DetCLIPv3: Towards Versatile Generative Open-vocabulary Object Detection

Lewei Yao[1,2], Renjie Pi[1], Jianhua Han[2], Xiaodan Liang[3],
Hang Xu[2†], Wei Zhang[2], Zhenguo Li[2], Dan Xu[1†]

[1]Hong Kong University of Science and Technology, [2]Huawei Noah's Ark Lab
[3]Shenzhen Campus of Sun Yat-Sen University

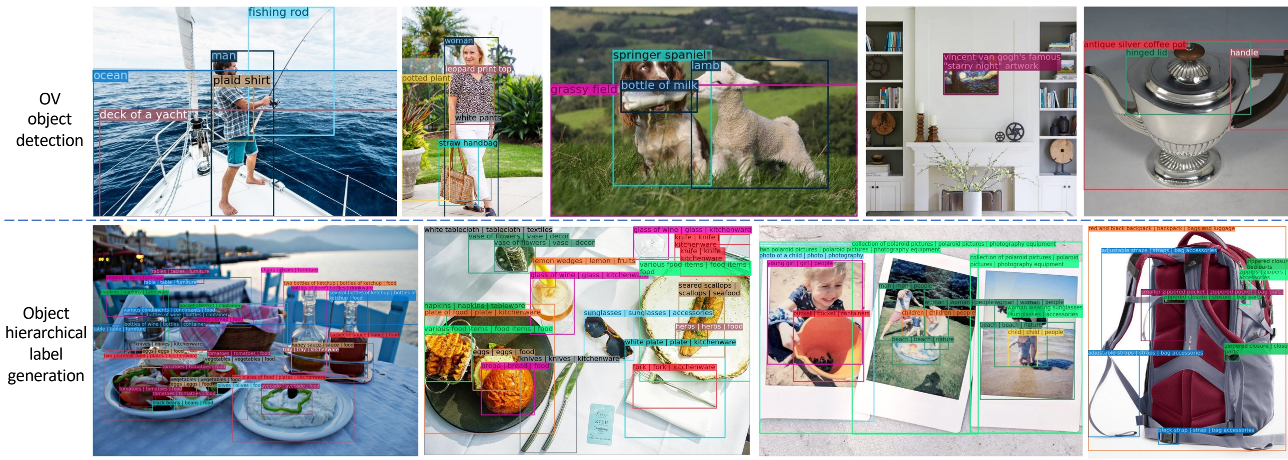


Figure 1. The versatility of DetCLIPv3 supports both open-vocabulary object detection (OVD) and the generation of hierarchical object labels. **Top**: when provided with extracted noun phrases from image-text pair captions as input, DetCLIPv3 can detect a broad spectrum of visual concepts. **Bottom**: In the absence of predefined categories as input, DetCLIPv3 detects potential objects and generates multi-granularity hierarchical labels for them, formatted as '*phrase* | *category* | *parent category*'. DetCLIPv3 offers a more comprehensive interpretation of objects, significantly expanding the application scope of OVD systems. Zoom in for the best viewing.

Abstract

Existing open-vocabulary object detectors typically require a predefined set of categories from users, significantly confining their application scenarios. In this paper, we introduce DetCLIPv3, a high-performing detector that excels not only at both open-vocabulary object detection, but also generating hierarchical labels for detected objects. DetCLIPv3 is characterized by three core designs: 1. Versatile model architecture: we derive a robust open-set detection framework which is further empowered with generation ability via the integration of a caption head. 2. High information density data: we develop an auto-annotation pipeline leveraging visual large language model to refine captions for large-scale image-text pairs, providing rich, multi-granular object labels to enhance the training. 3. Efficient training strategy: we employ a pre-training stage with low-resolution inputs that enables the object captioner to efficiently learn a broad spectrum of visual concepts from extensive image-text paired data. This is followed by a fine-tuning stage that leverages a small number of high-resolution samples to further enhance detection performance. With these effective designs, DetCLIPv3 demonstrates superior open-vocabulary detection performance, e.g., our Swin-T backbone model achieves a notable 47.0 zero-shot fixed AP on the LVIS minival benchmark, outperforming GLIPv2, GroundingDINO, and DetCLIPv2 by 18.0/19.6/6.6 AP, respectively. DetCLIPv3 also achieves a state-of-the-art 19.7 AP in dense captioning task on VG dataset, showcasing its strong generative capability.

1. Introduction

Recent progress in open-vocabulary object detection (OVD) has achieved the ability to identify and localize a diverse range of objects [2, 11, 12, 20, 27, 34, 54, 55, 62]. However, these models are limited by their reliance on a predefined object category list during inference, which hinders their usage in practical scenarios.

In contrast to current open-vocabulary object detection (OVD) methods that recognizes objects solely based on category names, human cognition demonstrates much more versatility. As illustrated in Figure 2, humans are able to understand objects from different granularities, in a hierarchical manner. This multi-level recognition ability showcases the rich visual understanding that humans possess, which is yet to be achieved in contemporary OVD systems.

To address the above limitations, we introduce DetCLIPv3, a novel object detector that enhances the scope of open-vocabulary object detection. DetCLIPv3 is able to not only recognize objects based on provided category names but also generate hierarchical labels for each detected ob-

†Corresponding author: xu.hang@huawei.com, danxu@cse.ust.hk

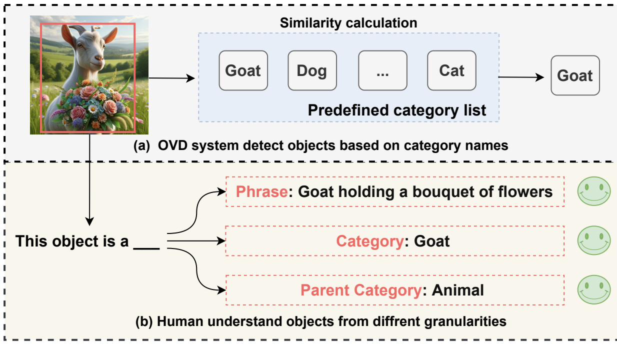


Figure 2. (a): Existing open-vocabulary object detectors recognize objects based on category names; (b): Humans interpret visual concepts from multiple hierarchies and granularities.

ject. This feature offers two advantages: 1) owing to the superior generative ability, the detector is applicable even in the absence of the appropriate input object category; 2) the model is able to provide a comprehensive and hierarchical description about the objects, rather than simply recognizing them based on given categories. Specifically, DetCLIPv3 is characterized by three core designs:

Versatile model architecture: DetCLIPv3 is grounded on a robust open-vocabulary (OV) detector, which is further empowered with a object captioner to provide generative capabilities. Specifically, the object captioner leverages the foreground proposals provided by the OV detector and is trained to generate hierarchical labels for each detected object through a language modeling training objective. This design allows for not only accurate localization, but also detailed descriptions of the visual concepts, and thereby offering a richer interpretation of the visual contents.

High information density data: The development of strong generative ability necessitates abundant training data enriched with detailed object-level descriptions. The scarcity of such comprehensive datasets (*e.g.*, Visual Genome [23]) presents a substantial obstacle in training effective object captioners. On the other hand, while large-scale image-text pair data are plentiful, they lack fine-grained annotations for each object. To benefit from such data, we design an auto-labeling pipeline leveraging state-of-the-art vision large language models [6, 33], which is able to provide refined image captions containing rich hierarchical object labels. With this pipeline, we derive a large-scale dataset (termed as **GranuCap50M**) for bolstering DetCLIPv3 's abilities in both detection and generation.

Efficient multi-stage training: The prohibitive training costs associated with high-resolution inputs for object detectors present a significant barrier to learning from extensive image-text pairs. To address the issue, we propose an efficient multi-stage alignment training strategy. This method initially harnesses knowledge from large-scale, low-resolution image-text datasets, followed by fine-tuning on high-quality, fine-grained, high-resolution data. The approach ensures comprehensive visual concept learning while maintaining manageable training demands.

With the effective designs, DetCLIPv3 achieves outstanding detection and object-level generation capabilities, *e.g.*, with a Swin-T backbone, it achieves a remarkable 47.0 zero-shot *fixed* AP [8] on the LVIS minival benchmark, significantly outperforming predecessors like GLIPv2 [60], DetCLIPv2 [55], and GroundingDINO [34]. Besides, it achieves 18.4 mAP on dense captioning task, surpassing the previous SOTA method GRiT [52] by 2.9 mAP. Extensive experiments further demonstrate the superior domain generalization and downstream transferbility of DetCLIPv3.

2. Related works

Open-vocabulary object detection. Recent advancements in Open-vocabulary Object Detection (OVD) allow for identifying objects across unlimited range of categories, as seen in [14, 15, 53, 58, 64]. These approaches achieve OVD by incorporating pre-trained VL models like CLIP [42] into the detector. Alternatively, expanding the detection training dataset has shown promise [22, 27, 29, 34, 54, 55, 60, 65], which combine the datasets from various task, such as classification and visual grounding. Moreover, pseudo labeling has emerged as another effective strategy for augmenting training datasets, as demonstrated in [13, 27, 39, 54, 63, 64]. However, previous OVD methods still require a predefined object categories for detection, limiting their applicability in diverse scenarios. In contrast, our DetCLIPv3 is capable of generating rich hierarchical object labels even in the absence of category names.

Dense captioning. Dense captioning aims at generating descriptions for specific image areas [21, 26, 28, 47, 56]. Recently, CapDet [36] and GRiT [52] both equip the object detector with generative ability by introducing a captioner. However, they are only able to generate descriptions for limited visual concepts due to the scarcity of training data contained in *i.e.*, Visual Genome [23]. In contrast, we harness the rich knowledge in the large-scale image-text pairs, and enable the model to generate hierarchical label information for much broader spectrum of concepts.

Re-captioning for image-text paris. Recent studies [5, 24, 40, 57] highlight the issues present in current image-text pair data and have shown that recaptioned high-quality image-text pairs can significantly enhance the learning efficiency of various visual tasks, such as text-to-image generation [5, 40], image-text retrieval [24, 25] and image captioning [24, 57]. We extend this idea to open-vocabulary object detection and explore how to effectively utilize the object entity information contained in image-text pairs.

3. Method

In this section, we introduce the core designs of DetCLIPv3, which includes: (1) Model Architecture (Sec. 3.1) - eluci-

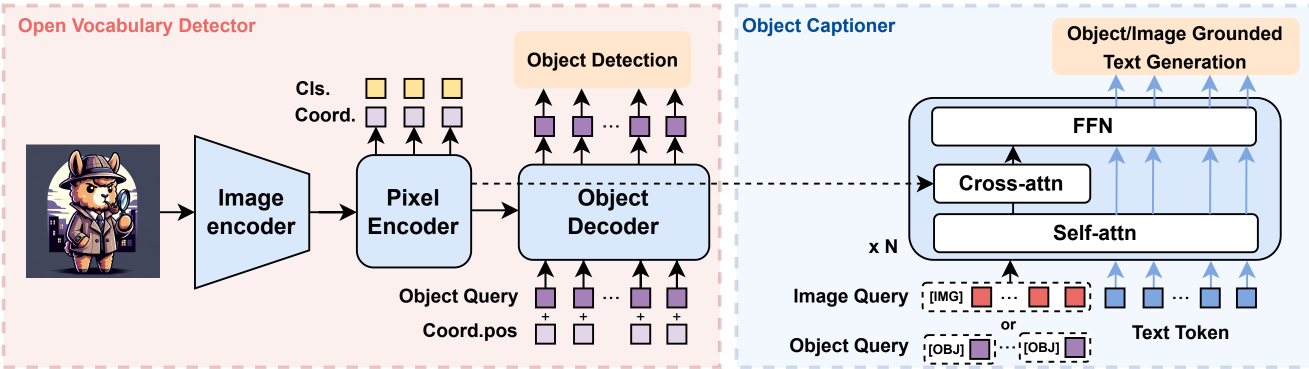


Figure 3. The illustration for DetCLIPv3 framework. Left: the OV detector is responsible for localizing objects given category names, as well as providing object proposals for the object captioner. Right: The object captioner is designed to generate hierarchical labels for detected objects and also learns to generate image-level descriptions as an aid to its training.

dating how our model enables both open-vocabulary object detection and generation of object descriptions; (2) Auto-Annotation Data Pipeline (Sec. 3.2) - detailing our approach for curating large-scale, high-quality image-text pairs, encompassing objects information across a spectrum of granularities; and (3) Training Strategy (Sec. 3.3) - outlining how we effectively leverage large-scale image-text datasets to facilitate object concept generation, which in turn boosts open-vocabulary detection capabilities.

3.1. Model Design

Figure 3 illustrates the overall framework of DetCLIPv3. In essence, the model is grounded on a powerful open-vocabulary detection object detector, and is equipped with an object captioner dedicated to generating hierarchical and descriptive object concepts. The model is able to function under two modes: 1) When a pre-defined category vocabularies list is provided, DetCLIPv3 predicts the localization of objects that are mentioned in the list; 2) in the absence of the vocabulary list, DetCLIPv3 is able to localize the objects and generate hierachical for each of them.

Data formulation. DetCLIPv3's training leverages datasets from multiple sources, including detection [46, 51], grounding [22], and image-text pairs [4, 44, 48, 49] with bounding-box pseudo-labels (as detailed in Sec. 3.2). Following DetCLIPv1/v2 [54, 55], we employ a *parallel formulation* to unify text inputs from various data sources into a uniform format. Specifically, each input sample is structured as a triplet, $(x, \{\mathbf{b}_i\}_{i=1}^N, y_{i=1}^M)$, where $x \in \mathbb{R}^{3\times H\times W}$ is the input image, $\{\mathbf{b}_i | \mathbf{b}_i \in \mathbb{R}^4\}_{i=1}^N$ represents a set of bounding boxes, and $y_{i=1}^M$ denotes a set of concept texts, comprising both positive and negative concepts.

For detection data, y_j comprises class names along with their definitions (as in [54, 55]), applicable in both training and testing phases. The negative concepts are sampled from categories within the dataset. For grounding and image-text pair data, the positive concepts are object descriptions, while the negatives are sampled from a large-scale noun corpus (details in Sec. 3.2). During training, to increase the number of negative concepts, we collect them across all training nodes and implement a deduplication process.

Open vocabulary detector. We present a compact yet powerful detector architecture for DetCLIPv3, which is depicted within the red box of Figure 3. Specifically, it is a dual-path model comprising a visual object detector Φ_v and a text encoder Φ_t. The visual object detector employs a transformer-based detection architecture [3, 61, 66], composed of a backbone, a pixel encoder, and an object decoder. The backbone and pixel encoder are responsible for extracting visual features, conducting fine-grained feature fusion, and proposing candidate object queries for the decoder. Similar to GroundingDINO [34], we utilize text features to select the top-k pixel features based on similarity, and later using their coordinate predictions to initialize the positional part of the decoder object query. However, distinctively, we abandon the computationally intensive cross-modal fusion modules designed in [34]. Following previous DETR-like detectors [3, 61, 66], our training loss is composed of three components: $\mathcal{L}_{det} = \mathcal{L}_{align} + \mathcal{L}_{box} + \mathcal{L}_{iou}$, where $\mathcal{L}_{align}$ is a contrastive focal loss [32] between regional visual features and textual concepts, while $\mathcal{L}_{box}$ and $\mathcal{L}_{iou}$ are the L1 loss and GIOU [43] loss, respectively. To boost the performance, auxiliary losses are employed at each layer of the decoder and at the output of the encoder.

Object captioner. The object captioner empowers DetCLIPV3 to generate detailed and hierarchical label for objects. To acquire the rich knowledge contained in image-text pairs, we further incorporate image-level captioning objective during training to enhance the generation capability. As illustrated in the blue box of Figure 3, the design of the object captioner is inspired by Qformer [25]. Specifically, it adopts a multi-modal Transformer-based architecture with its cross-attention layer replaced by deformable attention [66] tailored for dense prediction task. The captioner's input comprises both visual (object or image) queries and text tokens. The visual queries interact with features from the pixel encoder via cross-attention, while the self-attention layers and FFN layers are shared across different modalities. Furthermore, a multimodal causal self-attention mask [9, 25] is adopted to control the interaction between visual queries and text tokens. The training of the captioner is guided by the conventional language modeling loss $\mathcal{L}_{lm}$, with distinct input formats for object-level and image-level generation:

- **Object-Level generation**. The object query and the reference points required for the deformable cross-attention are derived from the final layer output of the object decoder. The input is structured as: $< objectquery, \texttt{[OBJ]}, text >$, where `[OBJ]` is a special task token indicating the object generation task. During training, we compute the loss using positive queries matching the ground truth. During inference, to obtain foreground proposals, we select the top-k candidate object queries based on their similarity to the most frequent 15K noun concepts from our curated noun corpus (Sec. 3.2). After generating hierarchical labels for these objects, we re-calibrate their objectness scores, using the OV detector to calculate the similarity between object queries and their generated 'phrase' and 'category' fields. The higher of these 2 similarities is then adopted as objectness score.
- **Image-Level generation**. Inspired by Qformer [25], we initialize 32 learnable image queries and use a set of fixed reference points. Specifically, we sample 32 locations from equal intervals from the reference points of the pixel encoder. Similar to object-level generation, the input is structured as $< imagequery, \texttt{[IMG]}, text >$, with `[IMG]` being a special task token indicating image generation. The inference process of image-level generation is consistent with the training.

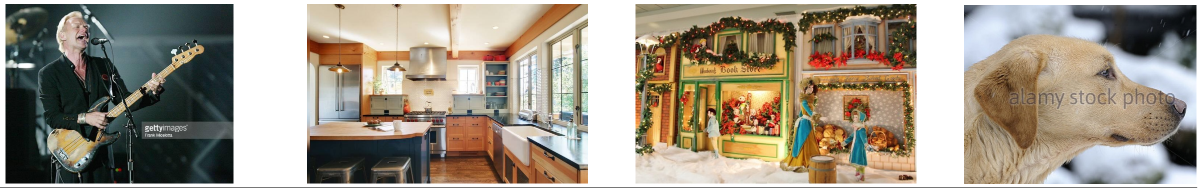
3.2. Dataset Construction

Auto-annotation data pipeline. Leveraging vast, cost-effective image-text pairs for visual concept learning is pivotal in enhancing the generalization capabilities of open-vocabulary object detectors. However, existing image-text pair datasets exhibit significant deficiencies that hinder their utility for OVD, as depicted in Figure 4: (1) **Misalignment**: Internet-sourced image-text pair data frequently contain substantial noise. Even with CLIP [42] score-based filtering [44, 45], many texts still fail to accurately describe the content of images, as shown in the 2nd and 3rd images of Figure 4. (2) **Partial annotation**: The majority of texts describe only the primary objects in images, resulting in sparse object information, and consequently, hurting learning efficiency of OVD systems, as observed in the 1st image. (3) **Entity extraction challenge**: Prior works [22, 30, 39, 55] primarily employ conventional NLP parsers, such as NLTK [1, 38] or SpaCy [19], to extract noun concepts from image-text pairs. Their limited capability can result in nouns that poorly align with the images' content, as illustrated in the second row of Figure 4. This mismatch poses further complications for subsequent learning process or pseudo-labeling workflow.

An ideal image-text pair dataset for OVD should encompass accurate and comprehensive descriptions of images, providing information about objects within images across a spectrum of granularity, from detailed to coarse. Motivated by this, we propose the use of a Visual Large Language Model (VLLM) [6, 33] to develop an automated annotation pipeline, improving the quality of data. VLLMs possess the capability to perceive the contents of images, as well as robust language skills, enabling them to generate precise and detailed captions as well as object descriptions. Specifically, our pipeline comprises the following processes:

1. **Recaptioning with VLLM**: We sample 240k image-text pairs from commonly used datasets [4, 48, 49] and conducted recaptioning using the InstructBLIP [6] model. To leverage the information from the original caption, we incorporate it into our prompt design, which is structured as: *"Given a noisy caption of the image:* {*raw caption*}*, write a detailed clean description of the image."*. This method effectively enhances the quality of the caption texts while maintaining the diversity of noun concepts in the original captions.
2. **Entity extraction using GPT-4**: We harness the exceptional language capability of GPT-4 [41] to process entity information in refined captions. Specifically, it is first utilized to filter out non-entity descriptions from the VLLM-generated captions, such as atmospheric or artistic interpretations of the images. Subsequently, it is tasked with extracting object entities present in the captions. Each entity was formatted into a triplet: {*phrase, category, parent category*}, representing object descriptions at three distinct levels of granularity.
3. **Instruction tuning of VLLM for large-scale annotation**: Considering the substantial costs of the GPT-4 API, its use for large-scale dataset generation is impractical. As a solution, we perform a further instruction tuning phase on a LLaVA [33] model, utilizing the improved captions and object entities obtained through prior steps. This finetuned model is then employed to produce captions and entity information for an extensive dataset comprising 200M image-text pairs, sampled from CC15M [4, 48], YFCC[49] and LAION [44].
4. **Auto-labelling for bounding boxes**: To automatically derive bounding box annotations for image-text paired data, we apply a pre-trained open-vocabulary object detector (Sec. 3.3) to assign pseudo bounding box labels given object entities derived from the previous steps. The accuracy of the detector can be greatly improved when provided with accurate candidate object entities

Input image



Raw text	rock artist performs on stage at awards held	8 Questions To Consider Before Meeting With A Home Designer Fox News	the Woodward's Windows	blonde labrador retriever in snow on a shoot day
Extracted nouns	1. rock; 2. artist; 3. stage; 4. awards	1. questions; 2. meeting; 3. home; 4. designer; 5. fox; 6. news	1. woodward; 2. windows	1. labrador; 2. snow; 3. shoot; 4. day
Reception text	A man is playing a bass guitar on stage during an awards ceremony. He is wearing a black suit and appears to be singing into a microphone while holding his guitar.	The image depicts a spacious kitchen with wooden cabinets, countertops, and appliances. There is a large island in the center. The kitchen also features a stainless steel refrigerator, oven, and dishwasher ...	The image features a Christmas-themed display in a store window, showcasing a variety of decorations and figurines. There are several mannequins dressed in Victorian-style clothing. Additionally, there are various Christmas trees and wreaths ...	The image features a blonde labrador retriever standing in the snow, looking up and away from the camera. The dog's head is tilted slightly to the side.
Extracted entities	1. 'Man playing a bass guitar' | 'Man' | 'Human' 2. 'Bass guitar' | 'Guitar' | 'Musical Instrument' 3. 'Stage' | 'Stage' | 'Location' 4. 'Black suit' | 'Suit' | 'Clothing' 5. 'Microphone' | 'Microphone' | 'Electronics' ...	1. 'Spacious kitchen' | 'kitchen' | 'Rooms in a house' 2. 'Wooden cabinets' | 'cabinets' | 'Furniture' 3. 'Countertops' | 'countertops' | 'Kitchen appliances' 4. 'Appliances' | 'appliances' | 'Kitchen appliances' 5. 'Large island' | 'island' | 'Kitchen furniture' ...	1. 'Christmas-themed display' | 'display' | 'Store Items' 2. 'Store window' | 'window' | 'Building Parts' 3. 'Figurines' | 'figurines' | 'Decorative Items' 4. 'Several mannequins' | 'mannequins' | 'Store Items' 5. 'Mannequins dressed in Victorian-style clothing' | 'mannequins' | 'Store Items' ...	1. 'Blonde labrador retriever' | 'labrador retriever' | 'Dog breeds' 2. 'Snow' | 'Snow' | 'Weather conditions' 3. 'Dog's head' | 'Head' | 'Body parts'

Figure 4. The illustration of quality issues existing in image-text pair data. **Row 1**: Existing image-text pair dataset typically suffer from significant partial annotation and image-text misalignment problems. **Row 2**: Limited by capabilities, traditional NLP parsers [1, 19] extract nouns do not correspond to actual object in the images. **Row 3**: Our data pipeline provides refined captions with highly detailed image descriptions, preserving effective visual concepts from the original captions while supplementing missing concepts. **Row 4**: Our data pipeline provides rich, multi-granularity object entity information.

from VLLM. Specifically, we utilize the fields *'phrase'* and *'category'* as the textual inputs for the detector and employ a predefined score threshold to filter the resulting bounding boxes. If either of the two fields are matched, we assign the entire entity {*phrase, category, parent category*} for that object. After filtering with a predefined confidence threshold, approximately 50M data are sampled for subsequent training, which we refer to as **GranuCap50M**. For training the detector, we use the fields of *'phrase'* and *'category'* as the textual label; while for object captioner, we concatenate the three fields – *'phrase* | *category* | *parent category'* – to serve as the object's ground truth description.

None concept corpus. Similar to DetCLIP [54], we develop a noun concept corpus using the information of extracted object entity. This corpus is mainly designed to provide negative concepts for grounding and image-text pair data (Sec. 3.1). Specifically, we collect the *'category'* field of entities from the 200M recaptioned data. Post frequency analysis, concepts with a total frequency below 10 are omitted. The resulting noun concept corpus of DetCLIPv3 consists of 792k noun concepts, expanding nearly 57 times beyond the 14k concepts built in DetCLIP.

3.3. Multi-stage Training Scheme

Learning to generate diverse object descriptions requires extensive training on large-scale datasets. However, dense prediction tasks such as object detection demand high-resolution inputs to effectively handle scale variance across different objects. This substantially raises the computational cost, posing a challenge to scaling up the training. To mitigate this issue, we develop training strategy based on 'pretraining+finetuning' paradigm to optimize training costs, Specifically, it consists of 3 steps:

1. **Training the OV detector (stage 1)**: In the initial phase, we train the OV detector with annotated datasets, *i.e.*, Objects365 [46], V3Det[51] and GoldG [22]. To prepare the model for learning from lower-resolution inputs in later training stages, we apply large-scale jittering augmentation to the training data. Additionally, the model with Swin-L backbone developed during this phase is utilized to generate pseudo bounding box for image-text pairs, as described in Sec. 3.2.
2. **Pretraining the object captioner (stage 2)**: To enable the object captioner to generate diverse object descriptions, we conduct its pretraining using GranuCap50M. To boost the efficiency of this training phase, we freeze all parameters of the OV detector, including the backbone, pixel encoder, and object decoder, and adopt a lower input resolution of 320×320. This strategy facilitates the captioner to efficiently acquire the visual concept knowledge from large-scale image-text pairs.
3. **Holistic finetuning (stage 3)**: This phase aims to adapt the captioner for high-resolution input while simultaneously improving the OV detector. Specifically, we sample 600k samples from GranuCap50M with balanced concepts. These samples, along with detection and grounding datasets, are utilized to further fine-tune the model. All parameters are released during this phase to maximize the effectiveness, with the training objective set to the combination of detection and captioning losses, *i.e.*, $\mathcal{L} = \mathcal{L}_{det} + \mathcal{L}_{lm}$. The supervision for the captioner solely comes from dataset constructed using our auto-annotation pipeline, whereas all data contribute to the training of the OV detector. Since both the detec-

	Method	Backbone	Pre-training data	LVIS$^{\text{minival}}$				LVIS$^{\text{val}}$			
				AP_{all}	AP_r	AP_c	AP_f	AP_{all}	AP_r	AP_c	AP_f
1	GLIP [27]	Swin-T	O365,GoldG,Cap4M	26.0	20.8	21.4	31.0	17.2	10.1	12.5	25.2
2	GLIPv2 [60]	Swin-T	O365,GoldG,Cap4M	29.0	–	–	–	–	–	–	–
3	CapDet [36]	Swin-T	O365,VG	33.8	29.6	32.8	35.5	–	–	–	–
4	GroundingDINO [34]	Swin-T	O365,GoldG,Cap4M	27.4	18.1	23.3	32.7	–	–	–	–
5	OWL-ST [39]	CLIP B/16	WebLI2B	34.4	38.3	–	–	28.6	30.3	–	–
6	DetCLIP [54]	Swin-T	O365,GoldG,YFCC1M	35.9	33.2	35.7	36.4	28.4	25.0	27.0	28.4
7	DetCLIPv2 [55]	Swin-T	O365,GoldG,CC15M	40.4	36.0	41.7	40.4	32.8	31.0	31.7	34.8
8	DetCLIPv3	Swin-T	O365,V3Det,GoldG,GranuCap50M	**47.0**	**45.1**	**47.7**	**46.7**	**38.9**	**37.2**	**37.5**	**41.2**
9	GLIP [27]	Swin-L	FourODs,GoldG,Cap24M	37.3	28.2	34.3	41.5	26.9	17.1	23.3	36.4
10	GLIPv2 [60]	Swin-H	FiveODs,GoldG,CC15M,SBU	50.1	–	–	–	–	–	–	–
11	GroundingDINO [34]	Swin-L	O365,OI,GoldG,Cap4M,COCO,RefC	33.9	22.2	30.7	38.8	–	–	–	–
12	OWL-ST [39]	CLIP L/14	WebLI2B	40.9	41.5	–	–	35.2	36.2	–	–
13	DetCLIP [54]	Swin-L	O365,GoldG,YFCC1M	38.6	36.0	38.3	39.3	28.4	25.0	27.0	31.6
14	DetCLIPv2 [55]	Swin-L	O365,GoldG,CC15M	44.7	43.1	46.3	43.7	36.6	33.3	36.2	38.5
15	DetCLIPv3	Swin-L	O365,V3Det,GoldG,GranuCap50M	**48.8**	**49.9**	**49.7**	**47.8**	**41.4**	**41.4**	**40.5**	**42.3**

Table 1. Zero-shot *fixed* AP [8] on LVIS val [16] and minival [22]. Gray numbers indicate including COCO [31] into training, which shares the identical image set with LVIS, thus not representing truly zero-shot. DetCLIPv3 achieves state-of-the-art performance.

tor and captioner have been pretrained, the model can be efficiently adapted in a few epochs.

4. Experiments

Training detail. We train 2 models with Swin-T and Swin-L [35] backbones. The training settings for the object detector primarily follows DetCLIPv2 [55]. We use 32/64 V100 GPUs to train swin-T/L-based models, respectively. The training epochs for the three phases are 12, 3, and 5, respectively. For the model with the Swin-T backbone, the respective training times for these stages amount to 54, 56, and 35 hours. Refer to Appendix for additional training details.

4.1. Zero-Shot Open-Vocabulary Object Detection

Following previous works [27, 39, 54, 55, 60], we evaluate our model's open-vocabulary capability with the zero-shot performance on the 1203-class LVIS [16] dataset. We report performance of *fixed* AP [8] on both val (LVIS$^{\text{val}}$) and minival [22] (LVIS$^{\text{minival}}$) splits. In this experiment, we only use the OV detector component of the model, with class names of the datasets serving as the input.

Table 1 presents a comparison of our method with existing approaches. DetCLIPv3 significantly outperforms its counterparts, demonstrating superior open-vocabulary object detection capabilities. For instance, on LVIS minival, our models with Swin-T (row 8) and Swin-L (row 15) backbones achieve an AP of 47.0 and 48.8, respectively, improving upon the previous state-of-the-art method, DetCLIPv2, by 6.6 (row 7) and 4.1 AP (row 14). Notably, the performance of our Swin-L model on the rare category (49.9 AP) even surpasses that on the base category (47.8 AP in frequent and 49.7 AP in common). This indicates that comprehensive pretraining with high-quality image-text pairs substantially broadens the model's capacity to recognize various visual concepts, leading to significant improved detection capabilities on long-tail distributed data.

4.2. Evaluation of Object Captioner

We adopt 2 tasks for evaluating our object captioner, *i.e.*, zero-shot generative object detection and dense captioning.

Zero-shot generative object detection. We conduct zero-shot object-level label generation on COCO [31] dataset with the inference process described in Sec. 3.1 and evaluate its detection performance. However, this evaluation poses significant challenges due to two key factors: (1) the absence of predefined categories for foreground selection results in discrepancies between the detector's proposed foreground regions and dataset's object patterns. (2) the generation results can be any arbitrary vocabulary, which may not align with the class names specified in the dataset. To mitigate these issues, we introduce several post-processing techniques. Specifically, we use the 'category' field from the generated labels as the object's class. To address issue (2), during evaluation, we compute similarity between the generated categories and COCO's class names using the text encoder of the evaluated model, substituting the generated object category with the best-matched COCO category. To resolve issue (1), we further filter out objects with a similarity score below a predefined threshold of 0.7.

To compare with existing methods, we adopt the OV COCO setting proposed in OVR-CNN [59], where 48 classes from COCO are selected as base classes and 17 as novel classes. The evaluation metric used is the mAP at an IoU of 0.5. Contrary to previous methods, *we perform zero-shot generative OV detection across all settings without conducting training on the base categories*. Table 2 presents

Method	Generative	Novel AP_{50}	Base AP_{50}	Overall AP_{50}
OVR-CNN [59]	✗	22.8	46.0	39.9
Detic [65]	✗	27.8	47.1	42.0
RegionCLIP [64]	✗	26.8	54.8	47.5
ViLD [15]	✗	27.6	59.5	51.3
VLDet [30]	✗	32.0	50.6	45.8
DetPro [10]	✗	43.3	61.9	55.7
DetCLIPv3 (Swin-T)	✔	54.7	42.8	46.9
DetCLIPv3 (Swin-L)	✔	**57.3**	44.2	49.3

Table 2. Zero-shot generative object detection on COCO [31]. Gray numbers indicate training on COCO's base classes. Our method significantly outperforms previous OV detectors in novel categories in a generative manner.

	Method	VG V1.2 AP	VG-COCO AP
1	JIVC [21]	10.0	7.9
2	COCG [28]	10.4	8.9
3	CAG-Net [56]	10.5	–
4	TDC+ROCSU [47]	11.9	11.6
5	CapDet [36]	15.4	14.0
6	GRiT [52]	15.5	–
7	DetCLIPv3 (Swin-T)	18.4	17.7
8	DetCLIPv3 (Swin-L)	**19.7**	**18.9**

Table 3. Dense captioning on VG V1.2 [23] and VG-COCO [47].

the evaluation results. Our generative method can significantly outperform previous discriminative approaches in novel class performance. And our overall AP achieves a level comparable to previous methods without training on base classes. These results demonstrate the potential of generative-based OV detection as a promising paradigm.

Dense captioning. Leveraging visual concept knowledge acquired from extensive image-text pairs, DetCLIPv3 can be easily adapted to generate detailed object descriptions. Following [21, 47], we evaluate the dense captioning performance on the VG V1.2 [23] and VG-COCO [47] datasets. To ensure a fair comparison, we finetune our model on the training dataset. Similar to CapDet [36], during fine-tuning, we convert our OV detector into a class-agnostic foreground extractor, which is achieved by assigning the textual label of all foreground objects to the concept *'object'*. Table 3 compares our method with the existing methods. DetCLIPv3 significantly outperforms existing approaches. *E.g.*, on VG, our models with Swin-T (row 7) and Swin-L (row 8) backbones surpass the previous best method, GRiT [52] (row 6), by 2.9 AP and 4.2 AP, respectively.

4.3. Robustness to Distribution Shift

A robust OV object detector should be capable of recognizing a broad spectrum of visual concepts across various domains. The recent vision-language model CLIP [42] show-

Method	Backbone	COCO AP	COCO-O AP	Effective Robustness
GLIP [27]	Swin-T	46.1	29	+8.0
DetCLIPv3	Swin-T	**47.2**	**38.5**	**+17.3**
DINO [61]	Swin-L	58.5	42.1	+15.8
DyHead [7]	Swin-L	56.2	35.3	+10.0
GLIP [27]	Swin-L	51.4	48	+24.9
GRiT [52]	ViT-H	60.4	42.9	+15.7
DetCLIPv3	Swin-L	48.5	**48.8**	**+27.0**

Table 4. Distribution shift performance on COCO-O [37]. Gray numbers indicate include COCO [31] data into training.

Method	Backbone	$LVIS^{mini}_{base}$ AP_{all}	$LVIS^{mini}_{base}$ AP_{rare}	$LVIS^{mini}_{all}$ AP_{all}	$LVIS^{mini}_{all}$ AP_{rare}	ODinW13 AP
GLIP [27]	Swin-T	–	–	–	–	64.9
GLIPv2 [60]	Swin-T	–	–	50.6	–	66.5
DetCLIPv2 [55]	Swin-T	–	–	50.7	44.3	68.0
OWL-ST+FT[39]	CLIP B/16	48.7	42.1	–	–	–
DetCLIPv3	Swin-T	**54.3**	**53.7**	**56.5**	**55.1**	**71.1**
GLIP [27]	Swin-L	–	–	–	–	68.9
GLIPv2 [60]	Swin-H	–	–	59.8	–	70.4
DetCLIPv2[55]	Swin-L	–	–	60.1	58.3	70.4
OWL-ST+FT[39]	CLIP L/14	56.2	52.3	–	–	–
DetCLIPv3	Swin-L	**60.5**	**60.3**	**60.5**	**60.7**	**72.1**

Table 5. Fine-tuning performance. *Fixed* AP [8] on LVIS minival5k [22] and average AP on ODinW13 [27] are reported.

cases remarkable generalization to domain shifts in ImageNet variants [17, 18, 50] through learning from extensive image-text pairs. Similarly, we expect observing comparable phenomena in OV detection. To this end, we use COCO-O [37] to study our model's robustness to distribution shifts. Table 4 compares our method with several leading closed-set detectors and the open-set detector GLIP on both COCO and COCO-O. As COCO is not incorporated into our training, DetCLIPv3's performance trails behind those detectors that are specifically trained on it. However, our model significantly outperforms these detectors on COCO-O. *E.g.*, our Swin-L model achieves 48.8 AP on COCO-O, which even surpasses its COCO performance (48.5 AP) and attains the best effective robustness score of +27.0. Refer Appendix for qualitative visualizations.

4.4. Transfer Results with Fine-tuning

Table 5 explores the transferability of DetCLIPv3 by fine-tuning it on downstream datasets, *i.e.* LVIS minival [22] and ODinW [27]. For LVIS, two settings are considered: (1) $LVIS^{mini}_{base}$: only base (common and frequent) classes are used for training, as per [39]; and (2) $LVIS^{mini}_{all}$: entails training with all categories.

DetCLIPv3 consistently outperforms its counterparts across all settings. On ODinW13, the Swin-T based DetCLIPv3 (71.1 AP) even surpasses the Swin-L based DetCLIPv2 (70.4 AP). On LVIS, DetCLIPv3 demonstrates exceptional performance, *e.g.*, the Swin-L based model reaches 60.5 AP on both $LVIS^{mini}_{base}$ and $LVIS^{mini}_{all}$, surpass-

		LVIS^{mini}		COCO-O	VG V1.2
		AP_{all}	AP_{rare}	AP	AP
1	Baseline	30.8	28.7	24.1	–
2	+GoldG	41.4	37.5	32.5	–
3	+V3Det	42.5	39.4	30.7	–
4	+GranuCap600k	45.3	42.2	36.4	–
5	+Captioner	46.6	44.5	38.0	17.1
6	+Stage2 pretraining	**47.0**	**45.1**	**38.5**	**18.4**

Table 6. The evolution roadmap of DetCLIPv3. Each row introduces new changes building upon the results of the preceding row.

	Threshold	#Samples	LVIS^{mini}	
			AP_{all}	AP_{rare}
1	0.15	600k	45.0	**43.2**
2	0.2	600k	**45.3**	42.2
3	0.25	600k	44.8	42.3
4	0.3	600k	44.9	41.0
5	0.2	300k	44.0	40.1
6	0.2	600k	45.3	42.2
7	0.2	1200k	**46.1**	**44.2**

Table 7. Impact of filtering threshold and data volume of pseudo-labeled image-text pairs.

ing OWL-ST+FT [39] (56.2 AP on $\text{LVIS}^{\text{mini}}_{\text{base}}$) that pretrains with 2B pseudo-labeled data by a large margin. This indicates the high-quality image-text pairs construted by our auto-annotation pipeline effectively boosts the learning efficiency. Besides, we observe a conclusion parallel to that in [39]: with strong pretraining, even fine-tuning solely on base categories can substantially enhance performance of rare categories. This is exemplified by the Swin-L model's improvement from 49.8 AP_{rare} in row 15 of Table 1 to 60.3 AP_{rare} of Table 5.

4.5. Ablation Study

DetCLIPv3's evolution roadmap. Table 6 investigates the development roadmap of DetCLIPv3, from the baseline model to the final version. Our experiments utilize a model with a Swin-T backbone. For the OV detector, we evaluate the AP on LVIS minival (Sec. 4.1) and COCO-O (Sec. 4.3), and for the captioner, we report the fine-tuned performance on VG (Sec. 4.2). Our baseline (row 1) model is our OV detector (as described in Sec. 3.1) without the object captioner and is trained solely on the Objects365 [46]. This model exhibits limited capabilities, achieving a modest 30.8 AP on LVIS. Subsequently, we introduce a series of effective designs: **(1) Incorporating more human-annotated data** (rows 2 and 3), *i.e.*, GoldG [22] and V3Det [51], significantly boosts the LVIS AP to 42.5. **(2) Introducing image-text pair data**, *i.e.*, 600k samples from GranuCap50M (also the training data used in our stage 3 training, see Sec. 3.3), effectively further improve the LVIS AP to 45.3. More importantly, it significantly improve the model's domain generalization, bring COCO-O AP from 30.7 in row3 to 36.4 in row 4. **(3)** Row 5 further **integrates the object captioner**, yet without the stage 2 pretraining. It boosts the LVIS AP to 46.6, despite no new data is introduced. This improvement reveals the learning of captioner benefits OV detection – learning to generate diverse labels for objects encourages the object decoder to extract more discriminative object features. **(4) Integrating stage 2 captioner pretraining** efficiently acquires broad visual concept knowledge from the massive image-text pairs of GranuCap50M. This design significantly enhances the generative capability of the captioner, boosting the VG AP from 17.1 of row 5 to 18.4 of row 6. Furthermore, it modestly improve the OV detection performance from 46.6 AP to 47.0 AP on LVIS.

Pseudo-labeling for Image-text Pairs. Table 7 investigates two critical factors in utilizing pseudo-labeled image-text pairs: the filtering threshold and the data volume. We experiment with Swin-T model for stage 1 training, with pseudo-label data incorporated. A filtering threshold of 0.2 achieved the best results, and increasing the volume of data continuously improved the performance of the OV detection. Though incorporating 1200k data achieving better results, we opt for 600k data for stage 3 training for efficiency consideration. Notably, when assisted with the captioner's learning in generative tasks, the effectiveness of 600k data samples (row 5 of Table 6, 46.6 AP) surpasses the result of 1200k samples without the captioner (46.1 AP).

4.6. Visualization

Figure 1 provides visualization results for both OV detection and object label generation of DetCLIPv3. Our model demonstrates superior visual understanding capabilities, capable of detecting or generating a broad range of visual concepts. Refer to Appendix for more visualization results.

5. Limitation and Conclusion

Limitation. The evaluation of DetCLIPv3's generative capability remains incomplete, as existing benchmarks fall short in effectively evaluating generative detection results. Moreover, the detection process in DetCLIPv3 currently does not support control via instructions. Moving forward, an important research direction will be to develop comprehensive metrics for evaluating generative open-vocabulary detectors and to integrate large language models (LLMs) for instruction-controled open-vocabulary detection.

Conclusion. In this paper, we present DetCLIPv3, an innovative OV detector that is capable of localizing objects based on category names, as well as generating object label that is hierarchical and multi-granular. Such enhanced visual capability enables more comprehensive fine-grained visual understanding, which expands the application scenarios for OVD model. We hope our method provides insights for future development of visual cognition systems.

References

[1] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc., 2009. 4, 5

[2] Zhaowei Cai, Gukyeong Kwon, Avinash Ravichandran, Erhan Bas, Zhuowen Tu, Rahul Bhotika, and Stefano Soatto. X-detr: A versatile architecture for instance-wise vision-language tasks. In *ECCV*, 2022. 1

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 3

[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 *CVPR*, 2021. 3, 4

[5] 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. 2

[6] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023. 2, 4

[7] Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. Dynamic head: Unifying object detection heads with attentions. In *CVPR*, 2021. 7

[8] Achal Dave, Piotr Dollár, Deva Ramanan, Alexander Kirillov, and Ross Girshick. Evaluating large-vocabulary object detectors: The devil is in the details. *arXiv preprint arXiv:2102.01066*, 2021. 2, 6, 7

[9] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. In *NeurIPS*, 2019. 4

[10] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *CVPR*, 2022. 7

[11] Dario Fontanel, Matteo Tarantino, Fabio Cermelli, and Barbara Caputo. Detecting the unknown in object detection. *arXiv preprint arXiv:2208.11641*, 2022. 1

[12] Mingfei Gao, Chen Xing, Juan Carlos Niebles, Junnan Li, Ran Xu, Wenhao Liu, and Caiming Xiong. Towards open vocabulary object detection without human-provided bounding boxes. *arXiv preprint arXiv:2111.09452*, 2021. 1

[13] Mingfei Gao, Chen Xing, Juan Carlos Niebles, Junnan Li, Ran Xu, Wenhao Liu, and Caiming Xiong. Open vocabulary object detection with pseudo bounding-box labels. In *ECCV*, 2022. 2

[14] 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*, 2, 2021. 2

[15] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022. 2, 7

[16] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 6

[17] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, 2021. 7

[18] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, 2021. 7

[19] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spacy: Industrial-strength natural language processing in python. 2020. 4, 5

[20] Matthew Inkawhich, Nathan Inkawhich, Hai Li, and Yiran Chen. Self-trained proposal networks for the open world. *arXiv preprint arXiv:2208.11050*, 2022. 1

[21] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *CVPR*, 2016. 2, 7

[22] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *ICCV*, 2021. 2, 3, 4, 5, 6, 7, 8

[23] 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. *IJCV*, 2017. 2, 7

[24] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 2

[25] 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. 2, 3, 4

[26] 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. 2

[27] 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 *CVPR*, 2022. 1, 2, 6, 7

[28] Xiangyang Li, Shuqiang Jiang, and Jungong Han. Learning object context for dense captioning. In *AAAI*, 2019. 2, 7

[29] Chuang Lin, Peize Sun, Yi Jiang, Ping Luo, Lizhen Qu, Gholamreza Haffari, Zehuan Yuan, and Jianfei Cai. Learning object-language alignments for open-vocabulary object detection. *arXiv preprint arXiv:2211.14843*, 2022. 2

[30] Chuang Lin, Peize Sun, Yi Jiang, Ping Luo, Lizhen Qu, Gholamreza Haffari, Zehuan Yuan, and Jianfei Cai. Learning object-language alignments for open-vocabulary object detection. In *ICLR*, 2023. 4, 7

[31] 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 *ECCV*, 2014. 6, 7

[32] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 3

[33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 2, 4

[34] 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. 1, 2, 3, 6

[35] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 6

[36] Yanxin Long, Youpeng Wen, Jianhua Han, Hang Xu, Pengzhen Ren, Wei Zhang, Shen Zhao, and Xiaodan Liang. Capdet: Unifying dense captioning and open-world detection pretraining. In *CVPR*, 2023. 2, 6, 7

[37] Xiaofeng Mao, Yuefeng Chen, Yao Zhu, Da Chen, Hang Su, Rong Zhang, and Hui Xue. Coco-o: A benchmark for object detectors under natural distribution shifts. In *ICCV*, 2023. 7

[38] George A. Miller. Wordnet, an electronic lexical database. 1998. 4

[39] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. *arXiv preprint arXiv:2306.09683*, 2023. 2, 4, 6, 7, 8

[40] OpenAI. Improving image generation with better captions. *https://openai.com/dall-e-3*, 2023. 2

[41] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 4

[42] 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 *ICML*, 2021. 2, 4, 7

[43] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, 2019. 3

[44] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 3, 4

[45] 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. In *NeurIPS*, 2022. 4

[46] 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 *ICCV*, 2019. 3, 5, 8

[47] Zhuang Shao, Jungong Han, Demetris Marnerides, and Kurt Debattista. Region-object relation-aware dense captioning via transformer. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 2, 7

[48] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 3, 4

[49] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 2016. 3, 4

[50] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, 2019. 7

[51] Jiaqi Wang, Pan Zhang, Tao Chu, Yuhang Cao, Yujie Zhou, Tong Wu, Bin Wang, Conghui He, and Dahua Lin. V3det: Vast vocabulary visual detection dataset. In *ICCV*, 2023. 3, 5, 8

[52] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding. *arXiv preprint arXiv:2212.00280*, 2022. 2, 7

[53] Johnathan Xie and Shuai Zheng. Zsd-yolo: Zero-shot yolo detection using vision-language knowledge distillation. *arXiv preprint arXiv:2109.12066*, 2021. 2

[54] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. In *NeurIPS*, 2022. 1, 2, 3, 5, 6

[55] Lewei Yao, Jianhua Han, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, and Hang Xu. Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignmen. In *CVPR*, 2023. 1, 2, 3, 4, 6, 7

[56] Guojun Yin, Lu Sheng, Bin Liu, Nenghai Yu, Xiaogang Wang, and Jing Shao. Context and attribute grounded dense captioning. In *CVPR*, 2019. 2, 7

[57] Qiying Yu, Quan Sun, Xiaosong Zhang, Yufeng Cui, Fan Zhang, Xinlong Wang, and Jingjing Liu. Capsfusion: Rethinking image-text data at scale. *arXiv preprint arXiv:2310.20550*, 2023. 2

[58] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. *arXiv preprint arXiv:2203.11876*, 2022. 2

[59] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021. 6, 7

[60] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. In *NeurIPS*, 2022. 2, 6, 7

[61] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr

with improved denoising anchor boxes for end-to-end object detection. In *ICLR*, 2023. 3, 7

[62] Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianfeng Gao, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. *arXiv preprint arXiv:2303.08131*, 2023. 1

[63] Shiyu Zhao, Zhixing Zhang, Samuel Schulter, Long Zhao, BG Vijay Kumar, Anastasis Stathopoulos, Manmohan Chandraker, and Dimitris N Metaxas. Exploiting unlabeled data with vision and language models for object detection. In *ECCV*, 2022. 2

[64] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. Regionclip: Region-based language-image pretraining. In *CVPR*, 2022. 2, 7

[65] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 2, 7

[66] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. 3