# Weak-to-Strong Compositional Learning from Generative Models for Language-based Object Detection

Kwanyong Park[1], Kuniaki Saito[2], and Donghyun Kim[3⋆]

[1] ETRI
[2] OMRON SINIC X Corporation
[3] 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. Traditional approaches attempt to improve VL models using hard negative synthetic text, but their effectiveness is limited. In this paper, we harness the exceptional compositional understanding capabilities of generative foundational models. We introduce a novel method for structured synthetic data generation aimed at enhancing the compositional understanding of VL models in language-based object detection. Our framework generates densely paired positive and negative triplets (image, text descriptions, and bounding boxes) in both image and text domains. By leveraging these synthetic triplets, we transform 'weaker' VL models into 'stronger' models in terms of compositional understanding, a process we call "Weak-to-Strong Compositional Learning" (WSCL). To achieve this, 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.

**Keywords:** Compositionality · Language-based Object Detection

## 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 [22,46]. While traditional recognition models [14,29,48,53] are restricted to classifying or detecting pre-defined classes, image-text paired data allow models to easily generalize to new concepts and domains with language queries. For example, GLIP [27] can perform phrase grounding or detect multiple objects in language queries by learning to align words and regions in each modality.

⋆ Corresponding author

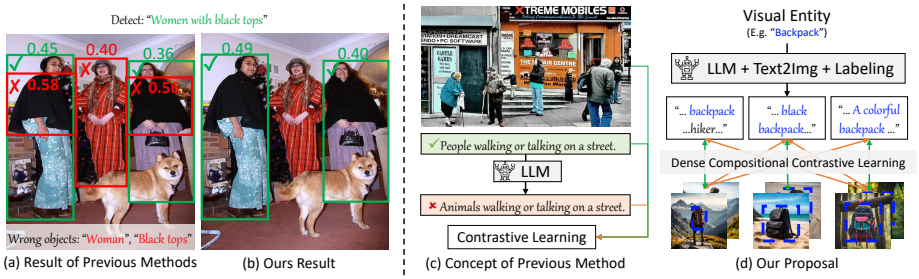


**Fig. 1:** (a-b) While previous models lack a compositional understanding of the given language query and localize wrong objects, resulting in higher scores for the wrong objects (*i.e.*, the middle woman and black tops) than the actual object, our method successfully localizes only correct objects corresponding to the query. (c) Previous VL methods apply (*e.g.*, [10, 26]) augmentation exclusively in the text domain. (d) The proposed method produces comprehensive synthetic triplets comprising <image, object description, bounding box>, incorporating compositional contrastive learning to improve the model's understanding of composition.

Despite advancements, VL models [27, 46] continue to face challenges in understanding complex language queries and structured vision-language concepts, such as detailed object attributes, shapes, textures, and their relationships [10, 56, 67]. A recent study [67] indicates that VL models often function like bags-of-words, lacking compositional understanding. This results in a significant performance drop in image-text retrieval tasks involving complex scenes and detailed captions with rich compositional structures. In the context of object detection, novel benchmarks like OmniLabel [52] and $D^3$ [64] have been introduced to assess the ability to interpret a broad range of complex object descriptions and accurately detect target objects (See Fig. 1-(a)). In such scenarios, VL models frequently overlook the complex and free-form textual descriptions provided, leading to incorrect detection results. To address this issue, previous work [26] has explored augmenting the text domain [10, 56, 67] by generating synthetic negative texts through noun swapping or creating new image captions (See Fig. 1-(c)). However, we observe that merely enriching the text domain is insufficient for models to learn dense relations between images and text.

To this end, we propose an innovative framework to distill the unprecedented compositional understanding of recent generative foundational models, such as large language models [1,3,59] and text-to-image diffusion models [6,25,44,50,51], into VL models. Within our framework, a series of generative models automatically generates synthetic data, from which a language-based object detector learns and inherits compositionality. Through this process, 'weaker' VL models evolve into 'stronger' models in terms of compositional understanding; we term this process "Weak-to-Strong Compositional Learning" (WSCL). Fig. 1-(d) illustrates the proposed framework.

To be specific, our framework consists of two steps: (1) Generating diverse and dense triplets. Instead of solely relying on difficult-to-obtainable real-world

data, we propose to generate dense triplets (*i.e.*, <image, object description, bounding box>) with the generative models (Sec. 4.1). We first use a large language model 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. 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. Note that our generation framework is scalable due to its automatic data construction process. (2) Effective learning from densely generated triplets (Sec. 4.2). 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. This forces the detector to be aware of the given 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. We call the two synergetic contrastive learning methods as compositional contrastive learning.

We utilize our method to enhance two advanced language-based object detection models, namely GLIP [27] and FIBER [8]. On the challenging OmniLabel benchmark [52], our proposal achieves a notable improvement of +5.0 and +4.8AP upon GLIP-T and FIBER-B. This suggests that our method effectively enhances its compositional understanding of visual objects and descriptions. Specifically, for long queries, the performance of the GLIP-T model is doubled from 8.2 to 16.4AP. Besides, our proposal is proven to be complementary to the previous text augmentation-based method, DesCo [26], and achieves the new state-of-the-art. **Our contribution** can be summarized as follows:

1. To our knowledge, this is the first work to generate diverse and dense synthetic triplets for language-based object detection, which are hard to obtain without expensive human annotations.
2. We present a novel compositional contrastive learning approach that efficiently learns to comprehend intricate compositions in images and text, and aligns image regions with the correct textual descriptions.
3. Our method is model-agnostic and can be applied to diverse prior language-based object detectors. We show that our method significantly improves the performance of the prior detectors on the two challenging benchmarks, OmniLabel and $D^3$, across diverse model architectures.

## 2 Related Work

**Vision-language Models.** Vision-language (VL) models (*e.g.*, CLIP [46], ALIGN [22], GLIP [27]) shows remarkable progress in diverse visual recognition tasks. CLIP and ALIGN are pre-trained on large-scale weakly supervised image-text pairs collected from the web with image-level contrastive learning objectives. In

order to gain a fine-grained understanding of images, several methods such as GLIP [27] propose region-level contrastive alignment between image regions and words in the text. GLIP additionally leverages detection and phrase grounding benchmarks and enables context-free object detection with language queries. However, as studies in [10, 19, 35, 56, 67, 68], VL models exhibit a limited compositional understanding of complex scenes and rich text descriptions for object attributes, texture, and their relations. In order to address these, hard negative and positive augmentation techniques on the language domain have been proposed in [9, 10, 67] and improve its ability of compositional understandings. On the other hand, we propose to generate synthetic triplets including synthetic data in both image and text domains, and automatically generate bounding boxes for language-based object detection.

**Object Detection.** Traditional detection models are trained to detect objects for a pre-defined set of categories [4, 47, 48, 62]. As a result, traditional models find it challenging to adapt to new tasks and domains, unable to differentiate between objects that vary in attributes such as texture, shape, and other characteristics. Recently, language-based object detection with vision-language models has demonstrated significant potential to enhance their adaptability by utilizing language queries. CLIP [46] opens a new research direction in open-vocabulary object detection [13, 23, 24, 70] demonstrating strong performances on unseen categories by leveraging text encoders like BERT [7]. MDETR [23] detects objects conditioned on complex language queries containing object attributes and relations. However, MDETR struggles to perform effectively on the Omnilabel benchmark [52], which presents queries with more intricate and challenging negative descriptions in free-form text. DesCo [26] employs large language models to generate synthetic rich language descriptions to improve the compositional understanding of language queries. Conversely, our research focuses on enhancing language-based object detection by utilizing synthetic triplets that incorporate pseudo bounding boxes for every object description. This is a significant challenge as existing detectors lack compositional understanding. To address this, we transform the complex task into several simpler detection tasks, thereby achieving precise bounding boxes for each description.

**Learning from Synthetic Data.** Deep learning models require massive labeled data to obtain strong performances. However, it is expensive to collect such labeled data. On the other hand, synthetic data can be obtained easily to train a model. Learning from synthetic data has been an active research topic for many years in diverse computer vision applications such as image classification [12, 21, 36, 37, 66], object detection [20, 28, 42, 45], and image segmentation [40, 41, 49, 54, 60, 61]. These models utilize graphics engines to generate images, which causes a domain gap from real data. Recently, several works utilize text-to-image diffusion models [44, 50] to generate synthetic images for visual recognition [2, 11, 16, 38, 57, 58, 63]. However, in our experiments, naively adding synthetic data as a set of training data does not necessarily improve the compositional reasoning ability of VL models. Therefore, we introduce a new compositional contrastive learning that effectively utilizes synthetic image-text paired data for our task.

## 3 Background: Language-based Object Detection

Language-based object detection takes free-form language queries and an image as inputs to identify and predict bounding boxes, aligning these boxes with the corresponding language queries. The task encompasses an open-set and multi-label framework, where queries may include descriptions of objects with unseen and intricate compositions [52, 64]. Furthermore, the descriptions may correspond to zero, one, or several instances within the image, diverging from typical object detection [29,53]. Such characteristics require a VL model to understand complex compositions in visual scenes and textual descriptions.

Several VL models (*e.g.*, GLIP [27] and FIBER [8]) are utilized to solve this task. We review GLIP [27], and its approach to addressing this task. GLIP redefines detection as a grounding task by matching each region or box $B$ in an image $I$ with phrases in a text query (prompt) $Q$ with a target alignment label $T$. The key is to transform existing data into a grounding format. For detection data, the query $Q$ contains a list of pre-defined object classes such as "Person. Bicycle, ..., glasses". For image-text paired data (*e.g.*, CC12M, SBU [5,39]), the query is a text caption containing entities in the image. Since $T$ is not available for this data, GLIP generates pseudo-grounding labels for the alignment between entities in the caption and regions in the image. Then a model is trained to align each word in the query $Q$ with each region $B$ in (pseudo) $T$ as follows:

$$O, P = G(I, Q), \quad S_{\text{ground}} = OP^{\top}, \quad \mathcal{L} = \mathcal{L}(S_{\text{ground}}, T) + \mathcal{L}_{loc} \tag{1}$$

where $G$ is the GLIP model, $O \in \mathbb{R}^{N \times d}$ are the regions features of $I$, $P \in \mathbb{R}^{M \times d}$ is the contextual word tokens features of $Q$, and $S_{\text{ground}} \in \mathbb{R}^{N \times M}$ is the alignment score. GLIP is trained to minimize the region-word matching and localization loss as in the standard object detection. GLIP struggles to identify the correct region in response to a complex query and fails to generate precise labels.

## 4 Method

We aim to improve the compositional understanding capabilities of a language-based object detector. 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 as training samples. Our approach involves two main steps: (1) dense synthetic triplet generation (Sec. 4.1) and (2) compositional contrastive learning with dense synthetic triplets (Sec. 4.2). In the first step, we introduce our method to generate diverse and semantically rich training triplets (*i.e.*, objects, object descriptions, and bounding boxes) in both image and text domains. Subsequently, we introduce compositional contrastive learning to effectively improve compositional understanding of visual objects and align with its complex object descriptions from our generated triplets for our language-based object detector.

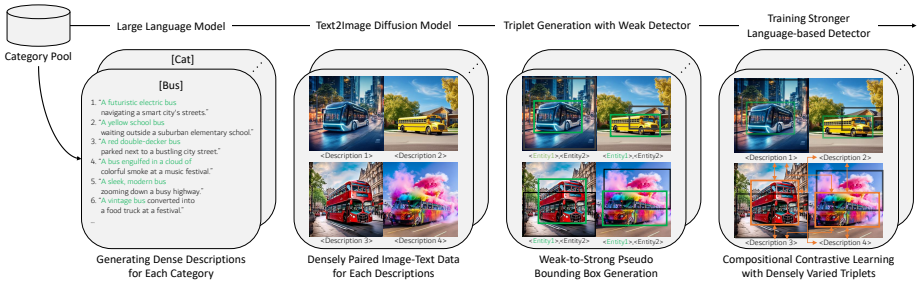


**Fig. 2:** Overview of our method. Our method consists of generating dense synthetic image-text paired triplets with generative models and creating bounding boxes. Finally, we introduce compositional conservative learning with our generated triplets which enhances the model's compositional ability in language-based object detection.

### 4.1 Synthetic Triplet Generation in Image and Text Domains

A traditional training data collection process for grounding data [29,43] is to *collect images*, and manually *annotate object bounding boxes with their text descriptions*. However, it would be prohibitively expensive to manually collect images that cover the full diversity of objects along with all their possible attributes, actions, and interactions with their environment. Additionally, localizing these objects and providing free-form textual descriptions of them is even more challenging. And, descriptions provided by human annotators are often brief and lack detail, which can hinder the effective learning of visual-language alignment. Furthermore, this process does not guarantee to obtain hard negatives (*i.e.*, dense triplets) which is crucial to improve the compositionality of VL models [10, 26]. In order to obtain diverse and dense triplets, we adopt a reversed approach which first *generates text descriptions* and then *collects corresponding images*: we begin by synthesizing diverse and plausible object descriptions, then proceed to generate corresponding images inspired by the recent breakthroughs in foundation models [1, 3, 6, 7, 25, 44, 50, 51], and finally, automatically localize the objects within these images. The overview of the proposed training data generation framework is depicted in Fig. 2. As a result, our method allows the automatic generation of dense triplets without requiring human annotation of text descriptions and bounding boxes.

**Generating Diverse Object Descriptions.** We aim to generate a collection of dense image-text pairs, where a wide variety of visual entities is depicted in the image and text domain. To achieve this, we initiate the process by generating diverse text descriptions for each entity with generative models. Recent advancements have demonstrated the remarkable capability of large language models (LLMs) [1, 3, 7, 59] to comprehend the real world in unprecedented detail. We capitalize on this knowledge by querying LLMs for plausible descriptions of objects under various conditions. 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. Representative examples are shown in Figs. 2 and 3.

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 [15, 30, 53]. The number of descriptions per entity is crucial for ensuring a comprehensive coverage of each entity's diversity, while the length of the descriptions influences the complexity of the resulting scenes. For example, longer descriptions tend to encompass more surrounding objects and intricate attributes, allowing us to tune the training samples' difficulty and quality.

**Generating Densely Paired Images with Diffusion Models.** While previous work focuses on synthetic text augmentation [9, 26], our objective is to acquire densely paired image-text data in both image and text domains with text-to-image generative models. Diffusion-based text-to-image generation models [17, 50, 55] have recently demonstrated their capability to produce high-fidelity, photo-realistic images. The latest breakthroughs [6, 44, 51] in foundational diffusion models enable the generation of complex scenes featuring multiple objects with detailed descriptions. Our research investigates the extent to which these diffusion models can enhance the task of language-based object detection.

We condition the image generation process on generated object descriptions. It is different from previous methods [65] that used simple, hand-written prompts (e.g., "a photo of a [NAME]"). 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.

We investigate the impact of generating a diverse set of images from a single description (*i.e.*, Image Density). By introducing varied initial noise into the diffusion model—achieved by manipulating random seeds—we generate different visuals of the same description. Examples of the variations are depicted in Fig. 3.

**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 often struggle to identify visual entities based on complex descriptions. Due to their compositional understanding capabilities, detectors like GLIP [27] inaccurately localize or completely overlook objects, as illustrated in Fig. 3-(b, left). This issue presents a new challenge in utilizing generated data for training purposes.

To this end, we delve into strategies for achieving precise object localization using weak detectors (in terms of compositional understanding), thereby facili-

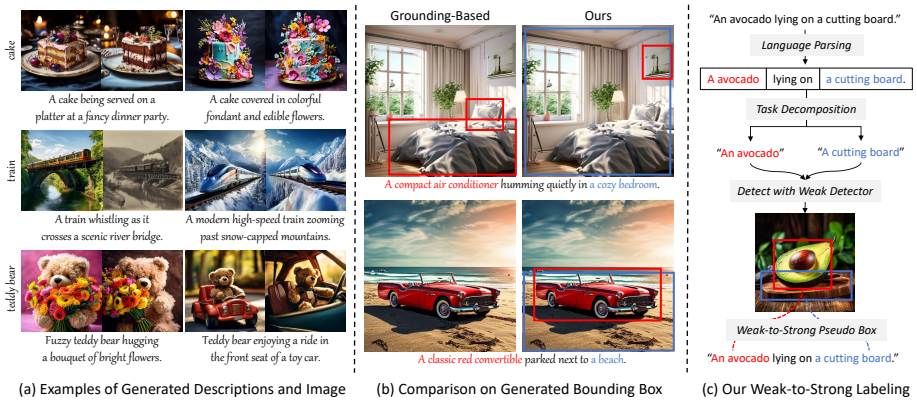


**Fig. 3:** (a) Qualitative examples of generated synthetic images and descriptions. (b) Comparison between grounding-based labeling and our weak-to-strong labeling. (c) Illustration of our weak-to-strong labeling, where we decompose the complex task into easy tasks. The bounding boxes collected from each task are combined to create strong compositional labels that train a strong detector.

tating the generation of rich supervision for training stronger detectors. We term this as a weak-to-strong labeling method. Our key idea is simple and intuitive: we decompose the complex phrase grounding problem into multiple manageable detection tasks. For this purpose, we make several key observations regarding the performance of recent visual-language detectors: 1) Although the detectors struggle to differentiate hard negative texts, they demonstrate proficiency in accurately localizing objects with positive texts (See the higher score for positive text (AP-dP) compared to overall detection Average Precision (AP-d) presented in Table 1.) 2) The model performs better at detecting objects described with concise text rather than complex descriptions. (See a higher score for short descriptions (AP-dS) compared to long descriptions (AP-dL).

Guided by the observations, we reformulate the complex phrase grounding problem into multiple tractable detection tasks with positive and short descriptions. An overview of our weak-to-strong labeling approach is depicted in Fig. 3-(c). For each pair of generated images and object descriptions, we initiate the process by identifying all noun phrases with an NLP parser [18]. We then treat each noun phrase as an independent description to detect the corresponding objects (task decomposition). This ensures satisfactory precision and recall, as demonstrated in Fig. 3-(b, right). Low-confidence predictions are filtered out based on a predetermined threshold p. The remaining predictions are re-assigned to the original position within the description, which results in a strong compositional label for the following step.

### 4.2 Description and Textural-structural Aware 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

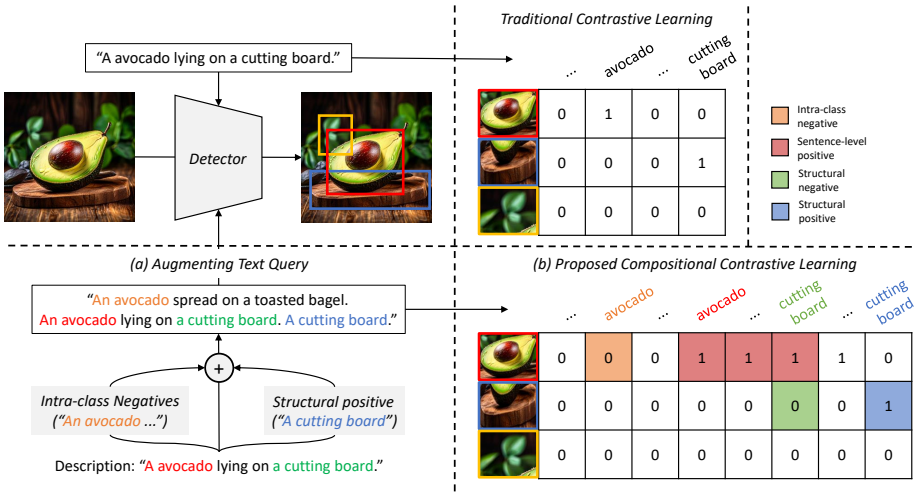


**Fig. 4:** Illustration of our compositional contrastive learning. (a) Intra-class negatives from other images of the same class and structural positives are introduced to learn the context of descriptions. (b) We associate the sentence-level positive (*i.e.*, the entire description sentence) with the 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".

alignment between noun phrases and detected object regions. However, our preliminary investigations reveal that models naively trained with these triplets often exhibit degraded performance. This raises the question: How can we effectively learn from the generated samples? Analyzing representative failure cases (See Fig. 1-(a)), we identify two critical functionalities for compositional understanding: description-awareness and textural-structural-awareness. We detail the methodologies for learning these functionalities using synthetic data and explore strategies to mitigate domain bias, thereby unlocking the synthetic data's full potential.

**Learning Description-awareness with Dense Contrastive Learning.** Traditional language-based detectors often lack description awareness, indiscriminately detecting entities, for example, detecting middle 'women' regardless of the provided descriptions in Fig. 1-(a). 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 Q with the negatives (*e.g.*, "An avocado spread on a toasted bagel" in Fig. 4). 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, indiscriminatively detecting all visual entities mentioned in the descriptions as detecting 'black top' in the Fig. 1-(a, left). To overcome this, we aim to distinguish between subject and non-subject entities within descriptions. We use textural relation [18] 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. 4) 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. 4) 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.

**Preventing Domain Bias.** While state-of-the-art diffusion models excel in producing high-quality, photo-realistic images, the synthesized images inevitably may exhibit artifacts. Moreover, even the most advanced diffusion models struggle to produce super-complex images with perfect text-to-image correspondence, leading to a loss of precise localization capabilities in complex scenes. This discrepancy raises the concern of language-based object detectors becoming overfitted to synthetic images, which could diminish their performance on real images.

To address these challenges, we propose two simple yet effective strategies: (1) Freezing the visual backbone while training detectors on synthesized data, which helps prevent the model's visual representations from overfitting to the synthetic distribution, and (2) Incorporating detection data as an additional training resource, which block the catastrophic forgetting of precise localization capabilities. These techniques collectively enable the model to seamlessly learn compositional visual language understanding without the risk of the domain gap.

## 5 Experiments

### 5.1 Experimental Setting

**Training Details.** We base our proposals on two recent language-based object detectors: GLIP and FIBER. Specifically, we utilize the GLIP-T and FIBER-B versions, which employ Swin Transformer [33] (Tiny and Base) as their visual backbones and BERT [7] and RoBERTa [32] as their language backbones. We finetune their official weights using a combination of our generated datasets and the Objects365 [53] object detection dataset. It should be noted that Objects365 has already been used in the training of both GLIP and FIBER. The inclusion of Objects365 aims to mitigate domain bias and preserve the detectors' innate ability to accurately localize objects within complex scenes, as detailed in the

| Model | Backbone | OmniLabel [52] AP | AP-c | AP-d | AP-dP | AP-dS | AP-dM | AP-dL | $D^3$ [64] Full | Pres | Abs |
|---|---|---|---|---|---|---|---|---|---|---|---|
| RegionCLIP [69] | ResNet-50 | 2.7 | 2.7 | 2.6 | 3.2 | 3.6 | 2.7 | 2.3 | - | - | - |
| Detic [70] | Swin-B | 8.0 | 15.6 | 5.4 | 8.0 | 5.7 | 5.4 | 6.2 | - | - | - |
| Grounding-DINO [31] | Swin-B | - | - | - | - | - | - | - | 20.7 | 20.1 | 22.5 |
| OFA-DOD [64] | Swin-B | - | - | - | - | - | - | - | 21.6 | 23.7 | 15.4 |
| GLIP-T [27] | 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 [8] | 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 [26] | 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 [26] | 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. We apply our method on top of diverse existing methods and significantly boost the performance.

methods section. By default, in synthetic data generation, we use the category pool from Object365, ChatGPT3.5-Turbo [3] for description generation, and Pixart [6] for image generation. For each category, we generate 20 descriptions and 8 images per description with different random seeds. In total, we generate 58,400 synthetic triplets. For additional details, please refer to the appendix.

**Evaluation Benchmarks.** We benchmark our proposed approach on the OmniLabel [52] and $D^3$ [64] datasets, following their official evaluation protocols. These datasets provide a comprehensive evaluation of the language-based object detector's proficiency in detecting objects specified by complex descriptions. Unlike traditional benchmarks in referring expressions, these datasets introduce scenarios with descriptions that either refer to no object or to multiple instances in an image, thereby facilitating a detailed compositional understanding in language-based object detection tasks.

Both benchmarks offer a suite of sub-metrics designed for an in-depth analysis. Specifically, for OmniLabel, the Average Precision for categories (AP-categ) and for descriptions (AP-descr) quantify detection accuracy for standard plain object categories and for free-form textual descriptions, respectively. The overall metric, AP, is computed as the harmonic mean between AP-categ and AP-descr, providing a balanced measure of both performances. Further dissecting description-based performance, the AP-descr-pos metric isolates the evaluation to positive descriptions, while AP-descr-S/M/L categorizes performance metrics according to the length of the descriptions (short, medium, and long), offering detailed insights into the detection efficacy relative to description complexity. The $D^3$ dataset categorizes descriptions into ABS ("absence") and PRES ("presence") based on whether the description includes expressions of absence (e.g., "without"). In addition to an overall evaluation metric encompassing all descriptions (referred to as FULL), $D^3$ provides distinct metrics for ABS and PRES.

### 5.2 Main Results

We evaluate the impact of the proposed learning framework with the densely generated triplets. Experimental results on OmniLabel and $D^3$ benchmarks are

summarized in Table 1. We first finetune two baseline models, GLIP and FIBER, and observe significant enhancements in 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 [26]). In this configuration, we apply their methods to enrich the language queries within the detection dataset 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.

### 5.3 Ablation Study and Analysis

To assess the impact of our proposed components, we conduct comprehensive ablation studies on the FIBER-B model.

| 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-only | 25.5 | 27.7 | 23.7 | 34.4 | 41.5 | 19.6 | 12.4 |
| (+) Det data | 26.3 | 30.2 | 23.3 | 34.2 | 41.0 | 19.7 | 11.5 |
| (+) Freeze vis-back | 26.8 | 31.3 | 23.4 | 34.4 | 40.8 | 19.5 | 11.8 |
| (+) Intra-neg | 29.0 | 30.9 | 27.4 | 36.6 | 44.2 | 24.0 | 14.9 |
| (+) Struct-neg | 29.0 | 31.0 | 27.3 | 37.1 | 43.7 | 24.4 | 16.2 |
| (+) Struct-pos | 30.5 | 31.6 | 29.5 | 40.3 | 43.7 | 26.3 | 21.3 |

**Table 2:** Ablation on compositional contrastive learning.

**Effective learning signals with synthetic data.** We validate the impact of the proposed learning methods. Experimental results are summarized in Table 2. We start by naive finetuning only on the densely generated triplets: treating these triplets similarly to conventional grounding data. (*i.e.*, Gen-only). While the description-based performance is improved, the precise localization capability with the given plain category is largely degraded. To mitigate the detrimental effects of the distributional discrepancies between generated and real-world data, we employ common object detection datasets [53] as a form of regularization and freeze the visual backbone during training. As shown in the table, each learning technique helps to maintain or even improve precise localization capability and thus enables solid learning from the synthetically generated datasets. Next, we explore the impact of the proposed contrastive learning methods. By contrasting dense descriptions from the same visual entity (*i.e.*, Intra-neg), 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. Naively treating the non-subject object as negative for the description doesn't bring notable improvements (*i.e.*, Struct-neg). However, when the concept of structural positive is included, the model is enforced to discriminate the same phrases according to their structural role in the description. This

| category | AP | AP-c | AP-d | AP-dS | AP-dL |
|---|---|---|---|---|---|
| COCO (80) | 29.7 | 31.0 | 28.5 | 43.8 | 18.6 |
| **O365 (365)** | 30.5 | 31.6 | 29.5 | 43.7 | 21.3 |
| LVIS (1203) | 31.1 | 30.9 | 31.3 | 45.3 | 23.3 |

**(a)** Entity Diversity

| num. des. | AP | AP-c | AP-d | AP-dS | AP-dL |
|---|---|---|---|---|---|
| 5 per ent. | 29.1 | 31.0 | 27.5 | 42.2 | 17.4 |
| 10 per ent. | 29.7 | 31.1 | 28.4 | 43.6 | 18.4 |
| **20 per ent.** | 30.5 | 31.6 | 29.5 | 43.7 | 21.3 |

**(b)** Description Density

| num. img. | AP | AP-c | AP-d | AP-dS | AP-dL |
|---|---|---|---|---|---|
| 2 per des. | 29.8 | 31.2 | 28.6 | 43.1 | 19.3 |
| 4 per des. | 29.7 | 31.3 | 28.2 | 42.0 | 19.2 |
| **8 per des.** | 30.5 | 31.6 | 29.5 | 43.7 | 21.3 |

**(c)** Image Density

**Table 3:** Analysis on scaling factors for generated triplets.

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 three factors: density of entity, description, and image.

We first study the density of the covered entity by scaling the category set. We borrow a well-curated list of classes from COCO [29], Object365 [53] and LVIS [14]. We generate dense synthetic triplets for each set and use them to train a detector. As shown in Table 3a, 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. On the contrary, for the plain-category name-based detection, the set of the Object365 class shows the best performance. This is because existing diffusion models also suffer from long-tailed issues and have trouble illustrating uncommon objects. Considering the balance between AP-c and AP-d, we use the category pool of Object365 as the default for other experiments. Our default setting is noted in bold.

We also explore the number of generated descriptions for each visual entity. We vary the number from 5 to 20 and report the performance of the detector trained on corresponding generated triplets in Table 3b. 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 an easy-to-scalable synthetic data generation framework.

Lastly, we study whether the density of generated images matters for the efficiency of the framework. To generate diverse images for a given description, we generate multiple variations by introducing different initial noises into the diffusion models, achieved by varying the random seed. We adjust the number of random seeds used for image generation from 2 to 8. As indicated in Table 3c, the diversity of images proves beneficial. The model benefits from learning across multiple visual variations of a single description, leading to a robust alignment between visual and linguistic representations.

**Pseudo box generation strategy.** We study the impact of the pseudo bounding box generation strategy on the final performance. As shown in Table 4, the proposed weak-to-strong

| strategy | AP | AP-c | AP-d | AP-dp | AP-dS | AP-dM | AP-dL |
|---|---|---|---|---|---|---|---|
| Grounding-based | 29.3 | 31.3 | 27.5 | 37.4 | 43.4 | 24.1 | 16.2 |
| **Weak-to-Strong** | 30.5 | 31.6 | 29.5 | 40.3 | 43.7 | 26.3 | 21.3 |

**Table 4:** Ablation on pseudo label generation strategies.

method brings notable improvements compared to conventional grounding-based technique. This shows the importance of the quality of the bounding box for compositional learning.

**Effective description length.** As highlighted in the methods section, the specified length of the descriptions affects the complexity of the object descriptions and the resultant images. To demonstrate this concept, we adjust the requested description lengths from 6 to 12 words and conduct a textural analysis with an NLP parser [34]. In Table 5, we report the average number of nouns and adjectives per description, which correlates with the number of objects and their specified attributes, respectively. Monotonic increased factors over the description length show a positive correlation between the requested description length and scene complexity.

| des. length | AP | AP-c | AP-d | AP-dS | AP-dL | NOUN | ADJ |
|---|---|---|---|---|---|---|---|
| 6 words | 29.7 | 31.0 | 28.5 | 43.8 | 18.6 | 2.89 | 0.87 |
| 8 words | 30.3 | 31.2 | 29.3 | 43.0 | 20.5 | 3.29 | 1.04 |
| **10 words** | 30.5 | 31.6 | 29.5 | 43.7 | 21.3 | 4.10 | 1.49 |
| 12 words | 29.9 | 31.5 | 28.5 | 43.8 | 18.3 | 4.24 | 1.52 |

**Table 5:** Additional analysis on the effective length of descriptions.

We then evaluate the effectiveness of our learning framework as the complexity of the generated image/text combinations varies. Although our approach performs robustly across all description lengths, optimal results were observed at a description length of 10 words. Short descriptions tend to generate overly simplistic descriptions and images, which are insufficient for learning nuanced description and structure sensitivity. Conversely, longer descriptions risk exceeding the capabilities of state-of-the-art models, potentially leading to generating images that are more likely to contain artifacts, such as missing objects or inaccurately depicted attributes. This may bring noise in the language-based object detector training.

**Additional analyses.** We present further ablation studies and analyses on various factors, such as pseudo box generation strategy, frozen backbones, the choice of diffusion models, and the efficiency of our framework. Additionally, we include qualitative detection results from our models and others in the supplementary materials.

## 6 Conclusion

Although vision-language (VL) models have made notable progress in language-based object detection, they continue to face challenges in comprehensively understanding the compositions of visual scenes and textual descriptions. This leads to a noticeable decline in performance when faced with complex language queries. To our knowledge, we first propose to automatically generate synthetic triplets containing diverse and complex text descriptions, corresponding images, and reliable pseudo-bounding boxes. These synthetic triplets lead a VL model to learn compositional capability with our proposed compositional contrastive learning. Our approach is model-agnostic, which can be applied to improve diverse existing VL models and significantly boost the performance on this challenging task.

## Acknowledgement

This study was supported by the following grants: 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, 70%), (No. RS-2019-II190079, Artificial Intelligence Graduate School Program(Korea University), 5%), the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(No. RS-2024-00341514, 15%), Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2024(Project Name: International Collaborative Research and Global Talent Development for the Development of Copyright Management and Protection Technologies for Generative AI, Project Number: (RS-2024-00345025, 10%)

## References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023) 2, 6
2. Azizi, S., Kornblith, S., Saharia, C., Norouzi, M., Fleet, D.J.: Synthetic data from diffusion models improves imagenet classification. arXiv preprint arXiv:2304.08466 (2023) 4
3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020) 2, 6, 11
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020) 4
5. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: 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. pp. 3558–3568 (2021) 5
6. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., et al.: Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. arXiv preprint arXiv:2310.00426 (2023) 2, 6, 7, 11
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018) 4, 6, 10
8. Dou, Z.Y., Kamath, A., Gan, Z., Zhang, P., Wang, J., Li, L., Liu, Z., Liu, C., LeCun, Y., Peng, N., et al.: Coarse-to-fine vision-language pre-training with fusion in the backbone. Advances in neural information processing systems **35**, 32942–32956 (2022) 3, 5, 11
9. Doveh, S., Arbelle, A., Harary, S., Herzig, R., Kim, D., Cascante-Bonilla, P., Alfassy, A., Panda, R., Giryes, R., Feris, R., et al.: Dense and aligned captions (dac) promote compositional reasoning in vl models. Advances in Neural Information Processing Systems **36** (2024) 4, 7

10. Doveh, S., Arbelle, A., Harary, S., Panda, R., Herzig, R., Schwartz, E., Kim, D., Giryes, R., Feris, R., Ullman, S., et al.: Teaching structured vision&language concepts to vision&language models. arXiv preprint arXiv:2211.11733 (2022) 2, 4, 6
11. Fan, L., Chen, K., Krishnan, D., Katabi, D., Isola, P., Tian, Y.: Scaling laws of synthetic images for model training... for now. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7382–7392 (2024) 4
12. Gan, C., Schwartz, J., Alter, S., Mrowca, D., Schrimpf, M., Traer, J., De Freitas, J., Kubilius, J., Bhandwaldar, A., Haber, N., et al.: Threedworld: A platform for interactive multi-modal physical simulation. arXiv preprint arXiv:2007.04954 (2020) 4
13. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. arXiv preprint arXiv:2104.13921 (2021) 4
14. Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5356–5364 (2019) 1, 13
15. Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5356–5364 (2019) 7
16. He, R., Sun, S., Yu, X., Xue, C., Zhang, W., Torr, P., Bai, S., Qi, X.: Is synthetic data from generative models ready for image recognition? arXiv preprint arXiv:2210.07574 (2022) 4
17. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020) 7
18. Honnibal, M., Montani, I.: spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing (2017), to appear 8, 10
19. Hsieh, C.Y., Zhang, J., Ma, Z., Kembhavi, A., Krishna, R.: Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. Advances in neural information processing systems **36** (2024) 4
20. Hsu, C.C., Tsai, Y.H., Lin, Y.Y., Yang, M.H.: Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. pp. 733–748. Springer (2020) 4
21. Hur, S., Shin, I., Park, K., Woo, S., Kweon, I.S.: Learning classifiers of prototypes and reciprocal points for universal domain adaptation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 531–540 (2023) 4
22. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q.V., Sung, Y., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision (2021) 1, 3
23. Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N.: Mdetr-modulated detection for end-to-end multi-modal understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1780–1790 (2021) 4
24. Kim, D., Angelova, A., Kuo, W.: Region-aware pretraining for open-vocabulary object detection with vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11144–11154 (2023) 4
25. Lee, Y., Park, K., Cho, Y., Lee, Y.J., Hwang, S.J.: Koala: self-attention matters in knowledge distillation of latent diffusion models for memory-efficient and fast image synthesis. arXiv preprint arXiv:2312.04005 (2023) 2, 6

26. Li, L., Dou, Z.Y., Peng, N., Chang, K.W.: Desco: Learning object recognition with rich language descriptions. Advances in Neural Information Processing Systems **36** (2024) 2, 3, 4, 6, 7, 11, 12
27. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10965–10975 (2022) 1, 2, 3, 4, 5, 7, 11
28. Lin, S., Wang, K., Zeng, X., Zhao, R.: Explore the power of synthetic data on few-shot object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 638–647 (2023) 4
29. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014) 1, 5, 6, 13
30. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014) 7
31. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023) 11
32. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019) 10
33. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021) 10
34. Loper, E., Bird, S.: Nltk: The natural language toolkit. arXiv preprint cs/0205028 (2002) 14
35. Ma, Z., Hong, J., Gul, M.O., Gandhi, M., Gao, I., Krishna, R.: Crepe: Can vision-language foundation models reason compositionally? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10910–10921 (2023) 4
36. Mikami, H., Fukumizu, K., Murai, S., Suzuki, S., Kikuchi, Y., Suzuki, T., Maeda, S.i., Hayashi, K.: A scaling law for synthetic-to-real transfer: How much is your pre-training effective? arXiv preprint arXiv:2108.11018 (2021) 4
37. Mishra, S., Panda, R., Phoo, C.P., Chen, C.F.R., Karlinsky, L., Saenko, K., Saligrama, V., Feris, R.S.: Task2sim: Towards effective pre-training and transfer from synthetic data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9194–9204 (2022) 4
38. Nguyen, Q., Vu, T., Tran, A., Nguyen, K.: Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. Advances in Neural Information Processing Systems **36** (2024) 4
39. Ordonez, V., Kulkarni, G., Berg, T.: Im2text: Describing images using 1 million captioned photographs. Advances in neural information processing systems **24** (2011) 5
40. Park, K., Woo, S., Oh, S.W., Kweon, I.S., Lee, J.Y.: Mask-guided matting in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1992–2001 (2023) 4

41. Park, K., Woo, S., Shin, I., Kweon, I.S.: Discover, hallucinate, and adapt: Open compound domain adaptation for semantic segmentation. Advances in Neural Information Processing Systems **33**, 10869–10880 (2020) 4
42. Peng, X., Sun, B., Ali, K., Saenko, K.: Learning deep object detectors from 3d models. In: Proceedings of the IEEE international conference on computer vision. pp. 1278–1286 (2015) 4
43. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE international conference on computer vision. pp. 2641–2649 (2015) 6
44. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023) 2, 4, 6, 7
45. Prakash, A., Boochoon, S., Brophy, M., Acuna, D., Cameracci, E., State, G., Shapira, O., Birchfield, S.: Structured domain randomization: Bridging the reality gap by context-aware synthetic data. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 7249–7255. IEEE (2019) 4
46. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 1, 2, 3, 4
47. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016) 4
48. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015) 1, 4
49. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: 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. pp. 102–118. Springer (2016) 4
50. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 2, 4, 6, 7
51. Sauer, A., Lorenz, D., Blattmann, A., Rombach, R.: Adversarial diffusion distillation. arXiv preprint arXiv:2311.17042 (2023) 2, 6, 7
52. Schulter, S., Suh, Y., Dafnis, K.M., Zhang, Z., Zhao, S., Metaxas, D., et al.: Omnilabel: A challenging benchmark for language-based object detection (2023) 2, 3, 4, 5, 11
53. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8430–8439 (2019) 1, 5, 7, 10, 12, 13
54. Shin, I., Park, K., Woo, S., Kweon, I.S.: Unsupervised domain adaptation for video semantic segmentation. arXiv preprint arXiv:2107.11052 (2021) 4
55. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020) 7
56. Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., Ross, C.: Winoground: Probing vision and language models for visio-linguistic composition-

ality. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5238–5248 (2022) 2, 4
57. Tian, Y., Fan, L., Chen, K., Katabi, D., Krishnan, D., Isola, P.: Learning vision from models rivals learning vision from data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15887–15898 (2024) 4
58. Tian, Y., Fan, L., Isola, P., Chang, H., Krishnan, D.: Stablerep: Synthetic images from text-to-image models make strong visual representation learners. Advances in Neural Information Processing Systems **36** (2024) 4
59. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023) 2, 6
60. Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7472–7481 (2018) 4
61. Wang, Z., Yu, M., Wei, Y., Feris, R., Xiong, J., Hwu, W.m., Huang, T.S., Shi, H.: Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12635–12644 (2020) 4
62. Woo, S., Park, K., Oh, S.W., Kweon, I.S., Lee, J.Y.: Bridging images and videos: A simple learning framework for large vocabulary video object detection. In: European Conference on Computer Vision. pp. 238–258. Springer (2022) 4
63. Wu, W., Zhao, Y., Chen, H., Gu, Y., Zhao, R., He, Y., Zhou, H., Shou, M.Z., Shen, C.: Datasetdm: Synthesizing data with perception annotations using diffusion models. Advances in Neural Information Processing Systems **36**, 54683–54695 (2023) 4
64. Xie, C., Zhang, Z., Wu, Y., Zhu, F., Zhao, R., Liang, S.: Described object detection: Liberating object detection with flexible expressions. Advances in Neural Information Processing Systems **36** (2024) 2, 5, 11
65. Xie, J., Li, W., Li, X., Liu, Z., Ong, Y.S., Loy, C.C.: Mosaicfusion: Diffusion models as data augmenters for large vocabulary instance segmentation. arXiv preprint arXiv:2309.13042 (2023) 7
66. Xu, T., Chen, W., Wang, P., Wang, F., Li, H., Jin, R.: Cdtrans: Cross-domain transformer for unsupervised domain adaptation. arXiv preprint arXiv:2109.06165 (2021) 4
67. Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., Zou, J.: When and why vision-language models behave like bags-of-words, and what to do about it? In: International Conference on Learning Representations (2023), https://openreview.net/forum?id=KRLUvxh8uaX 2, 4
68. Zhao, T., Zhang, T., Zhu, M., Shen, H., Lee, K., Lu, X., Yin, J.: Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. arXiv preprint arXiv:2207.00221 (2022) 4
69. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al.: Regionclip: Region-based language-image pretraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16793–16803 (2022) 11
70. Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., Misra, I.: Detecting twenty-thousand classes using image-level supervision. In: European Conference on Computer Vision. pp. 350–368. Springer (2022) 4, 11