Appendix for DetCLIPv3: Towards Versatile Generative Open-vocabulary Object Detection

A. Additional Implementation Details

Training. The training of DetCLIPv3 involves data from various sources. Table 1 summarizes detailed data information used in different training phases. Since the training process varies for different data type, (*e.g.*, the object captioner accepts only image-text pair data as input), we design each iteration's global batch to contain only one type of data.

For the training of the the open-vocabulary detector, following previous DetCLIP works [21, 23], we initialize the text encoder with the parameters of FILIP's [22] language model, and reduce its learning rate by 0.1 during training to preserve the knowledge obtained via FILIP's pre-training. To improve training efficiency, we set the maximum text token length for the text encoder to 16.

For the training of the object captioner, we initialize the captioner with the pre-trained weights of Qformer [8], whereas the deformable [26] cross-attention layers are randomly initialized. To preserve the knowledge acquired during Qformer [8] pretraining, the object captioner utilizes the same BERT [5] tokenizer for processing text input, different to the text encoder which employs the CLIP [15] tokenizer. The maximum text token length for the object captioner is set to 32.

In each training stage, to conserve GPU memory, automatic mixed-precision [13] and gradient checkpointing [2] are employed. Table 2 summarizes the detailed training settings for each training stage.

Training Stage	Dataset	Total Volume
Stage 1	O365 [17](0.66M), GoldG [9](0.77M), V3Det [20](0.18M)	1.61M
Stage 2	GranuCap50M	50M
Stage 3	O365 [17](0.66M), GoldG [9](0.77M), V3Det [20](0.18M), GranuCap600K (0.6M)	2.21M

Table 1. Dataset information for each training stage of DetCLIPv3. O365 refers to the Objects365 v2, from which we sample 0.66M data with balanced class for training, similar to previous DetCLIP v1/v2 [21, 23] works. GranuCap50M is developed with our proposed auto-annotation pipeline, using 50M image-text pairs sampled from a collection of CC3M [18], CC12M [1], YFCC100M [19] and LAION400M [16].

Inference. Inference process of DetCLIPv3's OV detector follows DINO [25], where the results for each image are derived from the predictions of 300 object queries with highest confidence scores. For the *fixed* AP [4] evaluation on the LVIS [6] dataset, it is required that each category within the entire validation set has at least 10,000 predictions. To ensure an sufficient number of predictions per image, we adopt an inference process similar to that of GLIP [9]. Specifically, during inference for each data sample, the 1203 categories are split into 31 chunks, with a chunk size of 40 categories. We conduct inference separately for each chunk and retain the top 300 predictions based on their confidence scores.

For the inference process of DetCLIPv3's object captioner, as described in the main paper, for each image, we utilize the most frequent 15k concepts from our developed noun concept corpus as text queries to extract top 100 foreground regions with highest similarity. After the generation of descriptive labels for these regions by the object captioner, their confidence scores are re-calibrated using the OV detector. A class-agnostic non-maximum suppression (NMS) operation is then performed for regions with re-calibrated scores higher than 0.05, the results of which are output as predictions. We set beam search's beam size equal to 1 for inference of object captioner.

Finetuning. We fine-tune DetCLIPv3 on 2 datasets, *i.e.*, LVIS [6] and ODinW13 [9]. Table 3 and 4 summarize the detailed fine-tuning settings for LVIS and ODinW13, respectively. For LVIS, when fine-tuning with base categories, we exclude novel categories while sampling negative concepts. For ODinW13, similar to DetCLIPv2[23], we employ an auto-decay learning rate schedule. Specifically, when performance reaches a plateau and persists for a tolerance period t_1, we reduce the learning rate by a factor of 0.1. If there is no improvement in performance for a tolerance period t_2, we then terminate the training process.

B. Additional Data Pipeline Details

Figure 1 illustrates an overview of DetCLIPv3's auto-annotation data pipeline.

Prompts. Here we provide the prompts used in each step, including those for the VLLMs as well as for GPT-4.

1. **Recaptioning with VLLM**: We employ Instruct-

Config	Stage1	Stage2	Stage3
GPUs (V100)	32 (Swin-T)/64 (Swin-L)		
training module	OV detector	object captioner	all modules
training objective	$\mathcal{L}_{det} = \mathcal{L}_{align} + \mathcal{L}_{box} + \mathcal{L}_{iou}$	$\mathcal{L}_{lm} = \mathcal{L}_{lm}^{obj} + \mathcal{L}_{lm}^{img}$	$\mathcal{L} = \mathcal{L}_{det} + \mathcal{L}_{lm}$
training epochs	12	3	5
input resolution	$320 \times 240 \sim 1333 \times 800$	320×320	$1333 \times 600 \sim 1333 \times 800$
batch size	128	2048	128
learning rate	2.8e-4	1e-4	1e-4
text encoder lr reduce factor	0.1	–	0.1
numper of concepts (grounding/image-text pairs)	4800	–	4800
optimizer	AdamW [11]		
optimizer momentum	$\beta_1 = 0.9, \beta_2 = 0.999$		
weight decay	0.05		
warmup iters	1000		
learning rate schedule	cosine annealing		
text token length (text encoder)	16		
text tokenizer (text encoder)	CLIP [15] Tokenizer		
text token length (object captioner)	32		
text tokenizer (object captioner)	BERT [5] Tokenizer		

Table 2. Detailed pre-training settings of DetCLIPv3. $\mathcal{L}_{lm}^{obj}$ and $\mathcal{L}_{lm}^{img}$ represent for language modeling training objective for object-level captioning and image-level captioning, respectively.

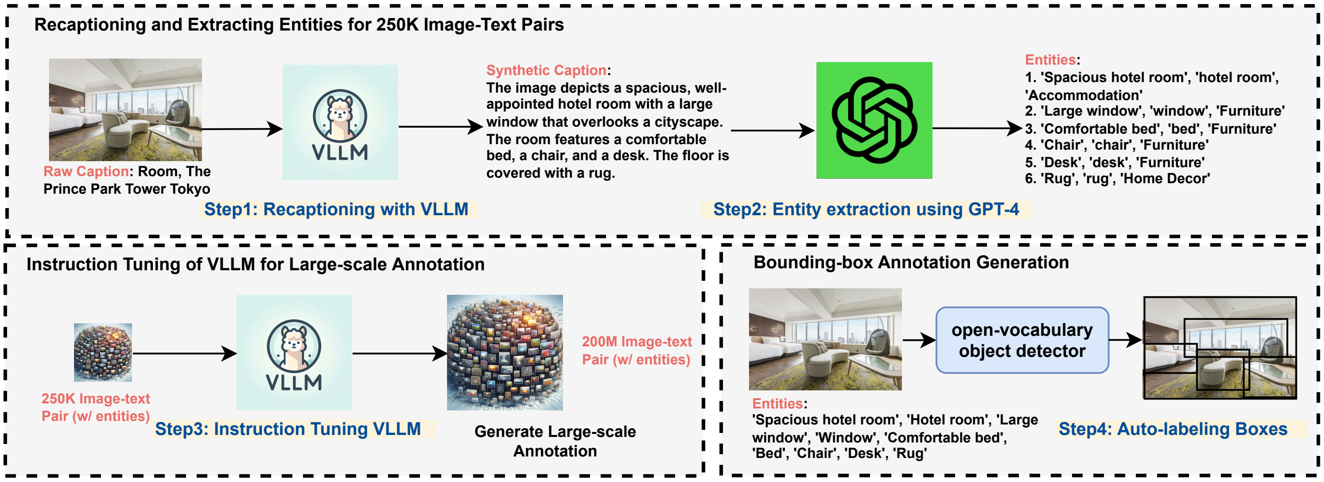

Recaptioning and Extracting Entities for 250K Image-Text Pairs
Raw Caption: Room, The Prince Park Tower Tokyo
VLLM
Synthetic Caption: The image depicts a spacious, well-appointed hotel room with a large window that overlooks a cityscape. The room features a comfortable bed, a chair, and a desk. The floor is covered with a rug.
Entities: 1. 'Spacious hotel room', 'hotel room', 'Accommodation' 2. 'Large window', 'window', 'Furniture' 3. 'Comfortable bed', 'bed', 'Furniture' 4. 'Chair', 'chair', 'Furniture' 5. 'Desk', 'desk', 'Furniture' 6. 'Rug', 'rug', 'Home Decor'
Step1: Recaptioning with VLLM
Step2: Entity extraction using GPT-4
Instruction Tuning of VLLM for Large-scale Annotation
250K Image-text Pair (w/ entities)
VLLM
200M Image-text Pair (w/ entities)
Step3: Instruction Tuning VLLM
Generate Large-scale Annotation
Bounding-box Annotation Generation
open-vocabulary object detector
Entities: 'Spacious hotel room', 'Hotel room', 'Large window', 'Window', 'Comfortable bed', 'Bed', 'Chair', 'Desk', 'Rug'
Step4: Auto-labeling Boxes

Figure 1. The illustration for DetCLIPv3's auto-annotation data pipeline. It initially utilizes a VLLM to recaption 240K image-text pairs, followed by the use of GPT-4 to extract object entities, formatted as {*phrase, category, parent category*}. Subsequently, these image-text pairs are used for instruction tuning of a VLLM, with the trained model providing annotations for a larger-scale of 200M image-text pairs. Finally, an OV detector is employed to provide pseudo-label bounding boxes for these data, and after applying a confidence score filtering, 50M data are sampled to form GranuCap50M.

BLIP [3] to recaption 240K image-text pairs. To leverage information from the original caption texts, we use the following prompt:
"Given a noisy caption of the image: {raw caption}, write a detailed clean description of the image."

2. **Entity extraction using GPT-4**: In this step, we first utilize GPT-4 to filter out non-entity descriptions from the captions generated by the VLLM. The prompt used is:
"Here is a caption for an image: {caption}. Extract the part of factual description related to what is directly observable in the image, while filtering out the parts that refer to inferred contents, description of atmosphere/appearance/style and introduction of history/culture/brand etc. Return solely the result without any other contents. If you think there is no factual description, just return 'None'."
Subsequently, we extract information about object entities from the filtered captions using the prompt:
"You are an AI tasked with developing an open-set object detection dataset from a large number of image captions, without access to the actual images. Your mission

Config	Value
GPUs (V100)	16
training epochs	16
optimizer	AdamW [11]
optimizer momentum	$\beta_1 = 0.9, \beta_2 = 0.999$
lr for image encoder	2e-5
lr for text encoder	2e-6
weight decay	0.05
warmup iters	1000
learning rate schedule	cosine decay
batch size	64
input resolution	1333×800
number of concepts per sample	150
augmentation	multi-scale training, random flip

Table 3. Detailed fine-tuning settings for LVIS [6].

Config	Value
GPUs (V100)	8
maximum training epochs	250
optimizer	AdamW [11]
optimizer momentum	$\beta_1 = 0.9, \beta_2 = 0.999$
lr for image encoder	4e-5
lr for text encoder	4e-7
weight decay	0.05
warmup iters	500
learning rate schedule	auto-step decay
lr decay tolerance t_1 (epochs)	10
training terminate tolerance t_2 (epochs)	15
minimum lr to stop decay	1e-8
batch size	32
input resolution	1333×800
augmentation	multi-scale training, random flip

Table 4. Detailed fine-tuning settings for ODinW13 [9].

is to accurately identify and extract 'objects' from these captions, following the principles below:
1. 'Objects' are physically tangible: They must be concrete entities that can be visually represented in an image. They are NOT (1) abstract concepts (like 'history', 'culture') or feelings (like 'sorrow', 'happiness'), (2) meta-references to the image itself (e.g., 'image', 'picture', 'photo') or the camera (e.g. something is facing the 'camera'), unless they are specifically referring to physical elements within the image. (3) any descriptors (like 'appearance', 'atmosphere', 'color'), (4) events/activities and processes (like 'game', 'presentation', 'performance') and specific event types (like 'country style wedding', 'film festival'), (5) compositional aspects (like 'perspective', 'focus', 'composition') or viewpoint/perspective (like 'bird's eye view').
2. 'Objects' are visually distinct: They are standalone entities that can be visually isolated from their environment. They do not include environmental characteristics (like 'colorful environment') and general location/scene descriptors (e.g., 'scene set indoors', 'country setting', 'sunny day', 'black and white illustration')
Adhere to these guidelines for the extraction process:
1. Consolidate duplicates: If multiple extracted 'objects' refer to the same entity in the caption, merge them into one while retaining conceptual diversity.
2. Categorize the descriptive variants: For 'objects' described with adjectives, provide both versions - with and without the adjective.
3. Identify the broader category: Assign a 'parent category' that each 'object' belongs to.
Present your results as a numbered list in this format: id. 'object with adjective', 'object without adjective', 'parent category'. Your response should consist exclusively of results, with no superfluous content.
Here's the caption: {caption}"

3. **Instruction tuning of VLLM for large-scale annotation**: In this phase, we use the caption texts and object entity information obtained from the above steps to fine-tune the LLaVA [10] model. Here, we combine the aforementioned information into a new concise prompt, and the question-answer pair is constructed as:
   Question:
   "From the noisy caption of the image: {raw caption}, generate a refined image description and identify all visible 'objects' – any visually and physically identifiable entity in the image. Keep the following guidelines in mind:
   1. Merge similar 'objects' from the caption, preserving conceptual diversity.
   2. For adjective-described 'objects', provide versions both with and without the adjective.
   3. Assign a 'parent category' for each 'object'.
   Present results as:
   Caption: {caption}
   Objects: {id. 'object with adjective', 'object without adjective', 'parent category'}.
   <image tokens>"
   Answer:
   Caption: {refined caption}
   Objects: {entity information}
   Here, the VLLM receives image tokens, *i.e.*, <image tokens>, along with their original captions, *i.e.*, {*raw caption*}, as inputs, and learns to generate refined captions and extract information about object entities.

Visualizations. Figure 2-a and 2-b depict refined caption and extracted entity information obtained via our proposed data pipeline. Additionally, Figure 3 displays the bounding box pseudo-labels generated by our Swin-L-based model after stage-1 training.

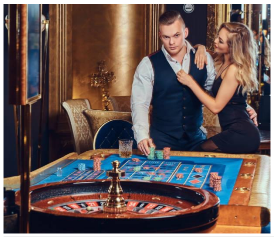
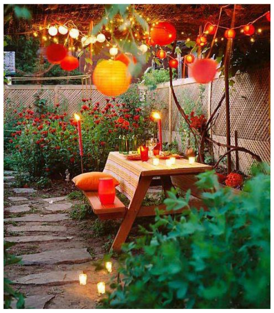
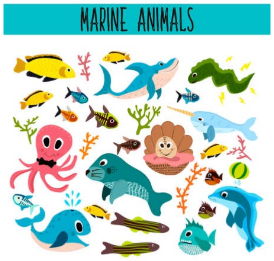
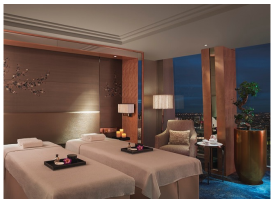
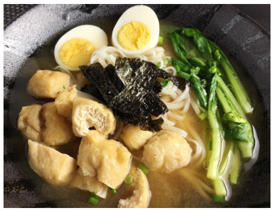
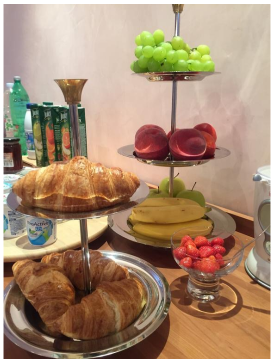
<table>
<tr><th></th><th>Recaption text</th><th>Extracted entities</th></tr>
<tr><td></td><td>A man and a woman are sitting at a casino table, surrounded by chips and cards. The man is wearing a vest, while the woman is dressed in a black dress. They appear to be engrossed in a game of roulette.</td><td>1. 'Man in a vest' | 'Man' | 'People'
2. 'Woman in a black dress' | 'Woman' | 'People'
3. 'Casino table' | 'Table' | 'Furniture'
4. 'Chips' | 'Chips' | 'Gaming Equipment'
5. 'Cards' | 'Cards' | 'Gaming Equipment'
6. 'Vest' | 'Vest' | 'Clothing'
7. 'Black dress' | 'Dress' | 'Clothing'
8. 'Game of roulette' | 'Roulette' | 'Gaming Equipment'</td></tr>
<tr><td></td><td>The image depicts a beautiful outdoor setting, featuring a wooden table and chairs placed in a garden area surrounded by lush greenery. The table is adorned with candles and lanterns. A walkway leads to the table.</td><td>1. 'Beautiful outdoor setting' | 'outdoor setting' | 'Locations'
2. 'Wooden table' | 'table' | 'Furniture'
3. 'Chairs' | 'chairs' | 'Furniture'
4. 'Garden area' | 'garden' | 'Locations'
5. 'Lush greenery' | 'greenery' | 'Nature'
6. 'Candles' | 'candles' | 'Decorations'
7. 'Lanterns' | 'lanterns' | 'Decorations'
8. 'Walkway' | 'walkway' | 'Locations'</td></tr>
<tr><td>
 </td><td>The image features a set of cute sea animals, including sharks, fish, piranha, octopus, manatee, whale, dolphin, narwhal, corals, and algae.</td><td>1. 'Cute sea animals' | 'sea animals' | 'Animals'
2. 'Sharks' | 'sharks' | 'Sea Animals'
3. 'Fish' | 'fish' | 'Sea Animals'
4. 'Piranha' | 'piranha' | 'Sea Animals'
5. 'Octopus' | 'octopus' | 'Sea Animals'
6. 'Manatee' | 'manatee' | 'Sea Animals'
7. 'Whale' | 'whale' | 'Sea Animals'
8. 'Dolphin' | 'dolphin' | 'Sea Animals'
9. 'Narwhal' | 'narwhal' | 'Sea Animals'
10. 'Corals' | 'corals' | 'Sea Animals'
11. 'Algae' | 'algae' | 'Plants'</td></tr>
<tr><td></td><td>The image depicts a luxurious spa room in the Shangri-La Hotel at The Shard, located in London. The room features two massage beds, each with a white cover and pillows. There are also two chairs in the room, one of which is placed next to one of the massage beds. A potted plant can be seen in the corner of the room.</td><td>1. 'Luxurious spa room' | 'Spa room' | 'Rooms'
2. 'Shangri-La Hotel' | 'Hotel' | 'Buildings'
3. 'The Shard' | 'Building' | 'Buildings'
4. 'Two massage beds' | 'Massage beds' | 'Furniture'
5. 'White cover' | 'Cover' | 'Textiles'
6. 'Pillows' | 'Pillows' | 'Textiles'
7. 'Two chairs' | 'Chairs' | 'Furniture'
8. 'Potted plant' | 'Plant' | 'Plants'</td></tr>
<tr><td></td><td>The image features a black bowl of Miso Noodle Soup, which contains tofu puffs, hard-cooked eggs, rice noodles, Asian green vegetables, roasted seaweed, and soup.</td><td>1. 'Black bowl of Miso Noodle Soup' | 'Bowl' | 'Kitchenware'
2. 'Miso Noodle Soup' | 'Soup' | 'Food'
3. 'Tofu puffs' | 'Tofu' | 'Food'
4. 'Hard-cooked eggs' | 'Eggs' | 'Food'
5. 'Rice noodles' | 'Noodles' | 'Food'
6. 'Asian green vegetables' | 'Vegetables' | 'Food'
7. 'Roasted seaweed' | 'Seaweed' | 'Food'</td></tr>
<tr><td></td><td>The image features a dining table with a variety of food items, including croissants, bananas, grapes, and apples. There is also a bowl of fruit on the table. A bottle of water can be seen in the background.</td><td>1. 'Dining table' | 'table' | 'Furniture'
2. 'Variety of food items' | 'food items' | 'Food'
3. 'Croissants' | 'croissants' | 'Food'
4. 'Bananas' | 'bananas' | 'Food'
5. 'Grapes' | 'grapes' | 'Food'
6. 'Apples' | 'apples' | 'Food'
7. 'Bowl of fruit' | 'bowl' | 'Kitchenware'
8. 'Bowl of fruit' | 'fruit' | 'Food'
9. 'Bottle of water' | 'bottle' | 'Kitchenware'
10. 'Bottle of water' | 'water' | 'Beverage'</td></tr>
</table>

Figure 2-a. Examples of refined captions and extracted object entities yield by DetCLIPv3's data pipeline.

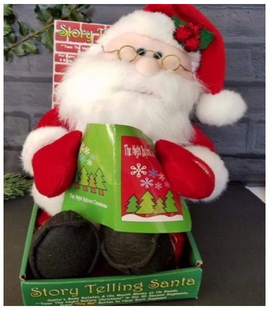
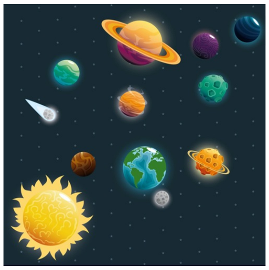
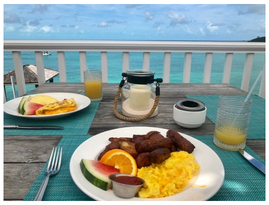
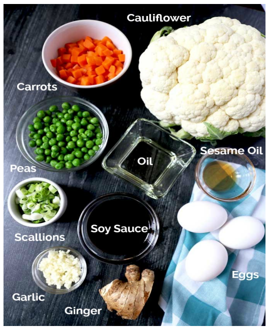
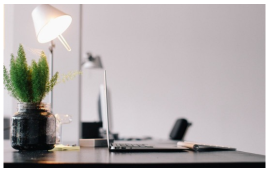
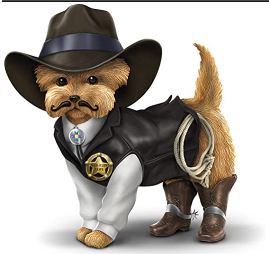
<table>
<tr><th></th><th>Recaption text</th><th>Extracted entities</th></tr>
<tr><td> </td><td>The image features a plush doll of Santa Claus sitting in a box, reading "The Night Before Christmas" book.</td><td>1. 'Plush doll of Santa Claus' | 'Plush doll' | 'Toys'
2. 'Santa Claus' | 'Santa Claus' | 'Fictional Characters'
3. 'Box' | 'Box' | 'Containers'
4. '"The Night Before Christmas" book' | 'Book' | 'Literature'</td></tr>
<tr><td></td><td>The image features a colorful and vibrant illustration of the planets of the solar system, including Earth, Mars, Venus, Jupiter, Saturn, Uranus, Neptune, and Pluto. These planets are arranged in a circular pattern on a dark background.</td><td>1. 'Colorful and vibrant illustration' | 'illustration' | 'Artwork'
2. 'Planets of the solar system' | 'planets' | 'Space Objects'
3. 'Earth' | 'Earth' | 'Planets'
4. 'Mars' | 'Mars' | 'Planets'
5. 'Venus' | 'Venus' | 'Planets'
6. 'Jupiter' | 'Jupiter' | 'Planets'
7. 'Saturn' | 'Saturn' | 'Planets'
8. 'Uranus' | 'Uranus' | 'Planets'
9. 'Neptune' | 'Neptune' | 'Planets'
10. 'Pluto' | 'Pluto' | 'Planets'</td></tr>
<tr><td></td><td>The image features a dining table set up on a balcony overlooking the ocean. The table is adorned with a white plate, which holds a delicious breakfast meal consisting of eggs, bacon, fruit, and orange juice. A fork, knife, and spoon are also present on the table, ready to be used for the meal.</td><td>1. 'Dining table set up on a balcony' | 'Dining table' | 'Furniture'
2. 'Balcony overlooking the ocean' | 'Balcony' | 'Architectural elements'
3. 'White plate' | 'Plate' | 'Tableware'
4. 'Delicious breakfast meal' | 'Meal' | 'Food'
5. 'Eggs' | 'Eggs' | 'Food'
6. 'Bacon' | 'Bacon' | 'Food'
7. 'Fruit' | 'Fruit' | 'Food'
8. 'Orange juice' | 'Juice' | 'Beverages'
9. 'Fork' | 'Fork' | 'Tableware'
10. 'Knife' | 'Knife' | 'Tableware'
11. 'Spoon' | 'Spoon' | 'Tableware'</td></tr>
<tr><td> </td><td>The image features a collection of ingredients for making fried rice, arranged on a black surface. These ingredients include various vegetables such as peas, carrots, cauliflower, and onions, as well as eggs, soy sauce, and sesame seeds.</td><td>1. 'Collection of ingredients' | 'ingredients' | 'Food items'
2. 'Fried rice' | 'rice' | 'Food items'
3. 'Black surface' | 'surface' | 'Household items'
4. 'Various vegetables' | 'vegetables' | 'Food items'
5. 'Peas' | 'peas' | 'Food items'
6. 'Carrots' | 'carrots' | 'Food items'
7. 'Cauliflower' | 'cauliflower' | 'Food items'
8. 'Onions' | 'onions' | 'Food items'
9. 'Eggs' | 'eggs' | 'Food items'
10. 'Soy sauce' | 'sauce' | 'Food items'
11. 'Sesame seeds' | 'seeds' | 'Food items'</td></tr>
<tr><td></td><td>The image depicts a well-organized and spacious desk in a virtual office space. The desk is adorned with various items, including a laptop, a lamp, a potted plant, and a vase. The lamp provides ample lighting for the workspace, while the potted plant adds a touch of greenery and life to the room.</td><td>1. 'Well-organized and spacious desk' | 'desk' | 'furniture'
2. 'Virtual office space' | 'office space' | 'location'
3. 'Laptop' | 'laptop' | 'electronics'
4. 'Lamp' | 'lamp' | 'lighting equipment'
5. 'Potted plant' | 'plant' | 'flora'
6. 'Vase' | 'vase' | 'decorative item'
7. 'Workspace' | 'workspace' | 'location'
8. 'Room' | 'room' | 'location'</td></tr>
<tr><td></td><td>The image features a small dog wearing a cowboy hat, sheriff's badge, and boots.</td><td>1. 'Small dog' | 'Dog' | 'Animals'
2. 'Cowboy hat' | 'Hat' | 'Clothing'
3. 'Sheriff's badge' | 'Badge' | 'Accessories'
4. 'Boots' | 'Boots' | 'Footwear'</td></tr>
</table>

Figure 2-b. Examples of refined captions and extracted object object entities yield by DetCLIPv3's data pipeline.

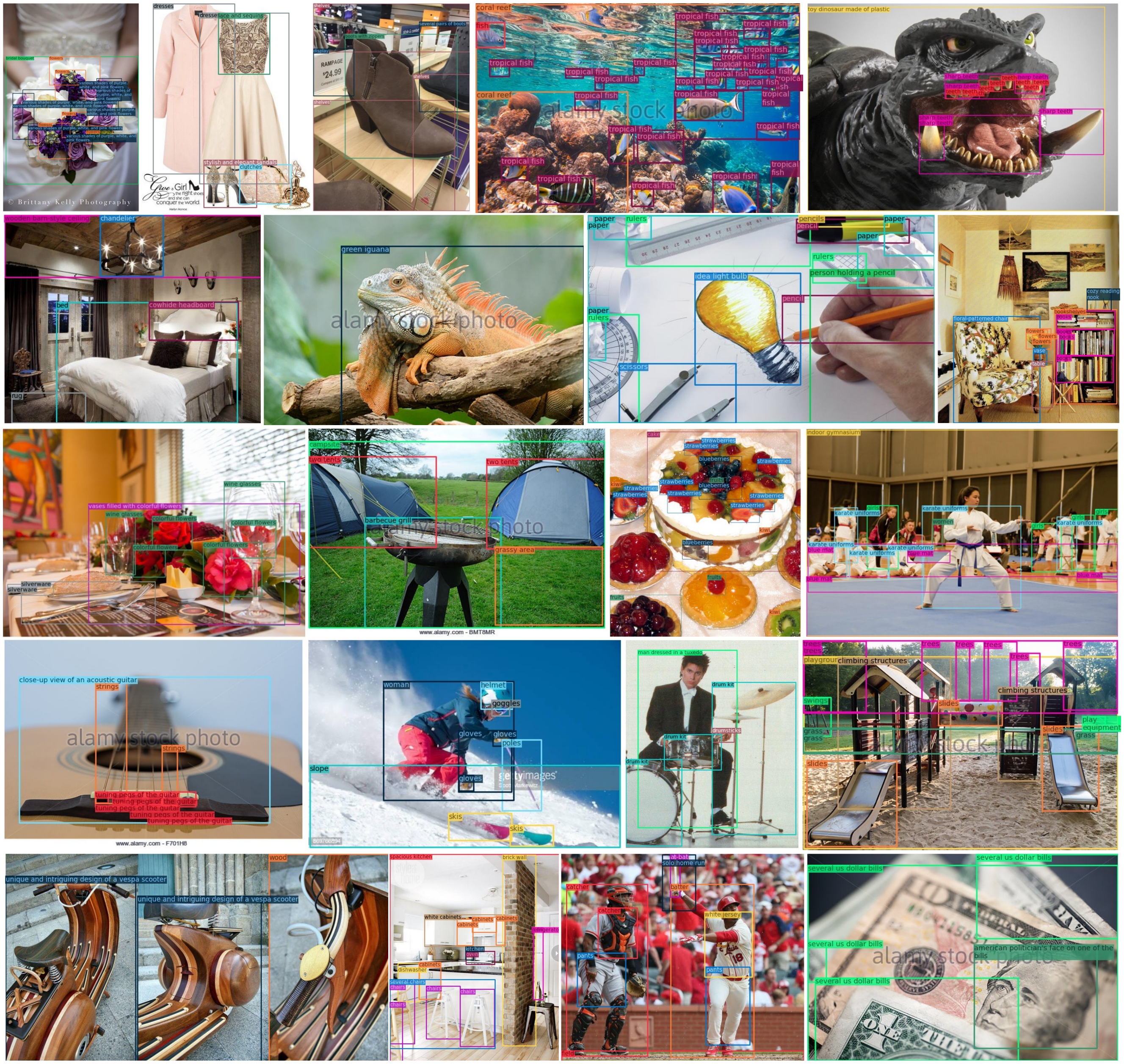


Figure 3. Examples of bounding box pseudo-labels generated by DetCLIPv3's Swin-L model after stage-1 training.

	Method	Backbone	Pre-training data	Fine-tuning data	LVIS$^{\text{minival}}$ AP$_{\text{all}}$	AP$_r$	AP$_c$	AP$_f$	LVIS$^{\text{val}}$ AP$_{\text{all}}$	AP$_r$	AP$_c$	AP$_f$
1	OWL-ST [14]	CLIP B/16	WebLI2B	–	31.8	35.4	–	–	27.0	29.6	–	–
2	OWL-ST [14]	CLIP L/14	WebLI2B	–	38.1	39.0	–	–	33.5	34.9	–	–
3	DetCLIPv3	Swin-T	O365,V3Det,GoldG,GranuCap50M	–	43.7	39.3	44.5	43.7	36.7	34.2	34.9	39.9
4	DetCLIPv3	Swin-L	O365,V3Det,GoldG,GranuCap50M	–	45.8	46.9	45.9	45.5	39.6	38.9	38.4	41.3
5	OWL-ST+FT [14]	CLIP B/16	WebLI2B	LVIS$_{\text{base}}$	47.2	37.8	–	–	41.8	36.2	–	–
6	OWL-ST+FT [14]	CLIP L/14	WebLI2B	LVIS$_{\text{base}}$	54.1	46.1	–	–	49.4	44.6	–	–
7	DetCLIPv3+FT	Swin-T	O365,V3Det,GoldG,GranuCap50M	LVIS$_{\text{base}}$	54.4	46.7	56.1	54.3	48.2	40.2	48.5	51.3
8	DetCLIPv3+FT	Swin-L	O365,V3Det,GoldG,GranuCap50M	LVIS$_{\text{base}}$	60.8	56.7	63.2	59.4	54.1	45.8	55.4	56.4

Table 5. Zero-shot and fine-tuning AP on LVIS val [6] and minival [7]. Results labeled without '+FT' represent zero-shot performance, whereas those with '+FT' indicate results of fine-tuning with LVIS base categories (LVIS$_{\text{base}}$). DetCLIPv3 significantly outperforms OWL-ST [14], which is pre-trained with 2 billion image-text pairs.

Method	Backbone	Sketch	Weather	Cartoon	Painting	Tattoo	Handmake	Average
DetCLIPv3	Swin-T	38.3	43.6	45.0	43.2	29.3	31.5	38.5
DetCLIPv3	Swin-L	50.8	48.6	56.9	53.7	44.5	38.2	48.8

Table 6. Detailed performance on COCO-O [12] dataset. Zero-shot AP is reported.

	Method	Backbone	PascalVOC	AerialDrone	Aquarium	Rabbits	EgoHands	Mushrooms	Packages	Raccoon	Shellfish	Vehicles	Pistols	Pothole	Thermal	**Average**
1	GLIP [9]	Swin-T	62.3	31.2	52.5	70.8	78.7	88.1	75.6	61.4	51.4	65.3	71.2	58.7	76.7	64.9
2	GLIPv2 [24]	Swin-T	66.4	30.2	52.5	74.8	80.0	88.1	74.3	63.7	54.4	63.0	73.0	60.1	83.5	66.5
3	GLIPv2 [24]	Swin-B	71.1	32.6	57.5	73.6	80.0	88.1	74.9	68.2	70.6	71.2	76.5	58.7	79.6	69.4
4	DetCLIPv2 [23]	Swin-T	67.5	41.8	50.8	80.4	79.8	90.1	73.7	70.8	54.8	66.5	77.7	54.8	82.2	68.5
5	DetCLIPv3	Swin-T	72.5	51.6	54.5	79.9	81.2	94.1	78.2	71.6	53.9	67.4	79.4	55.1	84.4	**71.1**
6	GLIP [9]	Swin-L	69.6	32.6	56.6	76.4	79.4	88.1	67.1	69.4	65.8	71.6	75.7	60.3	83.1	68.9
7	GLIPv2 [24]	Swin-H	74.4	36.3	58.7	77.1	79.3	88.1	74.3	73.1	70.0	72.2	72.5	58.3	81.4	70.4
8	DetCLIPv2 [23]	Swin-L	74.4	44.1	54.7	80.9	79.9	90	74.1	69.4	61.2	68.1	80.3	57.1	81.1	70.4
9	DetCLIPv3	Swin-L	76.4	51.2	57.5	79.9	80.2	90.4	75.1	70.9	63.6	69.8	82.7	56.2	83.8	**72.1**

Table 7. Detailed fine-tuned AP on ODinW13 [9] dataset. DetCLIPv3 outperforms its counterparts by a large margin.

C. More Qualitative Results

Figure 4-a, 4-b and 4-c present additional qualitative results showcasing multi-granular object labels generated by DetCLIPv3's object captioner. In the absence of candidate categories, DetCLIPv3's object captioner generates dense, fine-grained, multi-granular object labels, thus facilitating a more comprehensive image understanding.

D. More Experimental Results

More results on LVIS. To comprehensively evaluate the performance of DetCLIPv3, Table 5 provides the standard Average Precision (AP) on LVIS, comparing it with the state-of-the-art method OWL-ST [14], which is pretrained on 2 billion image-text pairs. Specifically, we assess two settings on the LVIS minival [7] and validation [6] datasets: the zero-shot performance and the performance after fine-tuning on LVIS base categories. Despite being pretrained with only 50M image-text pairs, DetCLIPv3 markedly outperforms OWL-ST, *e.g.*, DetCLIPv3 surpasses OWL-ST's counterparts by over 5 AP across all settings, demonstrating the superior learning efficiency of our method. Figure 5 provides the detection results on both zero-shot and LVIS$_{\text{base}}$ fine-tuning settings.

Detailed performance on COCO-O. Table 6 reports the detailed zero-shot AP performance on COCO-O [12]'s 6 domains, *i.e.*, sketch, weather, cartoon, painting, tattoo and handmake. Figure 6-a and 6-b visualizes the detection results, demonstrating DetCLIPv3's robust domain generalization capability.

Detailed performance on ODinW13. Table 7 reports the detailed fine-tuned performance on ODinW13 [9] dataset.

Method	Backbone	OV detector	Object captioner
GLIP [9]	Swin-T	2.5 FPS	–
DetCLIP [21]	Swin-T	2.3 FPS	–
DetCLIPv3	Swin-T	**14.5** FPS	1.2 FPS
GLIP [9]	Swin-L	0.3 FPS	–
DetCLIPv3	Swin-L	**8.2** FPS	0.9 FPS

Table 8. Inference speed. We test the speed with V100 GPU, using batch size=1 and FP16 inference. DetCLIPv3 can run significantly faster than previous methods like GLIP [9] and DetCLIP [21].

Inference speed. Table 8 reports the inference speed of DetCLIPv3 as well as a comparison with previous methods.

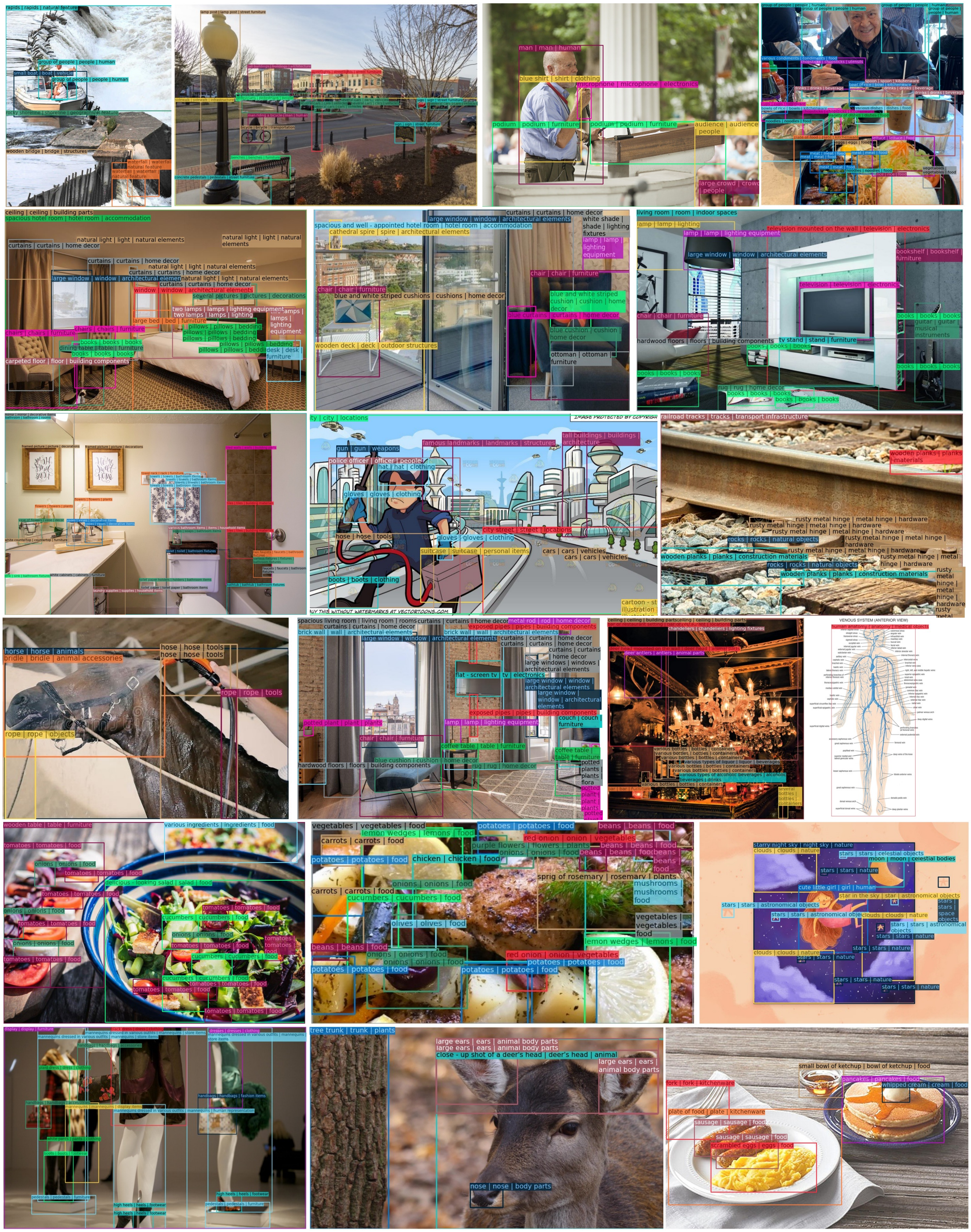

rapids | rapids | natural feature
group of people | people | human
small boat | boat | vehicle
rocky shoreline | shoreline | geographical feature
wooden bridge | bridge | structures
waterfall | waterfall | natural feature
lamp post | lamp post | street furniture
concrete pedestals | pedestals | street furniture
man | man | human
blue shirt | shirt | clothing
microphone | microphone | electronics
podium | podium | furniture
audience | audience | people
large crowd | crowd | people
group of people | people | human
ceiling | ceiling | building parts
spacious hotel room | hotel room | accommodation
natural light | light | natural elements
curtains | curtains | home decor
large window | window | architectural elements
window | window | architectural elements
several pictures | pictures | decorations
two lamps | lamps | lighting equipment
large bed | bed | furniture
chairs | chairs | furniture
pillows | pillows | bedding
books | books | books
dining table | table | furniture
desk | desk | furniture
carpeted floor | floor | building components
spacious and well - appointed hotel room | hotel room | accommodation
cathedral spire | spire | architectural elements
white shade | shade | lighting fixtures
lamp | lamp | lighting equipment
chair | chair | furniture
blue and white striped cushions | cushions | home decor
blue curtains | curtains | home decor
blue cushion | cushion | home decor
ottoman | ottoman | furniture
wooden deck | deck | outdoor structures
living room | room | indoor spaces
lamp | lamp | lighting
lamp | lamp | lighting equipment
television mounted on the wall | television | electronics
large window | window | architectural elements
bookshelf | bookshelf | furniture
television | television | electronics
chair | chair | furniture
books | books | books
guitar | guitar | musical instruments
hardwood floors | floors | building components
tv stand | stand | furniture
rug | rug | home decor
ty | city | locations
IMAGE PROTECTED BY COPYRIGHT
gun | gun | weapons
famous landmarks | landmarks | structures
tall buildings | buildings | architecture
police officer | officer | people
hat | hat | clothing
gloves | gloves | clothing
hose | hose | tools
city street | street | locations
suitcase | suitcase | personal items
cars | cars | vehicles
boots | boots | clothing
BUY THIS WITHOUT WATERMARKS AT VECTORTOONS.COM
cartoon - st illustration
railroad tracks | tracks | transport infrastructure
wooden planks | planks | materials
rusty metal hinge | metal hinge | hardware
rocks | rocks | natural objects
wooden planks | planks | construction materials
horse | horse | animals
bridle | bridle | animal accessories
hose | hose | tools
rope | rope | tools
rope | rope | objects
spacious living room | living room | rooms
curtains | curtains | home decor
metal rod | rod | home decor
exposed pipes | pipes | building components
brick wall | wall | architectural elements
large window | window | architectural elements
flat - screen tv | tv | electronics
potted plant | plant | plants
lamp | lamp | lighting equipment
couch | couch | furniture
chair | chair | furniture
coffee table | table | furniture
blue cushion | cushion | home decor
rug | rug | home decor
hardwood floors | floors | building components
ceiling | ceiling | building parts
chandeliers | chandeliers | lighting fixtures
deer antlers | antlers | animal parts
various bottles | bottles | containers
various types of liquor | liquor | beverages
various types of alcoholic beverages | alcoholic beverages | drinks
VENOUS SYSTEM (ANTERIOR VIEW)
human anatomy | anatomy | medical objects
wooden table | table | furniture
various ingredients | ingredients | food
tomatoes | tomatoes | food
onions | onions | food
delicious - looking salad | salad | food
cucumbers | cucumbers | food
vegetables | vegetables | food
lemon wedges | lemons | food
potatoes | potatoes | food
beans | beans | food
carrots | carrots | food
red onion | onion | vegetables
purple flowers | flowers | plants
chicken | chicken | food
sprig of rosemary | rosemary | plants
mushrooms | mushrooms | food
olives | olives | food
starry night sky | night sky | nature
clouds | clouds | nature
stars | stars | celestial objects
moon | moon | celestial bodies
stars | stars | nature
cute little girl | girl | human
star in the sky | star | astronomical objects
stars | stars | astronomical objects
display | display | furniture
mannequins dressed in various outfits | mannequins | store items
dresses | dresses | clothing
handbags | handbags | fashion items
pedestals | pedestals | furniture
high heels | heels | footwear
tree trunk | trunk | plants
large ears | ears | animal body parts
close - up shot of a deer's head | deer's head | animal
nose | nose | body parts
small bowl of ketchup | bowl of ketchup | food
pancakes | pancakes | food
whipped cream | cream | food
fork | fork | kitchenware
plate of food | plate | kitchenware
sausage | sausage | food
scrambled eggs | eggs | food

Figure 4-a. Qualitative results of multi-granular object labels generated by DetCLIPv3's object captioner. In the absence of candidate categories, DetCLIPv3's object captioner generates dense, fine-grained, multi-granular object labels, thus facilitating a more comprehensive image understanding.

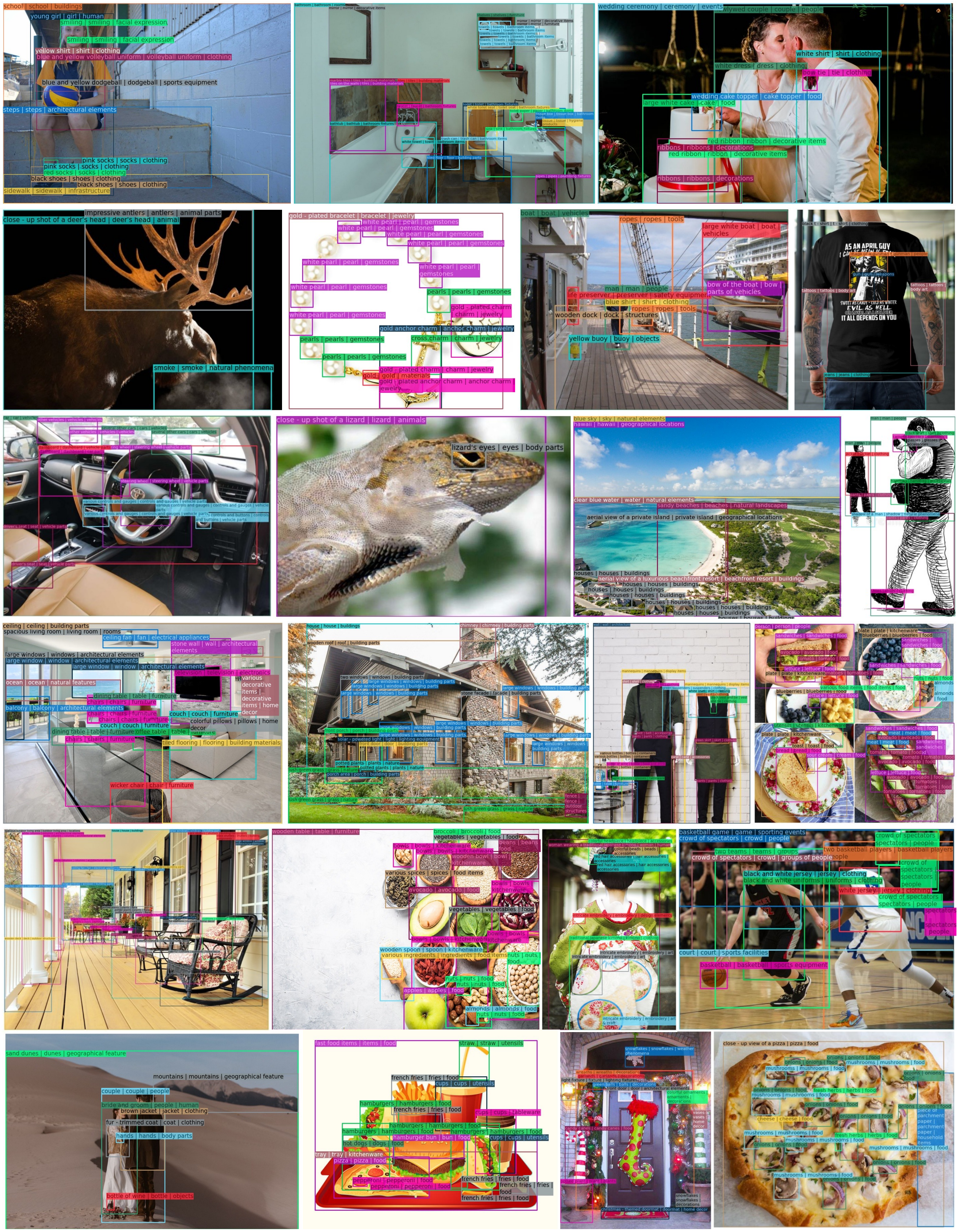

school | school | buildings
young girl | girl | human
smiling | smiling | facial expression
yellow shirt | shirt | clothing
blue and yellow volleyball uniform | volleyball uniform | clothing
blue and yellow dodgeball | dodgeball | sports equipment
steps | steps | architectural elements
pink socks | socks | clothing
red socks | socks | clothing
black shoes | shoes | clothing
sidewalk | sidewalk | infrastructure
wedding ceremony | ceremony | events
newlywed couple | couple | people
white shirt | shirt | clothing
white dress | dress | clothing
bow tie | tie | clothing
wedding cake topper | cake topper | food
large white cake | cake | food
red ribbon | ribbon | decorative items
ribbons | ribbons | decorations
impressive antlers | antlers | animal parts
close - up shot of a deer's head | deer's head | animal
smoke | smoke | natural phenomena
gold - plated bracelet | bracelet | jewelry
white pearl | pearl | gemstones
pearls | pearls | gemstones
gold anchor charm | anchor charm | jewelry
cross charm | charm | jewelry
gold - plated charm | charm | jewelry
gold | gold | materials
boat | boat | vehicles
ropes | ropes | tools
large white boat | boat | vehicles
man | man | people
life preserver | preserver | safety equipment
blue shirt | shirt | clothing
bow of the boat | bow | parts of vehicles
wooden dock | dock | structures
yellow buoy | buoy | objects
tattoos | tattoos | body art
close - up shot of a lizard | lizard | animals
lizard's eyes | eyes | body parts
blue sky | sky | natural elements
hawaii | hawaii | geographical locations
clear blue water | water | natural elements
sandy beaches | beaches | natural landscapes
aerial view of a private island | private island | geographical locations
houses | houses | buildings
aerial view of a luxurious beachfront resort | beachfront resort | buildings
ceiling | ceiling | building parts
spacious living room | living room | rooms
ceiling fan | fan | electrical appliances
stone wall | wall | architectural elements
large windows | windows | architectural elements
large window | window | architectural elements
television | television | electronics
ocean | ocean | natural features
dining table | table | furniture
chairs | chairs | furniture
balcony | balcony | architectural elements
couch | couch | furniture
colorful pillows | pillows | home decor
tiled flooring | flooring | building materials
wicker chair | chair | furniture
house | house | buildings
chimney | chimney | building parts
wooden roof | roof | building parts
stone facade | facade | building parts
potted plants | plants | nature
lush green grass | grass | nature
wooden table | table | furniture
broccoli | broccoli | food
vegetables | vegetables | food
bowls | bowls | kitchenware
wooden bowl | bowl | kitchenware
various spices | spices | food items
avocado | avocado | food
wooden spoon | spoon | kitchenware
various ingredients | ingredients | food items
nuts | nuts | food
apples | apples | food
almonds | almonds | food
basketball game | game | sporting events
crowd of spectators | crowd | people
two teams | teams | groups
two basketball players | basketball players
crowd of spectators | crowd | groups of people
black and white jersey | jersey | clothing
black and white uniforms | uniforms | clothing
white jersey | jersey | clothing
court | court | sports facilities
basketball | basketball | sports equipment
sand dunes | dunes | geographical feature
mountains | mountains | geographical feature
couple | couple | people
bride and groom | people | human
brown jacket | jacket | clothing
fur - trimmed coat | coat | clothing
hands | hands | body parts
bottle of wine | bottle | objects
fast food items | items | food
straw | straw | utensils
french fries | fries | food
cups | cups | utensils
hamburgers | hamburgers | food
hamburger bun | bun | food
hot dogs | dogs | food
tray | tray | kitchenware
pizza | pizza | food
pepperoni | pepperoni | food
close - up view of a pizza | pizza | food
onions | onions | food
mushrooms | mushrooms | food
cheese | cheese | food
fresh herbs | herbs | food

Figure 4-b. Qualitative results of multi-granular object labels generated by DetCLIPv3's object captioner. In the absence of candidate categories, DetCLIPv3's object captioner generates dense, fine-grained, multi-granular object labels, thus facilitating a more comprehensive image understanding.

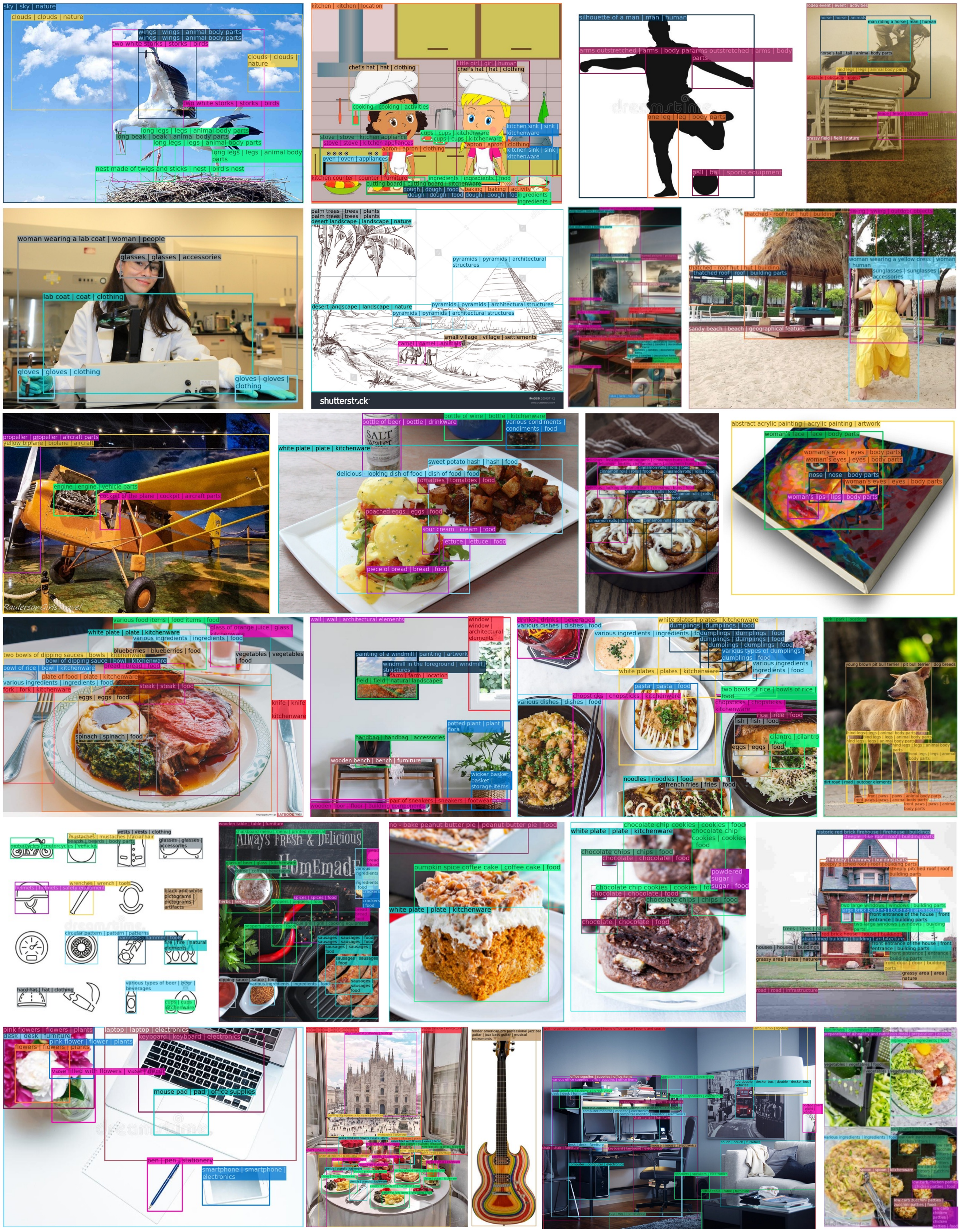

sky | sky | nature
clouds | clouds | nature
wings | wings | animal body parts
two white storks | storks | birds
long legs | legs | animal body parts
long beak | beak | animal body parts
nest made of twigs and sticks | nest | bird's nest
kitchen | kitchen | location
chef's hat | hat | clothing
little girl | girl | human
cooking | cooking | activities
stove | stove | kitchen appliances
cups | cups | kitchenware
kitchen sink | sink | kitchenware
apron | apron | clothing
oven | oven | appliances
kitchen counter | counter | furniture
ingredients | ingredients | food
cutting board | cutting board | kitchenware
dough | dough | food
baking | baking | activity
silhouette of a man | man | human
arms outstretched | arms | body parts
one leg | leg | body parts
ball | ball | sports equipment
rodeo event | event | activities
horse | horse | animal
man riding a horse | man | human
horse's tail | tail | animal body parts
grassy field | field | nature
woman wearing a lab coat | woman | people
glasses | glasses | accessories
lab coat | coat | clothing
gloves | gloves | clothing
palm trees | trees | plants
desert landscape | landscape | nature
pyramids | pyramids | architectural structures
small village | village | settlements
camel | camel | animal
shutterstock
thatched - roof hut | hut | building
woman wearing a yellow dress | woman | human
sunglasses | sunglasses | accessories
thatched roof | roof | building parts
sandy beach | beach | geographical feature
propeller | propeller | aircraft parts
yellow biplane | biplane | aircraft
engine | engine | vehicle parts
cockpit of the plane | cockpit | aircraft parts
bottle of beer | bottle | drinkware
bottle of wine | bottle | kitchenware
various condiments | condiments | food
white plate | plate | kitchenware
sweet potato hash | hash | food
delicious - looking dish of food | dish of food | food
tomatoes | tomatoes | food
poached eggs | eggs | food
sour cream | cream | food
lettuce | lettuce | food
piece of bread | bread | food
cinnamon rolls | rolls | food
abstract acrylic painting | acrylic painting | artwork
woman's face | face | body parts
woman's eyes | eyes | body parts
nose | nose | body parts
woman's lips | lips | body parts
various food items | food items | food
glass of orange juice | glass | kitchenware
white plate | plate | kitchenware
various ingredients | ingredients | food
blueberries | blueberries | food
two bowls of dipping sauces | bowls | kitchenware
bowl of dipping sauce | bowl | kitchenware
vegetables | vegetables | food
bowl of rice | bowl | kitchenware
bread | bread | food
plate of food | plate | kitchenware
steak | steak | food
fork | fork | kitchenware
eggs | eggs | food
knife | knife | kitchenware
spinach | spinach | food
wall | wall | architectural elements
window | window | architectural elements
painting of a windmill | painting | artwork
windmill in the foreground | windmill | structures
farm | farm | location
field | field | natural landscapes
potted plant | plant | flora
handbag | handbag | accessories
wooden bench | bench | furniture
wicker basket | basket | storage items
pair of sneakers | sneakers | footwear
drinks | drinks | beverages
white plates | plates | kitchenware
various dishes | dishes | food
dumplings | dumplings | food
various types of dumplings | dumplings | food
chopsticks | chopsticks | kitchenware
pasta | pasta | food
two bowls of rice | bowls of rice | food
rice | rice | food
fish | fish | food
cilantro | cilantro
noodles | noodles | food
french fries | fries | food
young brown pit bull terrier | pit bull terrier | dog breeds
dirt road | road | outdoor elements
front paws | paws | animal body parts
motorcycles | motorcycles | vehicles
mustaches | mustaches | facial hair
beards | beards | body parts
vests | vests | clothing
glasses | glasses | accessories
helmets | helmets | safety equipment
wrenches | wrench | tools
black and white pictograms | pictograms | artifacts
circular pattern | pattern | patterns
fire | fire | natural elements
hard hat | hat | clothing
various types of beer | beer | beverages
cups | cups | kitchenware
wooden table | table | furniture
herbs | herbs | food
spices | spices | food
sausages | sausages | food
dipping sauce | sauce | food
no - bake peanut butter pie | peanut butter pie | food
pumpkin spice coffee cake | coffee cake | food
white plate | plate | kitchenware
chocolate chip cookies | cookies | food
chocolate chips | chips | food
chocolate | chocolate | food
powdered sugar | sugar | food
historic red brick firehouse | firehouse | building
chimney | chimney | building parts
steeply pitched roof | roof | building parts
two large windows | windows | building parts
front entrance of the house | front entrance | building parts
front door | door | building parts
houses | houses | building
grassy area | area | nature
road | road | infrastructure
pink flowers | flowers | plants
desk | desk | furniture
laptop | laptop | electronics
keyboard | keyboard | electronics
pink flower | flower | plants
vase filled with flowers | vase | vase
mouse pad | pad | office supplies
pen | pen | stationery
smartphone | smartphone | electronics
fender american pro professional jazz bass guitar | jazz bass guitar | musical instruments
couch | couch | furniture

Figure 4-c. Qualitative results of multi-granular object labels generated by DetCLIPv3's object captioner. In the absence of candidate categories, DetCLIPv3's object captioner generates dense, fine-grained, multi-granular object labels, thus facilitating a more comprehensive image understanding.

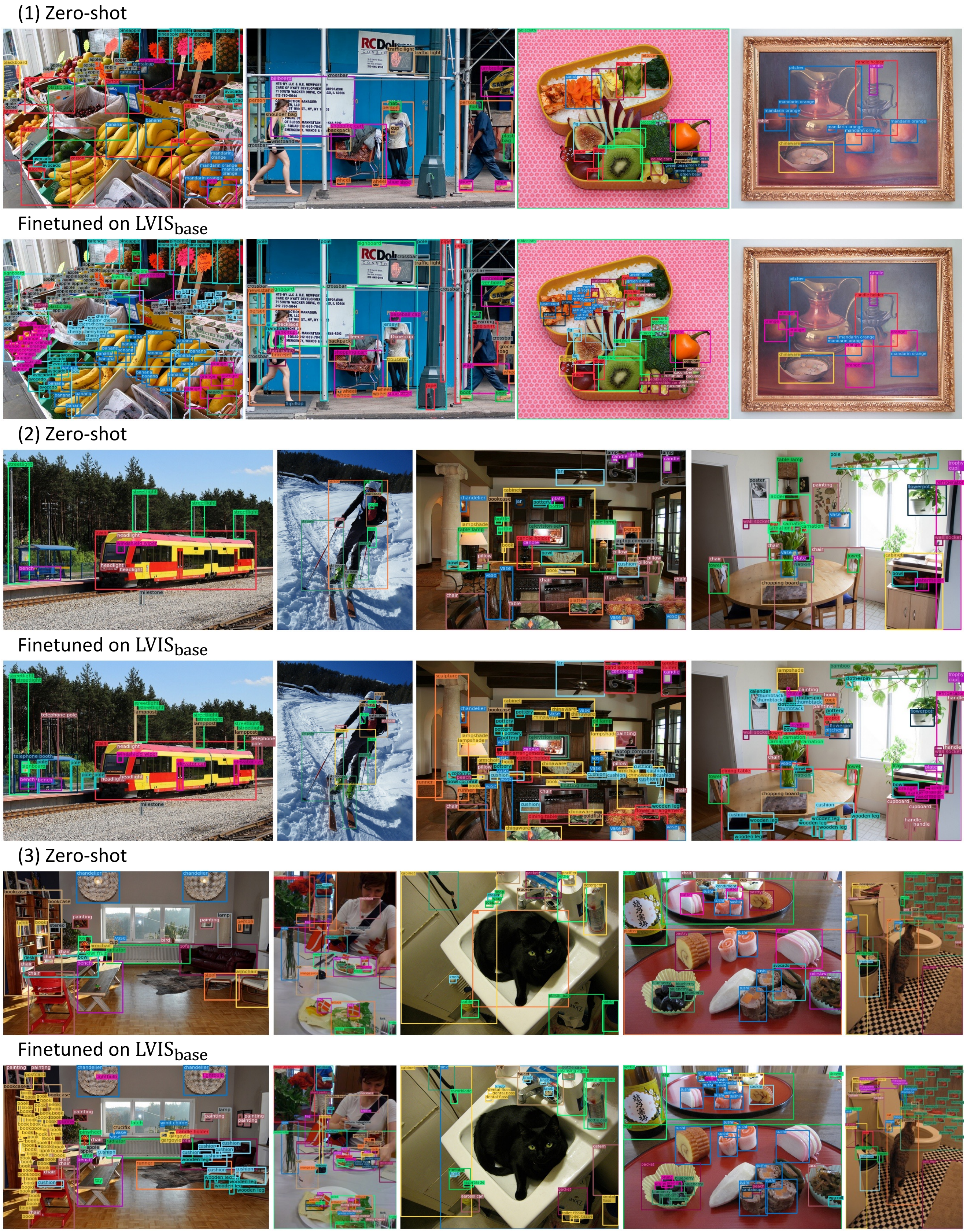

(1) Zero-shot
Finetuned on LVIS$_{base}$
(2) Zero-shot
Finetuned on LVIS$_{base}$
(3) Zero-shot
Finetuned on LVIS$_{base}$

Figure 5. Visualization of detection results on LVIS [6]. In each group, the first row represents the zero-shot results, while the second row indicates the results after fine-tuning on the LVIS base categories.

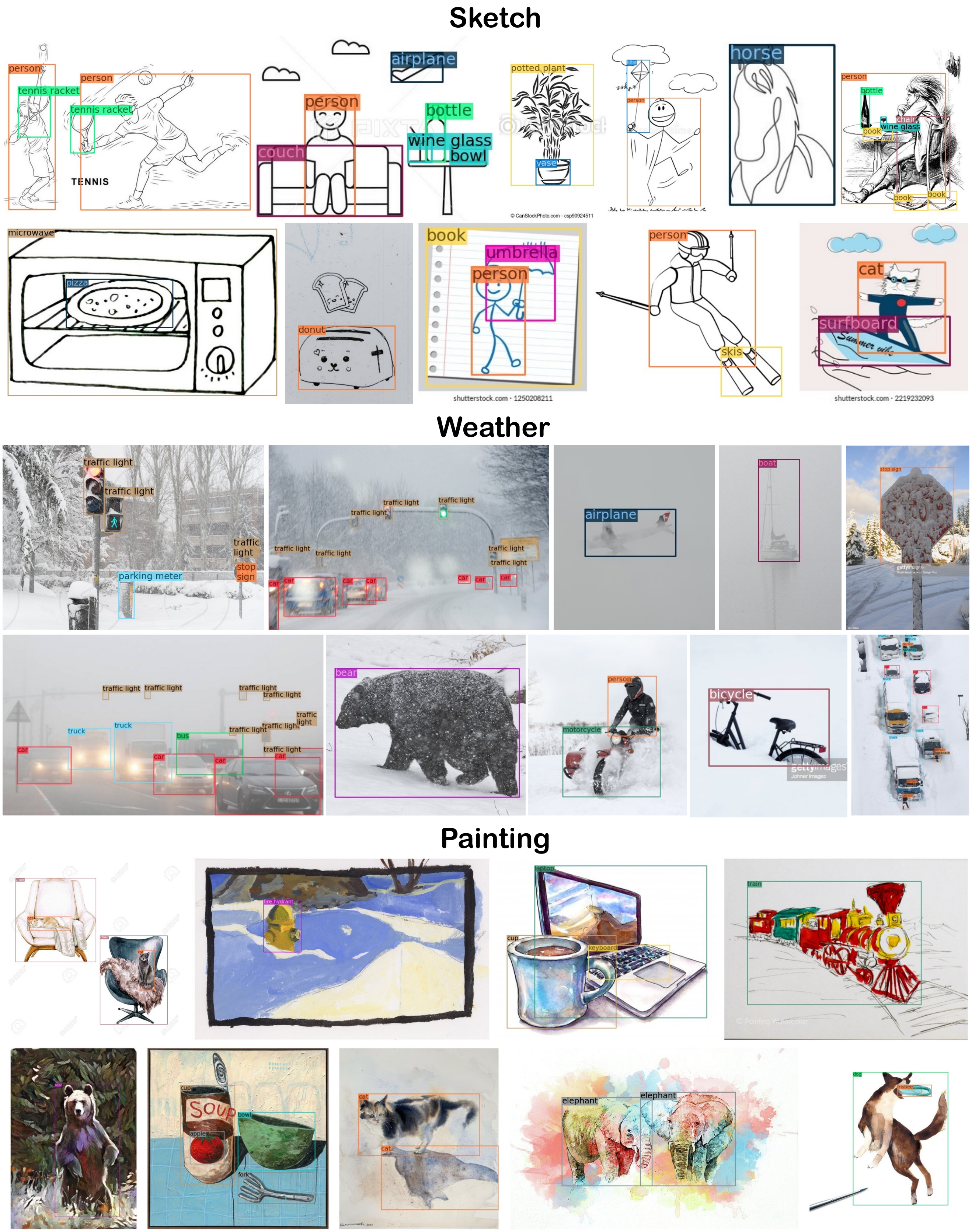


Figure 6-a. Zero-shot detection results on COCO-O[12] dataset. DetCLIPv3 exhibits a robust domain generalization capability.

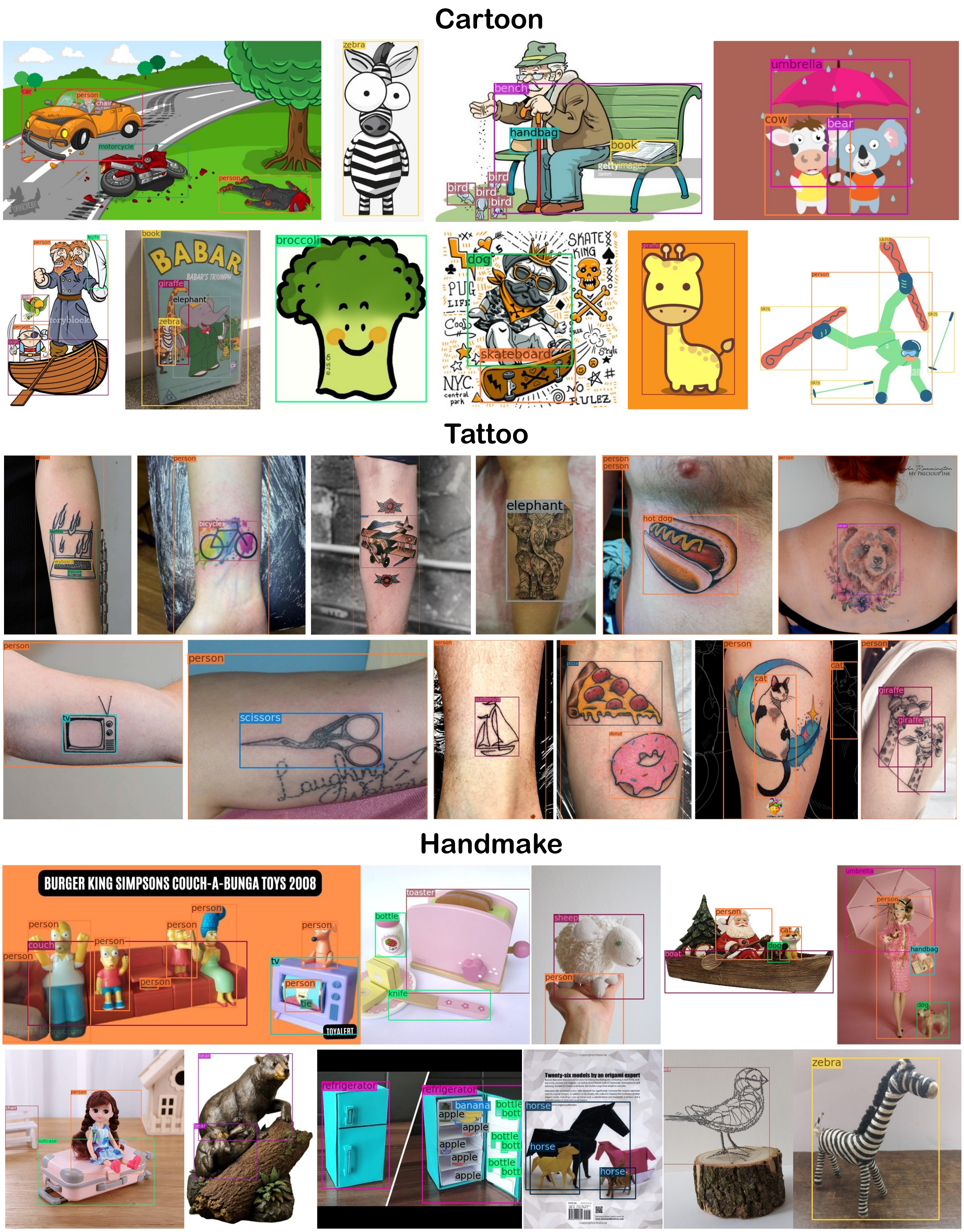


Figure 6-b. Zero-shot detection results on COCO-O[12] dataset. DetCLIPv3 exhibits a robust domain generalization capability.

References

[1] 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. 1

[2] Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. preprint arXiv:1604.06174, 2016. 1

[3] 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] 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. 1

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 1, 2

[6] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 1, 3, 6, 7, 11

[7] 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. 6, 7

[8] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1

[9] 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, 3, 7

[10] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 3

[11] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 2, 3

[12] 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, 12, 13

[13] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. Mixed precision training. In *ICLR*, 2018. 1

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

[15] 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. 1, 2

[16] 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. 1

[17] 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. 1

[18] 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. 1

[19] 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. 1

[20] 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. 1

[21] 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, 7

[22] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. In *ICLR*, 2022. 1

[23] 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, 7

[24] 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. 7

[25] 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. 1

[26] 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. 1