DoGA: Enhancing Grounded Object Detection via Grouped Pre-Training with Attributes

Yang Liu[1,2 *], Feng Hou[1,2 *], Yunjie Peng[3 *], Gangjian Zhang[4], Yao Zhang[4], Dong Xie[4], Peng Wang[4], Yang Zhang[4], Jiang Tian[4], Zhongchao Shi[4], Jianping Fan[4], Zhiqiang He[1,2,5 †]

[1] Institute of Computing Technology (ICT), Chinese Academy of Sciences,
[2] University of Chinese Academy of Sciences,
[3] Beihang University,
[4]AI Lab, Lenovo Research,
[5]Lenovo Ltd.
{liuyang20c, houfeng19}@mails.ucas.ac.cn, yunjiepeng@buaa.edu.cn, hezq@lenovo.com

Abstract

Recent advances in vision-language pre-training have significantly enhanced the model capabilities in open-vocabulary object detection. However, these studies often pre-train with coarse-grained text prompts, such as plain category names and brief grounded phrases. This limitation curtails the model's capacity for fine-grained linguistic comprehension and leads to a significant decline in performance when faced with detailed descriptions or contextual information. To tackle these problems, we propose DoGA: **D**etect **o**bjects with **G**rouped **A**ttributes, which employs commonly apparent attributes to bridge different granular semantics and uses specific attributes to identify the object discrepancy. Our DoGA incorporates three principal components: *1) Generation of attribute-based prompts*, consisting of linguistic definitions enriched with common-sense visible attributes and hard negative notations derived from the image-specific attribute features; *2) Paralleled entity fusion and optimization*, designed to manage long attribute-based descriptions and negative concepts efficiently; and *3) Prompt-wise grouped training* to accommodate model to perform many-to-many assignments, facilitating to process multiple attribute-based synonyms. Extensive experiments demonstrate that training with synonymous attribute-based prompts allows DoGA to generalize multi-granular prompts and surpass previous state-of-the-art approaches, yielding 50.2 on the COCO and 38.0 on the LVIS benchmarks under the zero-shot setting.

Code — https://github.com/liuyang-ict/DoGA

1 Introduction

Object detection is a fundamental task in computer vision, aiming to recognize and localize each object from input images. Traditional detectors are mostly constrained to a closed set of categories defined by training data, *e.g.*, 80 categories in COCO (Lin et al. 2014), failing to perceive a diverse range of visual concepts in an open world. Following the emerging vision-language pre-training approaches (Radford et al. 2021; Jia et al. 2021), GLIP (Li et al. 2022) considers object detection as a grounding problem and leverages additional grounding phrases to facilitate the alignment between visual embeddings and language prompts of objects. The following studies (Liu et al. 2023; Yao et al. 2022; Zhang et al. 2022) further demonstrate the efficacy of grounded pre-training.

Ideally, genuine vision-language detectors should generalize arbitrary linguistic comprehension and accurately locate objects/instances based on the given language prompts. Nevertheless, aligning with rudimentary language prompts (*i.e.*, plain category names, and concise grounding phrases) restricts the model within a coarse-grained level of linguistic comprehension and consequently confuses the model in complicated open-vocabulary detection. For example, even the highly generalized pre-trained detector (Liu et al. 2023) still experiences significant performance deterioration when subjected to categories having fine-grained level descriptions (the AP degrades from 48.2 to 19.2, see Figure 3). As shown in Figure 1(c), contextual definition leads to a mount of inferior boxes with incorrect categories. Furthermore, such rudimentary descriptions also lead to hallucinations and a large sensitivity in referring expression comprehension. As shown in Figure 1(d), when facing the confused prompt, the existing pre-trained detector easily neglects the specific contextual details and cursorily detects the *"black cat"* and *"wooden chair"* with relatively high scores.

To alleviate these problems, we propose **DoGA**, a novel vision-language detector through grouped pre-training with various attributes. The attribute feature offers two advantages: *1) semantic connection* builds the explicit commonality for the same category entity (intra-class compactness) and assists in identifying the different categories (inter-class separation); and *2) instance specificity* describes and represents different individuals within a same entity. Specifically, DoGA is characterized by three core designs:

1) Attribute-based definitions and hard-negative concepts

*This work was done when working as an intern at AI Lab, Lenovo Research, Beijing, China.

†Corresponding author.

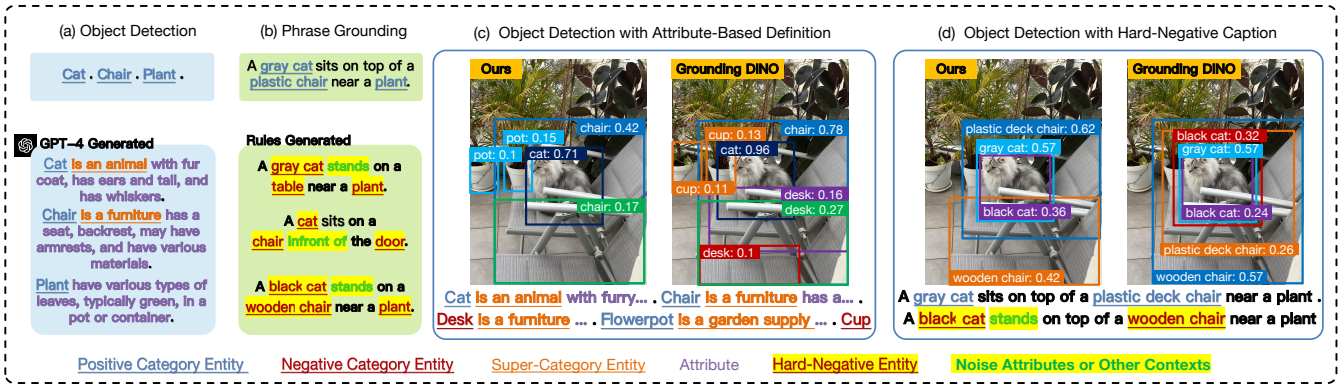


Figure 1: Comparison between Grounding DINO and our proposed DoGA with different attribute-based prompts. (a-b) Different from previous approaches with plain entity names and positive captions for both object detection and phrase grounding, we construct a series of attribute-based definitions for entity names and hard-negative captions with noise attributes and other contexts. (c-d) Compared with Grounding DINO, our DoGA can integrate context information to accurately identify target objects based on conditional definitions and effectively avoid hallucination in detecting hard-negative entities. All entities of negative captions are hard-negative entities. (Best viewed in color.)

construction. We first construct a vocabulary for category entity names where an attribute-based definition is provided, (including, parent category, shape, color, and parts) via large language models (LLMs) to facilitate general semantic class recognition. The attribute-based definition is randomly dropped and mixed to generate multi-granular concepts for each entity name. Moreover, a generation of hard-negative concepts is proposed for phrase grounding to augment the instance-level discrepancy from contextual attributes. Compared to the recent instruction prompt annotations (Ronghao et al. 2024), our goal is to use attributes and entities to construct hard-negative concepts to enhance the model's ability in contextual attributes comprehension and avoid hallucinations when prompting misguided information.

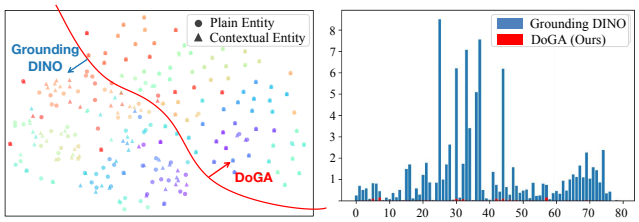
2) Paralleled entity fusion and optimization. Restricted by the max length of connected category prompts (Li et al. 2022; Liu et al. 2023), it is difficult to pre-train with a large number of negative concepts or long-detailed definitions for previous approaches. Instead of connecting whole prompts, DoGA splits each concept and parallelly feeds them into the text encoder. The entity names are then extracted from the resulting concepts and finally applied for text-image and text-query cross-attentions and loss calculation. The paralleled and extracted formulations allow the model to avoid unnecessary interaction, produce longer descriptions, and enforce the model to imbue attributes into the entity.

3) Group-wise synonymous prompts training and inference. During pre-training, one caption/entity often has various synonymous expressions and granularity of attribute-based description. To process the abundant synonyms efficiently, we propose a group-wise synonymous training strategy, extending the previous group-wise one-to-many matching (Chen et al. 2023) to a many-to-many setting. Compared to Group-DETR (Chen et al. 2023) conducting one-to-one assignments for duplicated labels within each group, our proposed group-wise synonymous training flexibly processes to different ground truth labels (box coordinates and prompts) for each group. Moreover, such synonymous training is directly adaptable to the inference time, preventing the extravagance of the pre-trained group parameters.

By training with the generated attribute-based definitions and hard-negative concepts, DoGA exhibits significant inter-class separation and intra-class compactness (Figure 2) at the semantic level. It also demonstrates notable instance specificity (Table 3) for different individual concepts with the same entity. As exhibited in Figure 1(c)-(d), our DoGA can accurately identify target objects based on conditional definitions and effectively avoid hallucination in detecting hard-negative entities. Without the wide variety of training data, it still significantly outperforms the existing state-of-the-art (SoTA) approaches. Based on the Bert-SwinT backbone, the proposed method yields 38.0 AP and 33.5 AP_r on LVIS minival zero-shot transferring, significantly outperforming GLIP (Li et al. 2022), GroundingDINO (Liu et al. 2023), and DetCLIP (Yao et al. 2022) by 10.0 / 9.6 / 1.1 AP, respectively. In the referring expression comprehension task, DoGA constantly promotes existing methods over 1.0 R@1.

2 Related Work

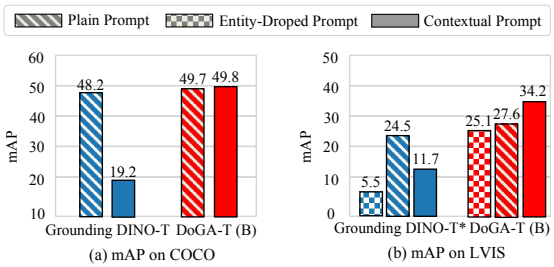
Open-Vocabulary and Grounded Object Detection. Open-vocabulary detection (OVD) aims to utilize the word embeddings to construct the connections between the base and novel classes, and then adapt the detector from existing categories to unseen ones. Benefiting from the success of vision-language pre-training methods, one way is to use the knowledge from pre-trained CLIP to facilitate the detection model in performing zero-shot transferring. VILD (Gu et al. 2021) trained object detectors by distilling visual features from a pre-trained model. ReginoCLIP (Zhong et al. 2022) made the model learn region-level visual representations for fine-grained alignments. On the other hand, some works try to use the semantic enrichment feature of grounding data to cover many rare categories and perform a generalized long-tailed detection where majority, minority, and unknown



(a) Overview of t-SNE visualization of plain/contextual embedding distributions

(b) The quantitative distance between two types of points within the same category

Figure 2: Macro view of entity embeddings by t-SNE.



(a) mAP on COCO

(b) mAP on LVIS

Figure 3: Comparison of multi-granular prompts.

classes coexist. MDETR (Kamath et al. 2021) first introduces phrase grounding into vision-language pre-training to perform both vision and language tasks (*e.g.*, phrase grounding and VQA). GLIP (Li et al. 2022) reformulates object detection as a phrase grounding problem and contracts a unified formulation, recognizing the objects mentioned in the text prompt. Grounding DINO (Liu et al. 2023) further extended this paradigm to the DETR-based model (Carion et al. 2020). Detic (Zhou et al. 2022) and DetCLIP (Yao et al. 2022) turned to solve a large-vocabulary detection problem.

Linguistic Knowledge Enhancement. The recent trend tries to use extra text knowledge of category labels to improve the zero-shot generalization capability of the pre-training model. In the classification task, K-Lite (Shen et al. 2022) utilized external category information from predefined knowledge datasets, such as WordNet (Miller 1995), to learn image representations. Another studies (Yang et al. 2023; Pratt et al. 2023; Menon and Vondrick 2022) applied LLMs to access the definitions of category names. (Doveh et al. 2023) focus on captioning data and generating hard negative (context-sensitive) captions by LLMs for the image classification task. Furthermore, OVAD (Bravo et al. 2023) introduced the attribute concept to the open-vocabulary setting but only focused on attributes instead of categories.

In the detection task, DetCLIP (Yao et al. 2022) provided substantial definitions for category names but leaked out the test-time categories information during the training processes. MM-OVOD (Kaul P 2023) offline trained text-based and visual-based classifiers for OVD. InstructDet (Ronghao et al. 2024) constructed instruction-guided text prompts for visual grounding datasets. DESCO (Li et al. 2023) focused on both plain and grounding datasets but only took a preliminary exploration of the hard negative prompting refinement. Compared with these studies, our proposed DoGA introduces the attribute into detection transformer pre-training, connects different semantic-level language prompts, and efficiently trains them with entity-extracted and grouped mechanisms.

3 How Contextual Entities are Distributed?

A robust open-vocabulary detection system should generalize across diverse semantic-granular linguistic understanding and accurately locate the object based on conditional prompts. To test this generalization ability, we pre-define two different granularity of category prompts: plain category c and contextual category $c^d = \{c, def\}$, where the former is the entity name of the category and the latter is the entity name with a contextual definition. For example, "ice cream" is the entity name c and its attribute-based definition $def = \{sc, \{a_s\}\}$ includes a super-category entity sc "a dessert" and S attributes $\{a_s, s = 1, \cdots, S\}$ "that is usually colorful, soft, served in a cone or cup, often with $\cdots$" As shown in Figure 3, the pre-trained Grounding-DINO (Liu et al. 2023) performs well with plain categories but declines markedly with contextual categories.

To explore its causes, we compare a distribution of two-type entities based on the COCO dataset (Lin et al. 2014) that contains 80 known classes. Concretely, we extract the entity embeddings from both plain and contextual prompts. Then, a macro view inter-class entity embedding distribution is further visualized by t-SNE (Van der Maaten and Hinton 2008) to measure its classification complexity. The detailed setting of t-SNE is presented in Appendix C.

The comparison of the distribution over entity embeddings is shown in Figure 2, where plain and contextual entities are denoted as circle points $\dot{P}_c$ and triangle points $\dot{P}_c^d$, respectively. Each point represents a category entity and we use different colors to distinguish them. In Grounding DINO (Liu et al. 2023), we observe that its output plain-entity points $\dot{P}_c$ are not equally distributed since the model can accurately identify them. When attached to contextual definitions, the distribution of contextual entities $\dot{P}_c^d$ becomes messier. The location of $\dot{p}_{c_i}^d$ may be closer to other plain entities' $\{\dot{p}_{c_j}, j \neq i\}$ rather than its corresponding one $\dot{p}_{c_i}$. Figure 2 further quantitates the offset distance $d_i = ||\dot{p}_{c_i}^d - \dot{p}_{c_i}||^2$ between the plain points $\dot{p}_{c_i}$ and the contextual ones $\dot{p}_{c_i}^d$ of the category i. It is witnessed that there is a large gap in embeddings' distribution after adding contextual definitions. Such a disordered distribution substantially confuses the detection model and causes worse results.

4 Method

In this section, we first review the baseline model Grounding DINO (Liu et al. 2023) (Section 4.1) and introduce the generating process of mixed attribute-based prompts for different tasks (Section 4.2). Subsequently, we propose DoGA to pre-train the detector with these generated prompts (Section 4.3), consisting of pipeline overview, paralleled text encoding, entity-extracted fusion and optimization, and a group-wise synonymous training strategy.

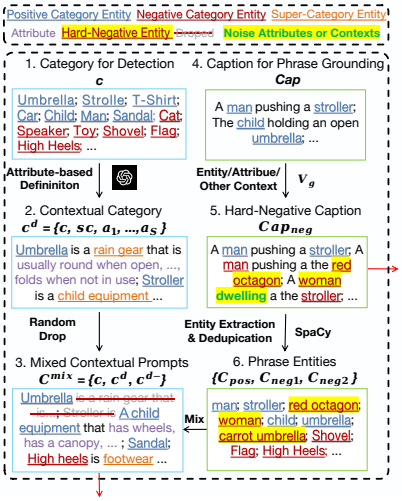


Figure 4: The generation of mixed attribute-based prompts. The processes include generating contextual definitions and hard-negative concepts (1→2→3, 4→5, and 4→5→6→3).

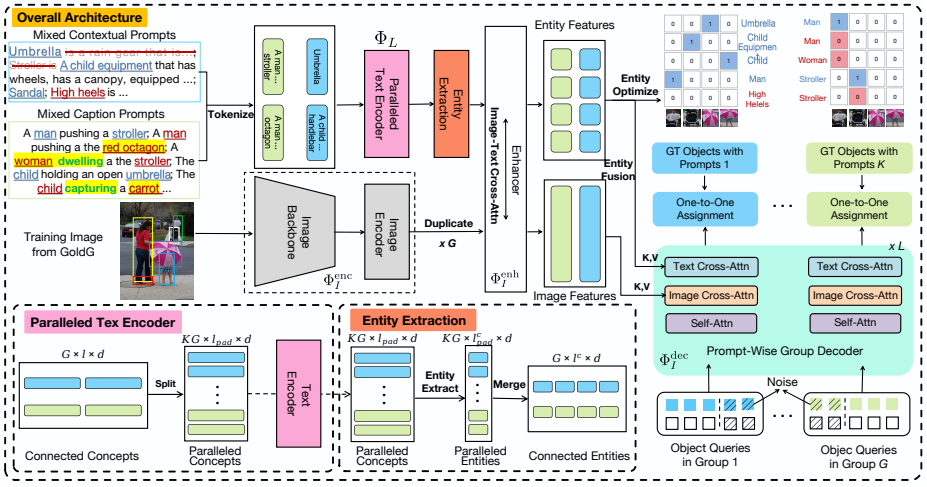


Figure 5: Overall architecture of DoGA with the group-wise synonymous training. DoGA contains a paralleled text encoder and an entity extraction to obtain entity embeddings. The resulting embeddings interact with duplicate image features in feature enhancer and G groups of object queries in the prompt-wise group decoder. Then the many-to-many entity loss is performed.

4.1 Preliminary

Given an image I, the goal of the Grounding DINO (Liu et al. 2023) is to identify the target objects based on the plain category prompts which is a concatenation of K category entity names $C = \{c_1, c_2, \cdots, c_K\}$. Each category c_k is first tokenized with l_k index of tokens $p_{c_k} \in N^{l_k}$ and then connected by a separation token $[SEP]$. For example, when training on the COCO dataset, the connected tokens are "person. bicycle. $\cdots$, toothbrush." and "." is $[SEP]$ by default. The final connected tokens $P_c \in N^l$are processed by a language encoder Φ_L separately according to a category-wise identity attention mask $M^I \in [0,1]^{l\times l}$, where $M^I[l^{j-1}:l^j, l^{j-1}:l^j] = 1$ prevent unnecessary interaction between category prompts in the attention module, $l = K + \sum_{k=1}^{K} l_k$ is the total length of the connected tokens, and $\{l^{j-1} = j-1+\sum_{k=1}^{j} l_k, 1 \le j \le K\}$ is the start length of the jth connected tokens. The resulting embeddings $\boldsymbol{P}_c \in \mathbb{R}^{l\times d}$ are aligned with n region features or object queries $\boldsymbol{O} \in \mathbb{R}^{n\times d}$ from image encoder Φ_I. The whole process and classification loss $\mathcal{L}_{\text{cls}}$ are calculated as

$$\begin{aligned} \boldsymbol{P}_c &= \Phi_L(P_c, M^I), \quad \boldsymbol{O} = \Phi_I(I, \boldsymbol{P}_c), \\ S &= \sigma(\boldsymbol{O}\cdot\boldsymbol{P}_c{}^{\top}), \quad \mathcal{L}_{\text{cls}} = loss(S, T(P_c)), \end{aligned} \tag{1}$$

where σ is *sigmoid* activation, $S \in [0,1]^{n\times l}$ is the token-wise dot-product score, and $T(P_c) \in [0,1]^{m\times l}$ is the prompt-wise one-hot label for m target (Li et al. 2022) based on the tokenized prompts P_c. $loss(S, T(P_c))$ is typically a focal loss (Lin et al. 2020), following DETR based detector.

4.2 Attribute-based Prompts Construction

Learning Category from Common Attributes. We leverage GPT-4 to generate a definition enriched with visual attributes for category names. The constructed vocabulary consists of about 1.6K categories (*i.e.*, 365 classes in Obj365 (Shao et al. 2019) and 1.3K classes in GQA (Hudson and Manning 2019)) appearing in pre-training datasets. For each category, we prompt GPT-4: "*Following the category, please output its super-category and some visual-specific attributes that an object detector needs to attention to*" — *e.g.*, "ice cream" (c) can be defined as "the dessert that is usually colorful, soft, served in a cone or cup, often with a creamy texture, could have toppings." ($def = \{sc, \{a_s\}\}$). More details of the prompts can be found in Appendix B.

We argue that only training with attribute-based definitions may erode the model's generalizability on coarse-grained prompts. As illustrated in Algorithm 1 and 1→2→3 of Figure 4, a mixed training strategy is then proposed to adapt to arbitrary semantic-granularity prompts. Given a set of category entities C, we leverage LLM to generate contextual prompts c^d. The mixture of plain c, contextual c^d, and pruned contextual prompts def/c^{d-} are controlled by α and β chance. There are three types of category prompts randomly mixed without overlaps: plain category $\{c\}$, super-category with the generated attributes $\{c^{d-}\} = \{def\}$, and contextual category $\{c^d\}$.

Learning Instance via Specific Attributes. To achieve real text-free inputs, Li et al. unify conventional object detection and phrase grounding problems, and thus enable the use of massive grounding data by localizing the instances according to specific captions of an image. However, such grounded captions lack enough negative concepts making models hard to understand instances deeply in an open-vocabulary system. To address this problem, we collect a vocabulary V_g from pre-training grounding datasets (Kamath et al. 2021) and attributes detection datasets (Bravo et al. 2023) via spaCy (Honnibal et al. 2020) and NLTK (Bird, Klein, and Loper 2009), covering all category/phrase entities and adjective attributes in the pre-training datasets. By replacing entities or attributes from V_g, hard-negative concepts

Algorithm 1: Mixed Category Prompts for Plain Detection

Data: C (Set of Categories)
$C^{\text{mix}} \leftarrow \emptyset$
for $c \leftarrow C$ **do**
  $c^d \leftarrow \text{LLM}(c)$
  if *random()* $< \alpha$ **then**
    $C^{\text{mix}} \leftarrow C^{\text{mix}} \cup \{c\}$
  else if *random()* $< \alpha + \beta$ **then**
    $c^{d-} \leftarrow \text{PruneCategory}(c^d)$
    $C^{\text{mix}} \leftarrow C^{\text{mix}} \cup \{c^{d-}\}$
  else
    $C^{\text{mix}} \leftarrow C^{\text{mix}} \cup \{c^d\}$
Return C^{mix}

Algorithm 2: Mixed caption Prompts for Phrase Grounding

Data: cap (captions), V_g (Grounding Vocabulary)
if *random()* $> \gamma$ **then**
  $C_{\text{pos}} \leftarrow \text{ExtractEntity}(cap)$
  $A_{\text{pos}} \leftarrow \text{ExtractAttribute}(cap)$
  $P_C, P_A \leftarrow \text{GetProperty}(C_{\text{pos}}, A_{\text{pos}})$
  $cap_{\text{neg}} \leftarrow \text{RandomReplace}(V_g, cap, C_{\text{pos}}, A_{\text{pos}}, P_C, P_A)$
  Return $\text{RandomClip}(cap \cup cap_{\text{neg}})$
else
  $C^i_{\text{pos}} \leftarrow \text{ExtractEntity}(cap_{i1}, \cdots, cap_{ik})$
  $C^i_{\text{pos}} \leftarrow \text{Deduplicate}(C^i_{\text{pos}}, B^i)$
  $C^i_{\text{neg}} \leftarrow \text{Choice}(V_g, \text{WordNet}(C^i_{\text{pos}}))$
  Return $\text{Algorithm1}(C^i_{\text{pos}}, C^i_{\text{neg}})$

are constructed to stimulate the model to understand the attributes and distinct instances even if they belong to the same category entity. Finally, we randomly clip the positive/hard-negative concepts to avoid prompt overfitting after deepfusion. Considering the imbalance of the number of annotations in different datasets, all captions from the same image are collected and converted to entity style with γ choice for plain object detection. In Algorithm 2, the hard-negative caption cap_{neg} is randomly replaced by either phrase entities C_{pos} or other context words A_{pos} according to their character property P_C, P_A. All captions from image i have γ chance to be extracted to phrase entities C^i_{pos} and deduplicated by the IoU among B^i. After collecting negative entities C_{neg} from the V_g, they are further processed by Algorithm 1.

4.3 Training with Attribute-based Prompts

Pipeline Overview. As illustrated in Figure 5, the connected prompts are split and individually encoded by a paralleled text encoder. The resulting prompt embeddings then undergo entity extraction to integrate the attributes and other contextual information into the category/phrase entity embeddings. Before the image-text cross-attention, the image features are duplicated by G to integrate different groups of text features accordingly. In the prompt-wise group decoder, both object and noise queries are initialized by grouped image features and subsequently integrate entity and image features across G different groups (where G = 2 in the figure). The final alignment losses focus solely on entity scores, where the red square and zero value indicate the hard-negative entity with noise contexts. For instance, the embedding of "the man" with "pushing a stroller" context is considered positive, whereas the same entity paired with "pushing a red octagon" represents the hard-negative.

Paralleled Text Encoder. When we pre-train DoGA on attribute-based prompts, we find it becomes low efficient and takes heavy memory cost with length increases of connected prompts $P \in N^l$. This is because it calculates all token-wised self-attention maps before using the identity attention mask M^I. As a practical solution, the token $p_k \in N^{l_k}$ is parallelly fed into the text encoder to obtain the corresponding embedding $\boldsymbol{p}_k \in \mathbb{R}^{l_{\text{pad}} \times d}$, where each class token is padded to l_{pad} tokens. Finally, the resulting embeddings are remerged to $\boldsymbol{P} \in \mathbb{R}^{l \times d}$. The process is formulated as

$$\boldsymbol{P} = \texttt{Merge}(\Phi_L(\texttt{Split}(P))). \tag{2}$$

Entity-extracted Fusion and Optimization. At the text fusion stage, the attention weights computation is of quadratic computation w.r.t. text token numbers. Thus, it is of very high computational and memory costs to process vast negative samples and long contextual prompt inputs. On the other hand, the generated common-sense attributes may not appear in the given images, which is ambiguous to optimize whether the agnostic attributes are aligned with the image-specific object. In the grounding dataset source, only the category (or phrase) entities are labeled, and the complicated attributes and other contexts make it hard to evaluate.

The above-mentioned issues can be mainly attributed to the deficit of the text encoder in processing contextual prompts. We hope it implicitly imbues attributes into entity embeddings, thus avoiding explicit interaction and optimization with contextual text features. To this end, we keep all prompts perform paralleled self-attention but extract the entity-related prompts/embeddings $\boldsymbol{P}_c \in \mathbb{R}^{K l^c_{\text{pad}} \times d}$ (*i.e.* category, phrase, or super-category (when without category) entities) for text-image/query interactions and loss optimization, where l^c_{pad} is the max length l^c_k of K entity tokens. The attribute-based loss is finally calculated by

$$\begin{gathered} \boldsymbol{P}^d_c = \texttt{Extract}(P^d, \boldsymbol{P}^d),\ \boldsymbol{O} = \Phi_I(I, \boldsymbol{P}^d_c), \\ S_c = \sigma(\boldsymbol{O} \cdot \boldsymbol{P}^{d\top}_c),\ \mathcal{L}_{\text{cls}} = loss(S_c, T(P^d_c)). \end{gathered} \tag{3}$$

The `Extract` operation is equal to an identity process when all categories are the original plain ones without attributes or other contextual captions.

Group-Wise Synonymous Training. Unlike standard object detection, vision-language pre-training enables models to detect specified objects in response to conditional prompts. However, the synonymous nature of language and the incompleteness of conditional prompts complicate the training process of one-to-one label assignment, making it challenging to handle multi-granular synonyms or different captions of the same image. To overcome this challenge, we propose a synonymous group-wise training mechanism that facilitates many-to-many matching. The prediction of each group is formulated as

$$\begin{gathered} I_i, \boldsymbol{P}^d_{c_i} = \Phi^{\text{enh}}_I(I, \boldsymbol{P}^d_{c_i}),\ \boldsymbol{O}_i = \Phi^{\text{dec}}_I(I_i, \boldsymbol{O}^{(0)}_i, \boldsymbol{P}^d_{c_i}), \\ \mathcal{L}_{\text{cls}} = \frac{1}{G} \sum_i^G loss\big(\sigma(\boldsymbol{O}_i \cdot \boldsymbol{P}^{d\top}_{c_i}), T(P^d_{c_i})\big), \end{gathered} \tag{4}$$

Different from the naive Group-DETR (Chen et al. 2023), where G group object queries $\{\boldsymbol{O}_i = (\boldsymbol{E}_i, \hat{B}_i) | i = 1, \cdots, G\}$

Model	Backbone	Pre-Training Data	Epoch	COCO val	LVIS minival		ODinW13 test
				AP	AP	AP_r / AP_c / AP_f	AP_{avg}
GLIP-T (B) (Li et al. 2022)	Swin-T	O365v1	30	44.9	17.8	13.5 / 12.8 / 22.2	**33.2**
Grounding-DINO-T (Liu et al. 2023)	Swin-T	O365v1	30+	46.7	16.1	- / - / -	-
DoRA-T (A) (**Ours**)	Swin-T	O365v1	12	**47.4**	**22.8**	**14.3 / 19.9 / 26.8**	31.9
DoRA-T (A) w/PTTA (**Ours**)	Swin-T	O365v1	12	**47.4**	**23.9**	**16.0 / 20.9 / 28.1**	-
GLIP-T (C) (Li et al. 2022)	Swin-T	O365v1, GoldG	30	46.7	24.9	17.7 / 19.5 / 31.0	44.4
Grounding-DINO-T (Liu et al. 2023)	Swin-T	O365v1, GoldG	30+	48.1	25.6	14.4 / 19.6 / 32.2	40.3
DoRA-T (B) (**Ours**)	Swin-T	O365v1, GoldG	12	**49.8**	**34.2**	**29.4 / 32.4 / 36.8**	**48.2**
DoRA-T (B) w/PTTA (**Ours**)	Swin-T	O365v1, GoldG	12	**49.8**	**35.6**	**30.6 / 33.3 / 38.4**	-
GLIP-T (Li et al. 2022)	Swin-T	O365v1, GoldG, Cap4M	30	46.3	26.0	20.8 / 21.4 / 31.0	46.5
Grounding-DINO-T (Liu et al. 2023)	Swin-T	O365v1, GoldG, Cap4M	30+	48.4	27.4	18.1 / 23.3 / 32.7	45.9
DetCLIP-T (Yao et al. 2022)	Swin-T	O365v2$^\diamond$, GoldG, YFCC1M	12	-	35.9	33.2 / 35.7 / 36.4	43.4
DesCo-GLIP-T (Li et al. 2023)	Swin-T	O365v1, GoldG, CC3M$^\diamond$	7†	45.8	34.6	30.8 / 30.5 / 39.0	-
DoRA-T (C) (**Ours**)	Swin-T	O365v1, GoldG, VAW+	12	**50.1**	**36.5**	32.0 / 35.2 / **39.1**	**49.4**
DoRA-T (C) w/PTTA (**Ours**)	Swin-T	O365v1, GoldG, VAW+	12	**50.2**	**38.0**	**33.5 / 36.9 / 39.8**	-

Table 1: Zero-shot transferring results on three mainstream detection tasks. † denotes fine-tuning. $^\diamond$ refers to the subsets.

Method	Backbone	Pre-Training Data	Flickr30k	RefCOCO			RefCOCO+			RefCOCOg	
			R@1 / R@5 / R@10	val	testA	testB	val	testA	testB	val	test
GLIP-T	Swin-T	O365v1, GoldG	84.8 / 94.9 / 96.3	49.96	54.69	43.06	49.01	53.44	43.42	65.58	66.08
Grounding-DINO-T	Swin-T	O365v1, GoldG	- / - / -	50.41	57.24	43.21	51.40	57.59	45.81	67.46	67.13
DoGA-T (B) (**Ours**)	Swin-T	O365v1, GoldG	**85.8 / 95.8 / 97.5**	**51.50**	**57.45**	**46.10**	**51.49**	57.57	**46.88**	67.42	**68.14**
GLIP-T	Swin-T	O365, GoldG, Cap4M	85.7 / 95.4 / 96.9	50.42	54.30	43.83	49.50	52.78	44.59	66.09	66.89
DesCo-GLIP	Swin-T	O365v1, GoldG, CC3M$^\diamond$	85.3 / 95.8 / 97.3	-	-	-	-	-	-	-	-
DoGA-T (C) (**Ours**)	Swin-T	O365v1, GoldG, VAW+	**86.3 / 96.2 / 98.1**	**52.20**	**58.19**	**46.72**	**51.83**	**58.21**	**47.36**	**68.17**	**69.25**

Table 2: Top-1 recall comparison on the referring expression comprehension tasks. $^\diamond$ refers to the subsets.

are independently optimized with duplicate target objects in the decoder Φ_I^{dec}, and Φ_I^{enh} is the enhancer for image-text cross-interaction, $\boldsymbol{O}_i^{(0)}$ is learned query embeddings for initialization of $\boldsymbol{O}_i$. We employ different synonyms/captions $P_{c_i}^{d/cap}$ to independently optimize G group queries based on the group-specific label $T(P_{c_i}^{d/cap})$ (as shown in Figure 5). To adapt to Grounding DINO (Liu et al. 2023), each group is assigned a series of noise queries whose embedding is initialized by group-specific prompt features. The noise initialization $\boldsymbol{O}_{n_i}^{(0)}$ is then calculated by

$$\boldsymbol{O}_{n_i}^{(0)} = \texttt{RandChoice}\big(T(P_{c_i}^d) \cdot f\big(\texttt{Detach}(\boldsymbol{P}_{c_i}^d)\big)^\top\big), \quad (5)$$

where f is a linear layer and `Detach` is detached gradient.

5 Experiments

We train DoGA-T with 12 training epochs and use Swin-Tiny (Liu et al. 2021) and BERT-base (Devlin et al. 2019) as backbone. We use 24 V100 GPUs and batch size 24 to train with group operations and 16 V100 GPUs and batch size 48 without them. Following previous studies (Li et al. 2022; Liu et al. 2023), DoGA is pre-trained on three mainstream data sources: O365v1 (Shao et al. 2019) consisting of 0.66M training images and 365 categories, 0.8M GoldG data collected by MDETR (Kamath et al. 2021), and VAW+ including 0.15M images from VAW (Pham et al. 2021) and GQA (Hudson and Manning 2019) with large attribute labels, both of which leverage and refine annotations from Visual Genome (Krishna et al. 2017). According to different pre-training datasets, we denote DoGA by (A), (B), and (C). More dataset details can be found in Appendix B.

5.1 Main Results

Object Detection. Table 1 comprehensively compares our DoGA with recent SoTA approaches over COCO (Lin et al. 2014), LVIS (Gupta, Dollar, and Girshick 2019), and ODinW 13 (Li et al. 2022) datasets. Concretely, we provide A, B, and C, three types of pre-training data sources for fair evaluation with previous approaches. DoGA constantly outperforms the current SoTA methods, averagely improving Grounding-DINO (Liu et al. 2023) by 1.3 AP on COCO, 9.0 AP and 15.0 AP_r on LVIS, and 5.7 AP on ODinW.

Phrase Grounding. Table 2 compares the proposed model on Referring Expression Comprehension (REC) tasks. Without the leakage of COCO images, DoGA retains the identified capability and achieves comparable results compared with recent studies. For more fine-tuning comparison with recent specific models, such as InstructDet (Ronghao et al. 2024) and UNINEXT (Yan et al. 2023), please refer to Table 14 of Appendix E for more details.

Semantic Consistency. In a macro-view, Figure 2 compares the entity distributions based on two different semantic level prompts over two pre-trained models. Based on the "large margin" principle, both entity prompts will show almost uniform distribution t-SNE$(P_c/P_c^d) \sim U(0,1)$ to support similar results on COCO. Moreover, Table 4 exhibits a high cosine-similarity between plain and contextual entities in known class, $\text{sim}(p_{c_i}, p_{c_i}^d) \ll \text{sim}(p_{c_i}, p_{c_j}^d)$. Therefore, a coincident uniform distribution is concluded in DoGA, which further demonstrates that such an attribute-based pre-training explicitly empowers the model to focus on the common attributes and imbues them to the entities.

In a micro-view, we evaluate this consistency on LVIS with four frequency segments: known, frequent, common, and rare. As reported in Table 4, the entity prompts present

Model	Persian	Bengal	Abyssinian	Ragdoll	Sphynx
Grounding-DINO	0.974	0.952	0.968	0.952	0.968
DoGA (**Ours**)	0.879	0.745	0.861	0.866	0.869

Table 3: Entity similarity of the "cat" in distinct definitions.

Info	Known	Frequent	Common	Rare
Count	183	213	324	666
Mean	0.85	0.81	0.79	0.71

Table 4: Entity similarity on different frequency segments.

Row	Components	Pre-Training Dataset	COCO	LVIS minival	
			AP	AP	AP_r
1	Baseline*	O365v1,GoldG	46.5	24.5	14.6
2	+Contextual	O365v1,GoldG	46.0	27.7	18.0
3	+Hard-Negative	O365v1,GoldG	46.4	30.6	24.0
4	+GST	O365v1,GoldG	49.8	34.2	29.4
5	+PTTA	O365v1,GoldG	49.8	35.6	30.6
6	+VAW+ Data	O365v1,GoldG,VAW+	50.2	38.0	33.5

Table 5: Effectiveness of each component. * is reproduced

LLM Version	AP	AP_r	Noise Scale	AP	AP_r
gpt3-babbage	27.8	19.8	1.0	22.0	14.5
gpt3-curie	28.7	21.9	0.8	23.1	13.8
gpt3-davinci	32.6	26.6	0.6	23.4	15.6
gpt3.5-turbo	33.4	27.1	0.4	24.7	15.9
lallma-3	33.6	28.0	0.2	24.9	17.8
gpt-4	34.2	29.2	0.0	34.2	29.2

Table 6: Bias of prompts with different qualities and noises.

a high degree of similarity in known class (0.85), while it gradually diverges as the frequency decreases (0.81→0.71). This result further demonstrates the promotion of AP_c and AP_r when DoGA inference with contextual categories.

Instance Specificity. As shown in Table 3, we adopt 12 kinds of cats with their specific attributes under the "cat" entity to construct instance prompts, *e.g.*, *Bengal: a cat, has striped and spotted coat patterns that... .* The cosine similarities over these prompts are calculated to indicate the relative distance within the same class of the "*cat*". It is witnessed that DoGA exhibits highly isolated distribution and accurately performs specific instance recognition while Grounding-DINO neglects the mostly contextual information and aggregates the approximate features. More details of the cat prompts are set out in Appendix G.

Prompt-Wise Test-Time Augmentation. We aim to build a complete grouped framework that enables the detector to leverage multi-granular synonymous concepts during both training and inference. To bridge this gap, we explore a prompt-wise test-time augmentation (PTTA) based on the mixture of synonymous prompts. By applying detailed synonymous prompts, ensemble queries, and weighted boxes fusion (Solovyev, Wang, and Gabruseva 2021), DoGA exhibits significant improvements, with an average increase of +1.1 AP and +2.0 AP_r on LVIS (Table 5). For COCO evaluation, the results remain stable because of the high compactness of known class features, indicating the robustness of our approach. Overall, integrating these results from different granular prompts is essential, especially when the most suitable prompt is unknown.

5.2 Ablation

Effectiveness of Each Component. Table 5 compares each component's efficacy. 1) Introducing the contextual prompts significantly improves performance, boosting baseline to 27.1. This demonstrates that DoGA effectively integrates attributes into the entity name. By constructing a correspondence between visual and textual attributes, the model's recognition of unknown classes with common attributes is facilitated. This hypothesis is further supported by the declining results in Table 6 with noised or inferior prompts. Moreover, attribute inclusion has also minimal impact on commonly recognized close-set categories, particularly the 80 most frequent categories in COCO. Excessive contextual information may complicate model optimization. This effect was previously observed in Grounding DINO, where training Grounding DINO yielded lower COCO mAP compared to DINO (class name *v.s.* one-hot). 2) This correspondence is further enhanced by hard-negative examples, increasing performance to 30.6. It helps DoGA handle more false rejections, thereby reducing hallucinations. 3) GST leads to substantial improvements for both known and rare classes. This likely occurs attributed to a more stable optimized direction from multiple positive synonyms.

Bias of Prompts. We assume that the definition quality is positively correlated with the capabilities of the language model, becoming stronger as the model is updated. Therefore, as shown in Table 6, we evaluate different types of prompts generated by various LLM versions and observe an improvement with more advanced LLMs. Additionally, a series of noise definitions are utilized to evaluate the model in prompts' correctness, which demonstrates that DoGA innately handles the hallucinations for noise prompts, and higher-correctness definitions significantly improve detection accuracy. This finding highlights the importance of high-quality prompts, as they effectively embed valuable visual attribute information. Finally, we explore the effect of prompt granularity in Figure 3(b). The entity name is indispensable (34.2 *v.s.* 25.1) and fine-grained definitions empower DoGA to detect accurately (27.6 *v.s.* 34.2).

6 Conclusion and Future Work

In this paper, we propose DoGA, aimed at applying attributes to group pre-train image-text models tailored for multi-granular text recognition. We construct two types of attribute-based prompts (*i.e.* contextual and hard-negative prompts) for both detection and grounding dataset sources. To efficiently pre-train with many long prompts, a paralleled text encoder and entity-extracted fusion and optimization are employed for the contextual information infusion. Moreover, we introduce a paralleled group training strategy to address multi-granular synonymous categories/captions, extending optimization to a many-to-many setting and bridging the gap between training, inference, and assembly. Experimentally, DoGA outperforms SoTA methods across multiple baselines. In the future, we will explore scaling DoGA to larger models and more pre-training datasets.

Acknowledgements

We thank the Lenovo Research AI Master platform for GPU supports without which this work would not be possible.

References

Bird, S.; Klein, E.; and Loper, E. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Bravo, M. A.; Mittal, S.; Ging, S.; and Brox, T. 2023. Open-vocabulary attribute detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7041–7050.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Chen, Q.; Chen, X.; Wang, J.; Zhang, S.; Yao, K.; Feng, H.; Han, J.; Ding, E.; Zeng, G.; and Wang, J. 2023. Group detr: Fast detr training with group-wise one-to-many assignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6633–6642.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

Doveh, S.; Arbelle, A.; Harary, S.; Schwartz, E.; Herzig, R.; Giryes, R.; Feris, R.; Panda, R.; Ullman, S.; and Karlinsky, L. 2023. Teaching structured vision & language concepts to vision & language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2657–2668.

Gu, X.; Lin, T.-Y.; Kuo, W.; and Cui, Y. 2021. Open-vocabulary Object Detection via Vision and Language Knowledge Distillation. In *International Conference on Learning Representations*.

Gupta, A.; Dollar, P.; and Girshick, R. 2019. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5356–5364.

Honnibal, M.; Montani, I.; Van Landeghem, S.; and Boyd, A. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Hudson, D. A.; and Manning, C. D. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6700–6709.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 4904–4916. PMLR.

Kamath, A.; Singh, M.; LeCun, Y.; Synnaeve, G.; Misra, I.; and Carion, N. 2021. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1780–1790.

Kaul P, Z. A., Xie W. 2023. Multi-modal classifiers for open-vocabulary object detection. In *International Conference on Machine Learning*, 15946–15969. PMLR.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123: 32–73.

Li, L. H.; Dou, Z.-Y.; Peng, N.; and Chang, K.-W. 2023. DesCo: Learning Object Recognition with Rich Language Descriptions. *arXiv preprint arXiv:2306.14060*.

Li, L. H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; et al. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10965–10975.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2020. Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2): 318–327.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Menon, S.; and Vondrick, C. 2022. Visual Classification via Description from Large Language Models. In *The Eleventh International Conference on Learning Representations*.

Miller, G. A. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11): 39–41.

Pham, K.; Kafle, K.; Lin, Z.; Ding, Z.; Cohen, S.; Tran, Q.; and Shrivastava, A. 2021. Learning to predict visual attributes in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13018–13028.

Pratt, S.; Covert, I.; Liu, R.; and Farhadi, A. 2023. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15691–15701.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ronghao, D.; Jiangyan, F.; Haodong, Z.; Chongjian, G.; Lin, S.; Lijun, G.; Chengju, L.; Qijun, C.; Feng, Z.; Rui, Z.; and Yibing, S. 2024. InstructDET: Diversifying Referring Object Detection with Generalized Instructions. In *The Thirteenth International Conference on Learning Representations*.

Shao, S.; Li, Z.; Zhang, T.; Peng, C.; Yu, G.; Zhang, X.; Li, J.; and Sun, J. 2019. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8430–8439.

Shen, S.; Li, C.; Hu, X.; Xie, Y.; Yang, J.; Zhang, P.; Gan, Z.; Wang, L.; Yuan, L.; Liu, C.; et al. 2022. K-lite: Learning transferable visual models with external knowledge. In *Advances in Neural Information Processing Systems*, volume 35, 15558–15573.

Solovyev, R.; Wang, W.; and Gabruseva, T. 2021. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, 107: 104117.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Yan, B.; Jiang, Y.; Wu, J.; Wang, D.; Luo, P.; Yuan, Z.; and Lu, H. 2023. Universal instance perception as object discovery and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15325–15336.

Yang, Y.; Panagopoulou, A.; Zhou, S.; Jin, D.; Callison-Burch, C.; and Yatskar, M. 2023. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19187–19197.

Yao, L.; Han, J.; Wen, Y.; Liang, X.; Xu, D.; Zhang, W.; Li, Z.; Xu, C.; and Xu, H. 2022. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. In *Advances in Neural Information Processing Systems*, volume 35, 9125–9138.

Zhang, H.; Zhang, P.; Hu, X.; Chen, Y.-C.; Li, L.; Dai, X.; Wang, L.; Yuan, L.; Hwang, J.-N.; and Gao, J. 2022. Glipv2: Unifying localization and vision-language understanding. In *Advances in Neural Information Processing Systems*, volume 35, 36067–36080.

Zhong, Y.; Yang, J.; Zhang, P.; Li, C.; Codella, N.; Li, L. H.; Zhou, L.; Dai, X.; Yuan, L.; Li, Y.; et al. 2022. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16793–16803.

Zhou, X.; Girdhar, R.; Joulin, A.; Krähenbühl, P.; and Misra, I. 2022. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, 350–368. Springer.