Learning from Synthetic Data for Visual Grounding

Ruozhen He[1] Ziyan Yang[1] Paola Cascante-Bonilla[2] Alexander C. Berg [3] Vicente Ordonez[1]
[1]Rice University [2]University of Maryland [3]University of California, Irvine

Abstract

This paper extensively investigates the effectiveness of synthetic training data in improving the capabilities of vision-and-language models for visual grounding. We explore various strategies to best generate image-text pairs and image-text-box triplets using a series of pretrained models. Through comparative analyses with synthetic, real, and web-crawled data, we identify factors that contribute to performance differences, and propose SynGround*, an effective pipeline for generating useful synthetic data for visual grounding. We show that data generated with SynGround improves the pointing game accuracy of pretrained ALBEF and BLIP models by* 4.81% *and* 17.11% *absolute percentage points, respectively, across the RefCOCO+ and the Flickr30k benchmarks.*

1. Introduction

Vision-and-language models pretrained on large-scale web-sourced image and text pairs have become exceedingly accurate across various tasks [3, 19, 26, 28, 29, 32, 34, 41, 46]. Our work focuses on the task of visual grounding, which consists of mapping arbitrary input text to image regions. Recent methods finetune pre-trained vision-and-language models with a large but more modest number of images annotated with bounding boxes or other region annotations; alternatively, these methods leverage pretrained object detectors that have been trained on such annotated data [6, 7, 13, 16, 20, 21, 29, 66, 68]. However, region annotations in the form of bounding boxes or segments can not be easily obtained from the web, and require more cognitive effort to annotate manually than providing a single label. For instance, curating the widely used Visual Genome [24] dataset involved contributions from 33,000 unique workers over 6 months, following 15 months of experimentation and refinement of the data representation. The massive manual demand restricts the scale of such region-annotated data compared to the image-text datasets at the billion scale [55]. Recent work has championed the use of synthetic data – *learning from models* – even for tasks that require only image-text pair supervision [58]. Our work takes this paradigm one step further by investigating whether synthetic data obtained from models is ready to make significant improvements for the visual grounding task, where we need to obtain high-quality image-text-region triplets.

In this paper, we take advantage of recent advancements in text-to-image generation [39, 48, 54], large language models [8, 59] and models for other vision-and-language tasks [27, 29, 31] to design an effective pipeline, *SynGround*, to supervise vision-and-language models for visual grounding. Our key findings and contributions are summarized as follows: (1) We 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. (2) Our results show that using our generated synthetic data outperforms using web-crawled data (Sec. 5.3). Additionally, our synthetic data can effectively augment real data (Sec. 3.1) and shows an upward trend in terms of scalability (Sec. 5.2).

2. Related Work

Visual Grounding. Visual grounding associates textual descriptions with relevant regions within images. Previous methods typically rely on expensive image-text-box triplets [6, 10–13, 15, 16, 21, 33, 63, 68]. Although some studies collect more data [65] or generate annotations for existing image-text datasets [44, 64, 70], we posit that our contribution is orthogonal as we aim to investigate the feasibility and limitations of generating and using synthetic data.

Learning from Synthetic Data. The use of synthetic data has been widely explored across various computer vision tasks, including image classification [14, 37, 43], semantic segmentation [5, 47, 49], object detection [42, 52], human pose estimation [22, 61], action recognition [62], and many other domains [1, 9, 18, 25, 35, 36, 38, 50, 51, 60, 67, 69]. 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 a generic vision-and-language model. This model 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 the base model on image-text pairs $\langle I, T\rangle$ and image-text-box triplets $\langle I, T, B\rangle$. Then, 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. 3.1 shows evaluation of SynGround when combined with Real Data.

Preliminaries and Setup

Image-Text Matching. We adopt ALBEF [26] as the main base model which incorporates image-text objectives including 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 ($\mathcal{L}_{\text{mlm}}$). 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 region-level box supervision on top of the ALBEF model [68]. This objective uses gradient-based explanation heatmaps G through Grad-CAM [56], and maximizes the consistency between this map and region annotations. This objective considers two terms. The first term $\mathcal{L}_{\max}$ encourages the maximum value of G inside a target box B to surpass the maximum value outside by a margin δ_1.

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

where $B_{i,j}$ is 1 when pixel location i, j is inside the box, and zero otherwise. The second term $\mathcal{L}_{\text{mean}}$ encourages the mean value of heatmap G inside the box to be larger than the mean value outside by a margin δ_2.

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

The full $\mathcal{L}_{\text{amc}}$ objective is $\mathcal{L}_{\text{amc}} = \mathcal{L}_{\max} + \lambda \cdot \mathcal{L}_{\text{mean}}$, 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 works for Visual Grounding methods that predict 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, 10, 13, 16, 17, 26, 33, 68]. We conduct evaluation across multiple benchmarks, including RefCOCO+ [71] and Flickr30k [45].

Image-Text-Box Generation Pipeline. Fig. 1 shows an overview of our proposed *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.

3.1. Using Real and/or Synthetic Data

SynGround can augment training with real data. Table 1 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) [68] on RefCOCO+ [71] Test A and B, and Flickr30k [45] 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 [24]. 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. Implementation Details

Image-Text-Box Synthesis. To favor reproducibility and accessibility, we adopted Stable Diffusion 2.1 [48] with guidance scale 10.0 as the text-to-image generator Ψ_g, an open-source LLM Vicuna-13B [8] as Ψ_t, and GLIP [29] as the object detector Ψ_d. We selected the box with top-1 confidence if it exceeds the default confidence threshold (0.7) in the official implementation. For image description generation Ψ_c, we experimented with BLIP-2 [27] and LLaVA 1.5 [31] for the *Image2Text* strategy. For the *Concept2Text* variant, we used Vicuna-13B [8] to generate image descriptions from a two-concept query with four randomly sampled in-context examples. The concept list contains nouns extracted from real VG captions.

Visual Grounding Tuning. The main base model ALBEF-14M [26] is the same as that adopted by the current SotA fully- [68] and weakly-supervised methods [17] for Pointing Game accuracy metric. ALBEF is pretrained on image-

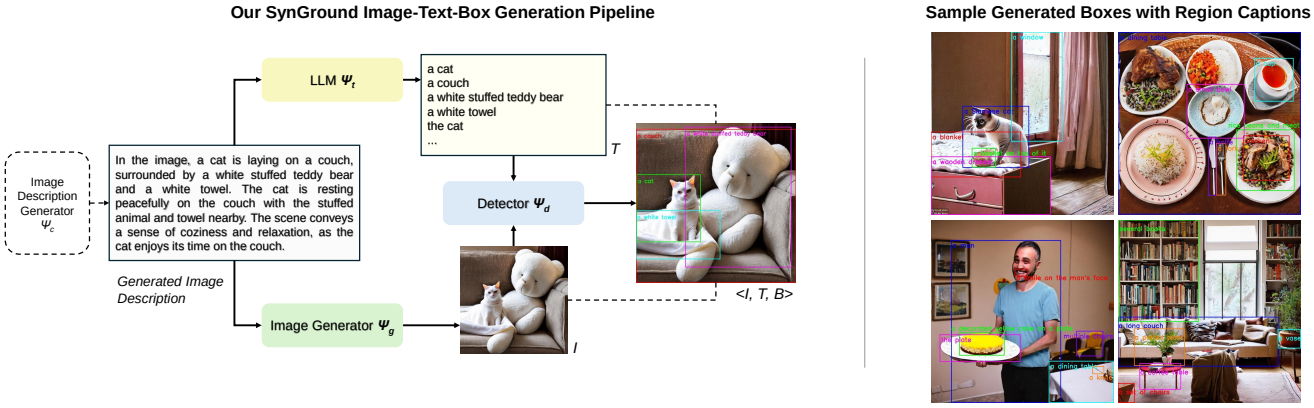

Figure 1. On the left, an overview of our SynGround image-text-box synthesis pipeline, and on the right some sample generated image-text-box triplets. We use an image description generator Ψ_c to output a description that serves as a prompt to an image generator Ψ_g to obtain synthetic image I. This 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 to obtain synthetic boxes B.

Table 1. 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 [26]	Off-the-Shelf	–	–	69.35	53.77	79.38	-
AMC [68]	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**

text pairs from CC [4], ImageNet-1k [53], MS-COCO [30], SBU Captions [40] and VG [24]. Tuning for visual grounding applies $\mathcal{L}_{\text{vl}}$ on image-text pairs and a combination of $\mathcal{L}_{\text{vl}}$ and $\mathcal{L}_{\text{amc}}$ on image-text-box triplets, adhering to the coefficient settings $\delta_1 = 0.5$, $\delta_2 = 0.1$, $\lambda_1 = 0.8$, and $\lambda_2 = 0.2$ as originally proposed in Yang *et al.* [68]. The training is conducted on a single node with 8 NVIDIA A40 GPUs. Input images are resized to 256×256 pixels and augmented with color jittering, horizontal flipping, and random grayscale conversion. All ALBEF-based experiments use an Adam optimizer [23] with a learning rate set to 1e-5 and a batch size of 512.

5. 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 data (Sec 5.1), performance at various data scales (Sec 5.2), and comparison against web-crawled data (Sec 5.3).

Table 2. 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.

Ex.	**Image**	**Text**	**Box**	$\langle I, T, B \rangle$	**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

5.1. Real-Synthetic Performance Gap Factors

Table 2 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

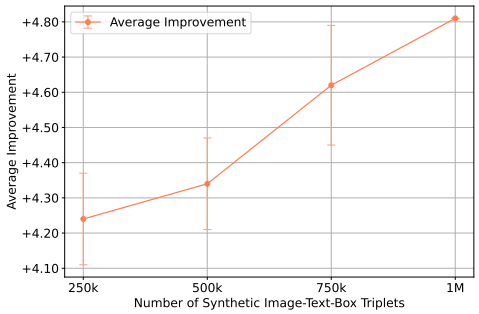

Figure 2. Pointing game accuracy improvement on RefCOCO+ and Flickr30k at various scales. The line denotes the mean improvement across 3 sampled subsets at each scale, and the error bars are corresponding standard deviations.

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 (∼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 entirely 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.

5.2. 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. 2 illustrates the average pointing game accuracy improvement across RefCOCO+ [71] and Flickr30k [45]. 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.

5.3. Synthetic Data vs. Web-Crawled Data

To showcase the challenge and necessity of generating effective synthetic data tailored for visual grounding, Table 3 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) [57], 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 [29], as the same in our method. Tuning the base model on it achieves (row 3) a 1.82% average performance gain. Additionally, He et al. [17] 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^H_S, 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.

Table 3. Comparisons of our synthetic data with web-crawled data. The first row is the off-the-shelf base model performance, and the second is the performance after tuning on real data. The third row ("CC") tunes on a subset of CC [57] image-text pairs with generated synthetic boxes, while "CC$_{Phrase}$" processes the text through LLM phrase extraction. SynGround^H_S and SynGround$_S$ refer to tuning on our synthetic data, relying less or more on the real data during synthesis, respectively.

Method	**Data**	$\langle I, T, B \rangle$	**RefCOCO+**		**Flickr30k**	Δ_{avg}
			Test A	Test B		
ALBEF [26]	-	–	69.35	53.77	79.38	-
AMC [68]	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^H_S	Synthetic	719,254	71.27	56.82	86.78	+4.12
SynGround$_S$	Synthetic	998,406	**73.70**	**56.35**	**86.89**	**+4.81**

6. Conclusion

We propose SynGround – an effective framework to generate synthetic 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.

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. 1

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

[3] 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. 1

[4] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021. 3

[5] 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. 1

[6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 1

[7] 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. 1

[8] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 1, 2

[9] 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. 1

[10] 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. 1, 2

[11] 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.

[12] 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.

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

[14] 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. 1

[15] 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. 1

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

[17] 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. 2, 4

[18] 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. 1

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

[20] 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. 1

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

[22] 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. 1

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[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. 1, 2, 3

[25] Varun Kumar, Ashutosh Choudhary, and Eunah Cho. Data augmentation using pre-trained transformer models. *arXiv preprint arXiv:2003.02245*, 2020. 1

[26] 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. 1, 2, 3, 4

[27] 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, 2

[28] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 1

[29] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 1, 2, 4

[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. 3

[31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1, 2

[32] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 1

[33] 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. 1, 2

[34] 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. 1

[35] 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. 1

[36] 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. 1

[37] 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. 1

[38] 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. 1

[39] 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. 1

[40] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011. 3

[41] 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. 1

[42] 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. 1

[43] 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. 1

[44] 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. 1

[45] 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

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

[47] 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. 1

[48] 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. 1, 2

[49] 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. 1

[50] 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. 1

[51] 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. 1

[52] 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. 1

[53] 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. 3

[54] 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. 1

[55] 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

[56] 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. 2

[57] 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. 4

[58] 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. 1

[59] 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. 1

[60] 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. 1

[61] 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. 1

[62] 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. 1

[63] 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. 1

[64] 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. 1

[65] 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. 1

[66] 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. 1

[67] 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. 1

[68] Ziyan Yang, Kushal Kafle, Franck Dernoncourt, and Vicente Ordonez. Improving visual grounding by encouraging consistent gradient-based explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19165–19174, 2023. 1, 2, 3, 4

[69] 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. 1

[70] 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. 1

[71] 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. 2, 4