Learning Compositional Language-based Object Detection with Diffusion-based Synthetic Data

Kwanyong Park
ETRI

Kuniaki Saito
OMRON SINIC X Corporation

Donghyun Kim
Korea University

Abstract

*Vision-language (VL) models often exhibit a limited understanding of complex expressions of visual objects (*e.g.*, attributes, shapes, and their relations), given complex and diverse language queries. While conventional methods try to enhance VL models through the use of hard negative synthetic text, their effectiveness remains restricted. In this paper, we introduce a structured synthetic data generation approach to improve the compositional understanding of VL models for language-based object detection. Specifically, our framework generates densely paired positive and negative triplets (image, text descriptions, bounding boxes) in both image and text domains. In addition, in order to train VL models effectively, we propose a new compositional contrastive learning formulation that discovers semantics and structures in complex descriptions from synthetic triplets. As a result, VL models trained with our synthetic data generation exhibit a significant performance boost in the Omnilabel benchmark by up to +5AP and the D^3 benchmark by +6.9AP upon existing baselines.*

1. Introduction

Recently, vision-language (VL) models have demonstrated significant advancements in visual recognition by learning from large-scale weakly supervised image-text pair datasets [10, 17]. While traditional recognition models [7, 13, 18, 21] are restricted to classifying or detecting predefined classes, image-text paired data allow models to easily generalize to new concepts and domains with language queries [6, 12].

Despite advancements, VL models [12, 17] continue to face challenges in understanding complex language queries and structured vision-language concepts, such as detailed object attributes, shapes, textures, and their relationships [5, 23, 25]. Related to this, novel object detection benchmarks like OmniLabel [20] and D^3 [24] have been introduced, tasking models with interpreting a broad range of complex object descriptions to accurately detect target objects. In such scenarios, VL models often overlook the complex and free-form textual descriptions provided, leading to wrong detection results. To address this issue, prior work [11] has explored augmenting the text domain [5, 23, 25] by generating synthetic negative texts through swapping nouns or generating new image captions. Nonetheless, we observe that merely enriching the text domain is insufficient for models to learn dense relations between images and text.

To overcome these difficulties, we propose an innovative framework to automatically generate the synthetic triplets and utilize them to improve the language-based object detector for better compositional understanding. The framework consists of two steps: (1) Generating diverse and dense triplets of (image, text descriptions, bounding boxes). Instead of solely relying on difficult-to-obtain real-world data [16], we propose to generate dense triplets with the generative models (Sec. 2.1). We first use a large language model [1, 2] to collect diverse and dense variations of visual entities (e.g., attributes, relations) in the text domain, then translate these descriptions to the image domain with the text-to-image diffusion models [3]. As a last piece, we localize depicted visual entities as a bounding box. In this step, we decompose the hard grounding problem into multiple easy detection problems, and this simple yet effective change enables us to obtain an accurate bounding box. (2) Effective learning from densely generated triplets (Sec. 2.2). For an image of a specific visual entity, we first contrast the dense variation of descriptions and the detector is trained to detect the object only for the corresponding descriptions. Besides, we use structural information in the textural description to identify the subject entity and use it to suppress the predictions for the non-subject entities in the descriptions. Both contrastive learning method largely improves compositional understanding, resulting in significant performance gain in description-based object detection. Thanks to the generality of the proposed framework, we show that our method significantly improves the performance of the diverse prior detectors on the two challenging benchmarks, OmniLabel [20] and D^3 [24].

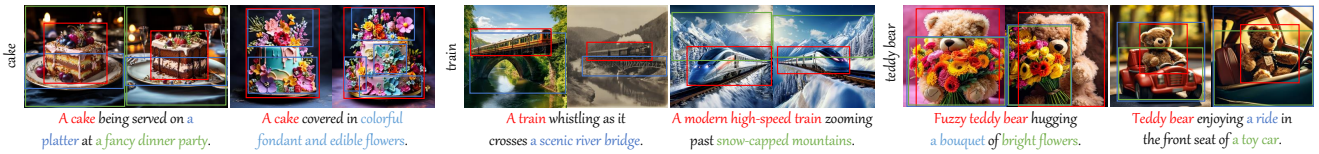

Figure 1. Qualitative examples of synthetic triplets for each visual entity. Best viewed in zoom.

2. Method

We aim to improve the compositional understanding capabilities of language-based object detectors. Instead of relying on difficult-to-obtain triplets (image, object descriptions, and bounding boxes), we harness the capabilities of foundational models by generating these triplets. Our approach involves two main steps: (1) dense synthetic triplet generation (Sec. 2.1) and (2) compositional contrastive learning with dense synthetic triplets (Sec. 2.2).

2.1. Synthetic Triplet Generation in Image and Text Domains

A traditional training data collection process for grounding data [13, 16] is to *collect images*, and manually *annotate object bounding boxes with their text descriptions*. However, it would be prohibitively expensive and does not guarantee to obtain hard negatives (*i.e.*, dense triplets), which is crucial to improve the compositionality of VL models [5, 11]. In order to obtain diverse and dense triplets, we adopt a reversed approach which first *generates text descriptions* and then *collects corresponding images* and *localize the depicted objects*.

Generating Diverse Object Descriptions. We initiate the process by generating diverse text descriptions for a wide variety of visual entities with large language models (LLM) [2]. For instance, we prompt an LLM with instructions such as, *"Please list $\{ND\}$ plausible visual object descriptions for $\{class\}$ that are around $\{NW\}$ words in length. Consider incorporating diverse visual attributes, actions, and spatial or semantic relations with other objects in each description."* This approach allows us to efficiently gather prior knowledge about specific visual entities (*i.e.* $\{class\}$), encompassing their likely attributes, natural co-occurrences with other objects, and the relationships between them.

The proposed LLM-based method for generating object descriptions is notable for its scalability and controllability. By adjusting parameters such as the pool size of visual entities (*i.e.*, entity density), the number of descriptions ($\{ND\}$) per entity (*i.e.*, description density), and the length of each description (*i.e.*, $\{NW\}$), we can easily manage the diversity and volume of the generated descriptions. We borrow the pool of visual entities from well-curated lists of everyday object categories from popular object detection datasets [14, 21].

Generating Densely Paired Images with Diffusion Models. Diffusion-based text-to-image generation models [8, 19, 22] have recently demonstrated their capability to produce high-fidelity, photo-realistic images. To acquire densely paired image-text data, we generate multiple images with the diffusion model [3] for each generated object description. This approach allows us to explicitly introduce diversity by specifying the objects in the descriptions. As a byproduct, this strategy provides pairs of object descriptions and images for training purposes.

Weak-to-Strong Pseudo Bounding Box Generation. Even if we have a collection of densely paired generated descriptions and images, accurate localization information of the depicted objects is crucial for training detectors on it. However, even recent pre-trained vision-language detectors [4, 12] often struggle to identify visual entities based on complex descriptions, due to their limited compositional understanding capabilities.

To this end, we delve into strategies for achieving precise object localization using weak detectors (in terms of compositional understanding), thereby facilitating the generation of rich supervision for training stronger detectors. We term this as a weak-to-strong labeling method. An overview of the process is depicted in Fig. 2-(a). We reformulate the complex phrase grounding problem into multiple tractable detection tasks with positive and short descriptions. For each pair of generated images and object descriptions, we first identify all noun phrases with an NLP parser [9]. Each noun phrase is treated as an independent description to detect the corresponding objects (*i.e.*, task decomposition). This ensures satisfactory precision and recall. Low-confidence predictions are filtered out based on a predetermined threshold. The remaining predictions are re-assigned to the original position within the description, which results in a strong compositional label for the following step. Fig. 1 illustrates the qualitative examples of synthetic triplets through the proposed generation pipeline.

2.2. Compositional Contrastive Learning for Language-based Object Detection

A straightforward approach to utilize the generated triplets (image, object descriptions, bounding boxes) is to use as additional grounding data: learning the alignment between noun phrases and detected object regions. However, our preliminary investigations reveal that models naively trained with these triplets show marginal improvements. We

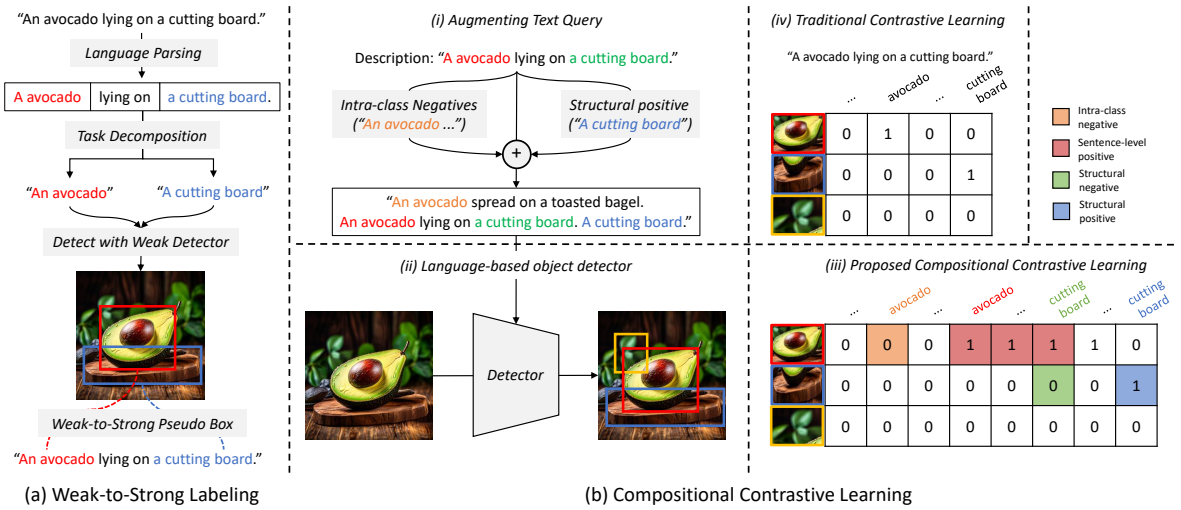

Figure 2. (a) Overview of proposed weak-to-strong labeling. (b) Illustration of our compositional contrastive learning. Intra-class negatives from other images of the same class and structural positives are augmented into the text query. We associate the sentence-level positive (*i.e.*, the entire description sentence) with the pseudo bounding box of the "an avocado" while differentiating the structure negative (*i.e.*, the noun phrase "a cutting board") from the pseudo bounding box of the "a cutting board". Best viewed in color.

identify two critical functionalities for compositional understanding: description-awareness and textural-structural-awareness.

Learning Description-awareness with Dense Contrastive Learning. Traditional language-based detectors often lack description awareness, indiscriminately detecting entities, regardless of the provided descriptions. To address this, we introduce supervisory signals that lead the model to pay attention to the given descriptions. Specifically, we select intra-class negative captions from the description pool that belong to the same object category as the image and augment the input query with the negatives (*e.g.*, "An avocado spread on a toasted bagel" in Fig. 2-(b)). Then the model is trained to disregard the visual entities for these negative captions. This approach demands that the model discerns between identical or similar noun phrases based solely on the context of entangled descriptions, significantly enhancing description-based detection accuracy. Notably, densely generated descriptions synergy well with this description-awareness training.

Learning Textural-Structural-Awareness. Existing language-based detectors often perform akin to a bags-of-words [25], indiscriminatively detecting all visual entities mentioned in the descriptions. To overcome this, we aim to distinguish between subject and non-subject entities within descriptions. We use textural relation [9] between noun phrases to identify subject and non-subject entities (i.e., visual entities within the descriptions). Then, the detector is instructed to ignore non-subject entities (*e.g.*, "lying on a cutting board" in Fig. 2-(b)) based on the description. We term this concept as a structural negative. For the subject noun entity, we ensure that the entire positive descriptions are positively aligned (*i.e.* sentence-level positive). In addition, to prevent the model from taking shortcuts that overlook later nouns, we introduce structural positives (*e.g.*, "A cutting board" in Fig. 2-(b)) by augmenting the model's textual input with the noun phrase of the non-subject entity. Then, the detector is trained to recognize the corresponding object for the structural positive query. Through this strategy, the model learns to differentiate identical noun phrases based on their structural role within the language query (subject vs. non-subject). This leads to significant improvements in performance, particularly for complex queries involving multiple visual entities.

3. Experiments

Training Details. By default, in synthetic data generation, we use the category pool from Object365 [21], ChatGPT3.5-Turbo [2] for description generation and Pixart [3] for image generation. We finetune pre-trained detectors [4, 11, 12] using a combination of our generated datasets and the Objects365 [21] object detection dataset.

Evaluation Benchmarks. We benchmark our proposed approach on the OmniLabel [20] and D^3 [24] datasets. For OmniLabel, the AP-c and AP-d quantify detection accuracy for standard plain object categories and for free-form textual descriptions, respectively. AP-d-S/M/L categorizes performance metrics according to the length of the descriptions (short, medium, and long). For more details, please refer to the original papers.

3.1. Main Results

We evaluate the impact of the proposed learning framework in Table 1. We first finetune two baseline models, GLIP [12] and FIBER [4], and observe significant enhancements in

Model	Backbone	OmniLabel [20] AP	AP-c	AP-d	AP-dP	AP-dS	AP-dM	AP-dL	D^3 [24] Full	Pres	Abs
RegionCLIP [26]	ResNet-50	2.7	2.7	2.6	3.2	3.6	2.7	2.3	-	-	-
Detic [27]	Swin-B	8.0	15.6	5.4	8.0	5.7	5.4	6.2	-	-	-
Grounding-DINO [15]	Swin-B	-	-	-	-	-	-	-	20.7	20.1	22.5
OFA-DOD [24]	Swin-B	-	-	-	-	-	-	-	21.6	23.7	15.4
GLIP-T [12]	Swin-T	19.3	23.6	16.4	25.8	29.4	14.8	8.2	19.1	18.3	21.5
w/ Ours	Swin-T	**24.3**	**23.9**	**24.7**	**34.4**	**39.3**	**21.6**	**16.4**	**26.0**	**25.6**	**27.1**
FIBER-B [4]	Swin-B	25.7	30.3	22.3	34.8	38.6	19.5	12.4	22.7	21.5	26.0
w/ Ours	Swin-B	**30.5**	**31.6**	**29.5**	**40.3**	**43.7**	**26.3**	**21.3**	**26.5**	**26.0**	**27.7**
Desco-GLIP [11]	Swin-T	23.8	**27.4**	21.0	30.3	33.7	19.0	13.7	24.2	22.9	27.8
w/ Ours	Swin-T	**26.5**	27.1	**25.9**	**35.6**	**38.1**	**23.2**	**18.7**	**29.3**	**29.1**	**30.1**
Desco-FIBER [11]	Swin-B	29.3	31.6	27.3	37.7	42.8	24.4	18.6	28.1	27.2	**30.5**
w/ Ours	Swin-B	**32.0**	**33.1**	**30.9**	**40.4**	**45.2**	**27.7**	**22.9**	**30.8**	**31.0**	30.4

Table 1. Performance comparison with state-of-the-art methods.

learning method	AP	AP-c	AP-d	AP-dp	AP-dS	AP-dM	AP-dL
FIBER-B	25.7	30.3	22.3	34.8	38.6	19.5	12.4
Gen-as-grounding	26.8	31.3	23.4	34.4	40.8	19.5	11.8
(+) Des.-aware	29.0	30.9	27.4	36.6	44.2	24.0	14.9
(+) Text.-struct.-aware	30.5	31.6	29.5	40.3	43.7	26.3	21.3

Table 2. Ablation on compositional contrastive learning.

language-based object detection performance across both datasets. This implies that the proposed learning framework is generic over different detection architectures and evaluation scenarios. Notably, the GLIP model's performance shows a substantial improvement, with an increase of +5.0AP and +6.9AP on the overall metrics for the OmniLabel and D^3 datasets, respectively. The enhancements are particularly pronounced for long queries (*i.e.*, AP-dL in OmniLabel), where the performance of the GLIP model doubles from 8.2 to 16.4.

We then explore the synergy between our proposals and the prior language augmentation-based method (*i.e.*, DesCo [11]). In this configuration, we apply their methods to enrich the language queries within the detection dataset [21] during training. As shown in the table, our proposal surpasses their models, DesCo-GLIP and DesCo-FIBER, by a considerable margin across both datasets. This shows that augmenting solely within the textual domain is insufficient. Our compositional contrastive learning on densely generated triplets offers distinct and substantial improvements.

3.2. Ablation Study and Analysis

Effective learning signals with synthetic data. We validate the impact of the proposed learning methods in Table 2. We start with naive finetuning, treating densely generated triplets similarly to conventional grounding data. (*i.e.*, Gen-as-grounding). The naive finetuning method only shows the marginal improvements. Upon this, we explore the impact of the proposed contrastive learning methods. By contrasting dense descriptions from the same visual entity (*i.e.*, Des.-aware), the model faithfully learns the description awareness, leading to the significant improvements of 4.0AP in the description-based performance. We then explore the text structural-based contrastive learning, enforcing the model to discriminate the same phrases according to their structural role in the description. This greatly improves description-based performance, especially the notable gain of 6.4AP for long queries. To sum up, all the proposed learning methods show their unique effect and the performance improvements of the final model over the baseline are significant.

Scaling factors for the generated dataset. The scale of a dataset is a crucial determinant of its effectiveness. We investigate various design choices that influence the size of the generated datasets, identifying the critical factors for efficient data scaling. We mainly explore two factors: density of entity and description.

COCO	O365
28.5 / 18.6	29.5 / 21.3

We first study the density of the covered entity by scaling the category set from COCO [13] to Object365 [21]. We generate dense synthetic triplets for each set and use them to train a detector. As shown in AP-d/AP-dL above, the description-based performance gradually improved as the scale of the visual entity grew. This implies that it is crucial to learn from dense triplets of diverse visual entities.

5 per ent.	10 per ent.	20 per ent.
27.5 / 17.4	28.4 / 18.4	29.5 / 21.3

We also explore the number of generated descriptions for each visual entity. We vary the number from 5 to 20 and report the AP-d/AP-dL of the trained detector above. The number of descriptions per entity greatly impacts overall scores, especially on the long query. This shows the importance of dense triplets and highlights the potential of our easy-to-scalable synthetic data generation framework.

4. Conclusion

In this paper, we propose to automatically generate synthetic triplets of diverse and complex text descriptions, corresponding images, and reliable pseudo-bounding boxes. With the synergy of the synthetic triplets and proposed compositional contrastive learning, our generic framework largely improves the compositional understanding of diverse language-based object detectors.

Acknowledgement This work was supported by the Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2022-0-00871, Development of AI Autonomy and Knowledge Enhancement for AI Agent Collaboration).

References

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

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1, 2, 3

[3] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 1, 2, 3

[4] Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. *Advances in neural information processing systems*, 35:32942–32956, 2022. 2, 3, 4

[5] Sivan Doveh, Assaf Arbelle, Sivan Harary, Rameswar Panda, Roei Herzig, Eli Schwartz, Donghyun Kim, Raja Giryes, Rogerio Feris, Shimon Ullman, et al. Teaching structured vision&language concepts to vision&language models. *arXiv preprint arXiv:2211.11733*, 2022. 1, 2

[6] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 1

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

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[9] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017. 2, 3

[10] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. 2021. 1

[11] Liunian Li, Zi-Yi Dou, Nanyun Peng, and Kai-Wei Chang. Desco: Learning object recognition with rich language descriptions. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 3, 4

[12] 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, 3, 4

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

[14] 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. 2

[15] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 4

[16] 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. 1, 2

[17] 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

[18] 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. 1

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

[20] Samuel Schulter, Yumin Suh, Konstantinos M Dafnis, Zhixing Zhang, Shiyu Zhao, Dimitris Metaxas, et al. Omnilabel: A challenging benchmark for language-based object detection. 2023. 1, 3, 4

[21] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 1, 2, 3, 4

[22] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[23] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022. 1

[24] Chi Xie, Zhao Zhang, Yixuan Wu, Feng Zhu, Rui Zhao, and Shuang Liang. Described object detection: Liberating ob-

ject detection with flexible expressions. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 3, 4

[25] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*, 2023. 1, 3

[26] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022. 4

[27] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pages 350–368. Springer, 2022. 4