findings from the default base model ALBEF. *Refer to the supplementary material for the base model selection and implementation details.* As shown in Table 5, our generated synthetic image-text and image-text-boxes significantly enhance its visual grounding performance (III, V), matching closely to the improvement from training on real data (II, IV). Additionally, we investigate the factors contributing to the degradation in Table 6. Similar to findings from ALBEF experiments in Sec. 4, most drop comes from the box and text synthesis. *In supplementary materials*, **consistent findings are observed by replacing ALBEF with BLIP for experiments presented in previous sections**. Moreover, our synthetic data also **helps the other type of grounding models [81] that predicts boxes** evaluated by Acc@0.5.

Effect of Data Scale on Visual Grounding. This section explores the scalability of synthetic data. We start the image-text-box synthesis at the scale of 250k and then extend it to 1M (SynGround). We sample 3 times from the 1M SynGround data and experiment with each scale to measure variance. Fig. 3 illustrates the average pointing game accuracy improvement across RefCOCO+ [90] and Flickr30k [57]. We plot the mean improvement at each scale with lines and their standard deviations with error bars. **The observed upward trend indicates a promising scaling-up ability to use synthetic data with SynGround.**

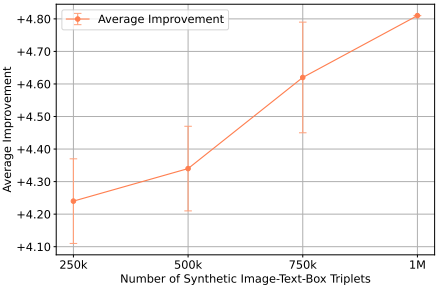

Figure 3: Pointing game accuracy improvement on RefCOCO+ and Flickr30k at various scales. The line denotes the mean improvement across 3 sampled subsets at each scale, and the error bars are corresponding standard deviations.

Table 7: Comparisons of our synthetic data with web-crawled data. The first row is the off-the-shelf base model performance while the second is after tuning on real data. The third row ("CC") tunes on a subset of CC [73] image-text pairs with generated synthetic boxes, while "CC_{Phrase}" processes the text through LLM phrase extraction. $SynGround_S^H$ and $SynGround_S$ refer to tuning on our synthetic data, relying less or more on the real data during synthesis.

Method	Data	$\langle I,T,B \rangle$	RefCOCO+		Flickr30k	Δ_{avg}
			Test A	Test B		
ALBEF [38]	-	–	69.35	53.77	79.38	-
AMC [37]	Real	1,649,546	78.89	61.16	86.46	+8.00
CC	Web	1,000,000	69.05	54.96	83.94	+1.82
CC_{Phrase}	Web	1,000,000	70.35	55.31	85.43	+2.86
$SynGround_S^H$	Synthetic	719,254	71.27	56.82	86.78	+4.12
$SynGround_S$	Synthetic	998,406	**73.70**	**56.35**	**86.89**	**+4.81**

Synthetic Data vs. Web-Crawled Data To showcase the challenge and necessity of generating effective synthetic data tailored for visual grounding, Table 7 compares our synthetic data and web-crawled data. The first and second rows are the off-the-shelf and tuning on real VG data, respectively. For fair comparisons, we randomly sample 1M web-crawled data from Conceptual Captions (CC) [73], approximately matching the scale of our synthetic data. As CC data only encompasses images and texts, we add synthetic boxes using an open-vocabulary detector [42], as the same in our method. Tuning the base model on it achieves (row 3) a 1.82% average performance gain. Additionally, experiments in Table 1 and other work [28] find that visual grounding ability can be enhanced more significantly with object-centric short phrases rather than generic image descriptions. Considering that CC text might describe entire scenarios, we further apply our LLM phrase extraction (row 4) and generate synthetic boxes for the synthetic text phrases, leading to a greater average improvement of 2.86%. However, to our best effort, we can not make the web-crawled data reach a similar enhancement with our synthetic data ($SynGround_S^H$, $SynGround_S$). Our experimental results indicate that it is non-trivial to curate or synthesize image-text-boxes for visual grounding. The image and text favored by visual grounding seem to feature specific properties, such as images with multiple objects and text for region descriptions.

5 Conclusion

This paper investigates various strategies and conducts extensive analyses for generating synthetic training data to improve the visual grounding ability of vision-and-language models. By leveraging exhaustive image descriptions for image synthesis, utilizing an LLM for phrase extraction, and adopting an open-vocabulary object detector for box synthesis, we propose SynGround– an effective framework to generate training data for improving visual grounding. SynGround can augment real data to yield further performance gains, and surpasses the efficacy of web-crawled data in visual grounding. Furthermore, SynGround is scalable and capable of generating theoretically infinite data using LLMs for image description generation.

Acknowledgements. This work was funded by NSF Award #2201710 and funding from the Ken Kennedy Institute at Rice University.

References

[1] Hassan Abu Alhaija, Siva Karthik Mustikovela, Lars Mescheder, Andreas Geiger, and Carsten Rother. Augmented reality meets computer vision: Efficient data generation for urban driving scenes. *International Journal of Computer Vision*, 126:961–972, 2018.

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

[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.

[4] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*, 2023.

[5] Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2616–2627, 2023.

[6] Paola Cascante-Bonilla, Hui Wu, Letao Wang, Rogerio S Feris, and Vicente Ordonez. Simvqa: Exploring simulated environments for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5056–5066, 2022.

[7] Paola Cascante-Bonilla, Khaled Shehada, James Seale Smith, Sivan Doveh, Donghyun Kim, Rameswar Panda, Gul Varol, Aude Oliva, Vicente Ordonez, Rogerio Feris, et al. Going beyond nouns with vision & language models using synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20155–20165, 2023.

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

[9] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

[10] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1841–1850, 2019.

[11] Zhihong Chen, Ruifei Zhang, Yibing Song, Xiang Wan, and Guanbin Li. Advancing visual grounding with scene knowledge: Benchmark and method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15039–15049, 2023.

[12] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

[13] Yabo Dan, Yong Zhao, Xiang Li, Shaobo Li, Ming Hu, and Jianjun Hu. Generative adversarial networks (gan) based efficient sampling of chemical composition space for inverse design of inorganic materials. *npj Computational Materials*, 6(1):84, 2020.

[14] 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.

[15] Celso M de Melo, Antonio Torralba, Leonidas Guibas, James DiCarlo, Rama Chellappa, and Jessica Hodgins. Next-generation deep learning based on simulators and synthetic data. *Trends in cognitive sciences*, 2022.

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

[17] 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.

[18] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. Redcaps: Web-curated image-text data created by the people, for the people. *arXiv preprint arXiv:2111.11431*, 2021.

[19] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017.

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

[21] Lijie Fan, Kaifeng Chen, Dilip Krishnan, Dina Katabi, Phillip Isola, and Yonglong Tian. Scaling laws of synthetic images for model training... for now. *arXiv preprint arXiv:2312.04567*, 2023.

[22] Chuang Gan, Jeremy Schwartz, Seth Alter, Damian Mrowca, Martin Schrimpf, James Traer, Julian De Freitas, Jonas Kubilius, Abhishek Bhandwaldar, Nick Haber, et al. Threedworld: A platform for interactive multi-modal physical simulation. *arXiv preprint arXiv:2007.04954*, 2020.

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

[24] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3749–3761, 2022.

[25] 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.

[26] 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.

[27] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? *arXiv preprint arXiv:2210.07574*, 2022.

[28] Ruozhen He, Paola Cascante-Bonilla, Ziyan Yang, Alexander C Berg, and Vicente Ordonez. Improved visual grounding through self-consistent explanations. *arXiv preprint arXiv:2312.04554*, 2023.

[29] Xuanli He, Islam Nassar, Jamie Kiros, Gholamreza Haffari, and Mohammad Norouzi. Generate, annotate, and learn: Nlp with synthetic text. *Transactions of the Association for Computational Linguistics*, 10:826–842, 2022.

[30] Naoto Inoue, Kotaro Kikuchi, Edgar Simo-Serra, Mayu Otani, and Kota Yamaguchi. Layoutdm: Discrete diffusion model for controllable layout generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10167–10176, 2023.

[31] 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.

[32] 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 Conference on Computer Vision and Pattern Recognition*, pages 15513–15523, 2022.

[33] 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.

[34] Kotaro Kikuchi, Edgar Simo-Serra, Mayu Otani, and Kota Yamaguchi. Constrained graphic layout generation via latent optimization. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 88–96, 2021.

[35] Donghyun Kim, Kaihong Wang, Kate Saenko, Margrit Betke, and Stan Sclaroff. A unified framework for domain adaptive pose estimation. In *European Conference on Computer Vision*, pages 603–620. Springer, 2022.

[36] 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.

[37] Varun Kumar, Ashutosh Choudhary, and Eunah Cho. Data augmentation using pre-trained transformer models. *arXiv preprint arXiv:2003.02245*, 2020.

[38] 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.

[39] 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.

[40] 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.

[41] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[42] 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.

[43] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023.

[44] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

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

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

[47] 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.

[48] Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. Generating training data with language models: Towards zero-shot language understanding. *Advances in Neural Information Processing Systems*, 35:462–477, 2022.

[49] Masato Mimura, Sei Ueno, Hirofumi Inaguma, Shinsuke Sakai, and Tatsuya Kawahara. Leveraging sequence-to-sequence speech synthesis for enhancing acoustic-to-word speech recognition. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 477–484. IEEE, 2018.

[50] Samarth Mishra, Rameswar Panda, Cheng Perng Phoo, Chun-Fu Richard Chen, Leonid Karlinsky, Kate Saenko, Venkatesh Saligrama, and Rogerio S Feris. Task2sim: Towards effective pre-training and transfer from synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9194–9204, 2022.

[51] Arthur Moreau, Nathan Piasco, Dzmitry Tsishkou, Bogdan Stanciulescu, and Arnaud de La Fortelle. Lens: Localization enhanced by nerf synthesis. In *Conference on Robot Learning*, pages 1347–1356. PMLR, 2022.

[52] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[53] Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3170–3180, 2023.

[54] Xingchao Peng, Baochen Sun, Karim Ali, and Kate Saenko. Learning deep object detectors from 3d models. In *Proceedings of the IEEE international conference on computer vision*, pages 1278–1286, 2015.

[55] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 102–118. Springer, 2016.

[61] 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.

[62] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016.

[63] Andrew Rosenberg, Yu Zhang, Bhuvana Ramabhadran, Ye Jia, Pedro Moreno, Yonghui Wu, and Zelin Wu. Speech recognition with augmented synthesized speech. In *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*, pages 996–1002. IEEE, 2019.

[64] Nick Rossenbach, Albert Zeyer, Ralf Schlüter, and Hermann Ney. Generating synthetic audio data for attention-based speech recognition systems. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7069–7073. IEEE, 2020.

[65] Artem Rozantsev, Vincent Lepetit, and Pascal Fua. On rendering synthetic images for training an object detector. *Computer Vision and Image Understanding*, 137:24–37, 2015.

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

[67] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[68] Mert Bulent Sariyildiz, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In *CVPR 2023–IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[69] 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.

[70] 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.

[71] 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.

[72] 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.

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

[74] Yonglong Tian, Lijie Fan, Kaifeng Chen, Dina Katabi, Dilip Krishnan, and Phillip Isola. Learning vision from models rivals learning vision from data. *arXiv preprint arXiv:2312.17742*, 2023.

[75] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *Advances in Neural Information Processing Systems*, 36, 2024.

[76] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[77] Allan Tucker, Zhenchen Wang, Ylenia Rotalinti, and Puja Myles. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *NPJ digital medicine*, 3(1):1–13, 2020.

[78] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 109–117, 2017.

[79] Gül Varol, Ivan Laptev, Cordelia Schmid, and Andrew Zisserman. Synthetic humans for action recognition from unseen viewpoints. *International Journal of Computer Vision*, 129(7):2264–2287, 2021.

[80] 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.

[81] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International conference on machine learning*, pages 23318–23340. PMLR, 2022.

[82] Weiyun Wang, Min Shi, Qingyun Li, Wenhai Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao Li, Xizhou Zhu, Zhiguo Cao, et al. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. *arXiv preprint arXiv:2308.01907*, 2023.

[83] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. *arXiv preprint arXiv:2311.06242*, 2023.

[84] 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.

[85] Li Yang, Yan Xu, Chunfeng Yuan, Wei Liu, Bing Li, and Weiming Hu. Improving visual grounding with visual-linguistic verification and iterative reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9499–9508, 2022.

[86] Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. Generative data augmentation for commonsense reasoning. *arXiv preprint arXiv:2004.11546*, 2020.

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

[88] Lin Yen-Chen, Pete Florence, Jonathan T Barron, Tsung-Yi Lin, Alberto Rodriguez, and Phillip Isola. Nerf-supervision: Learning dense object descriptors from neural radiance fields. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 6496–6503. IEEE, 2022.

[89] 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. *arXiv preprint arXiv:2310.07704*, 2023.

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

[91] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 519–535. Springer, 2020.