CONSISTENCY BEYOND CONTRAST: ENHANCING OPEN-VOCABULARY OBJECT DETECTION ROBUSTNESS VIA CONTEXTUAL CONSISTENCY LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in open-vocabulary object detection focus primarily on two aspects: scaling up datasets and leveraging contrastive learning to align language and vision modalities. However, these approaches often neglect internal consistency within a single modality, particularly when background or environmental changes occur. This lack of consistency leads to a performance drop because the model struggles to detect the same object in different scenes, which reveals a robustness gap. To address this issue, we introduce Contextual Consistency Learning (CCL), a novel framework that integrates two key strategies: Contextual Bootstrapped Data Generation (CBDG) and Contextual Consistency Loss (CCLoss). CBDG functions as a data generation mechanism, producing images that contain the same objects across diverse backgrounds. This is essential because existing datasets alone do not support our CCL framework. The CCLoss further enforces the invariance of object features despite environmental changes, thereby improving the model's robustness in different scenes. These strategies collectively form a unified framework for ensuring contextual consistency within the same modality. Our method achieves state-of-the-art performance, surpassing previous approaches by +16.3 AP on OmniLabel and +14.9 AP on D^3. These results demonstrate the importance of enforcing intra-modal consistency, significantly enhancing model generalization in diverse environments. Data, code and models will be made publicly available.

1 INTRODUCTION

Object detection has made significant strides in recent years. However, two advanced tasks based on this technology continue to present considerable challenges: open-vocabulary object detection (OVOD) and descriptive textual object detection, such as referring expression comprehension (REC) and visual grounding (VG). Open-vocabulary object detection aims to detect previously unseen objects in dynamic environments. Recent works Dou et al. (2022); Gu et al. (2021); Li et al. (2022b); Lin et al. (2022); Minderer et al. (2023); Zhao et al. (2022); Jin et al. (2024); Zang et al. (2024) have advanced training strategies for such tasks, others Kamath et al. (2021); Kuo et al. (2022); Minderer et al. (2022); Subramanian et al. (2022) have focused on enhancing model architectures. In parallel, tasks involving referring expressions and visual grounding, which require detecting objects based on complex natural language descriptions, have shown advances in training methodologies Xie et al. (2025); Chen et al. (2025); Zong et al. (2025); Lin et al. (2024); Peng et al. (2023), architectural improvements Yin et al. (2025); Lin et al. (2023); You et al. (2023) and the use of the capabilities of large models Shen et al. (2025); Xuan et al. (2024); Zhan et al. (2024).

Despite these advancements, there is still a crucial gap in addressing the internal consistency within each input image and query. We identify an issue in existing models Dou et al. (2022); Li et al. (2022b; 2023a): the features of the same object tend to vary significantly across different scenes, which indicates that current models may overfit to specific training backgrounds. This inconsistency not only affects the detection stability but might also degrade the model's generalization ability, raising an important question: *Can we obtain object features that are robust to environmental changes?* To validate this, we construct the D^3_{BC} test set by applying background replacement to the original D^3 dataset. Detailed in Section 4.3, baseline methods suffer notable performance drops under this setting, highlighting their limited robustness to contextual changes. In contrast, our method maintains

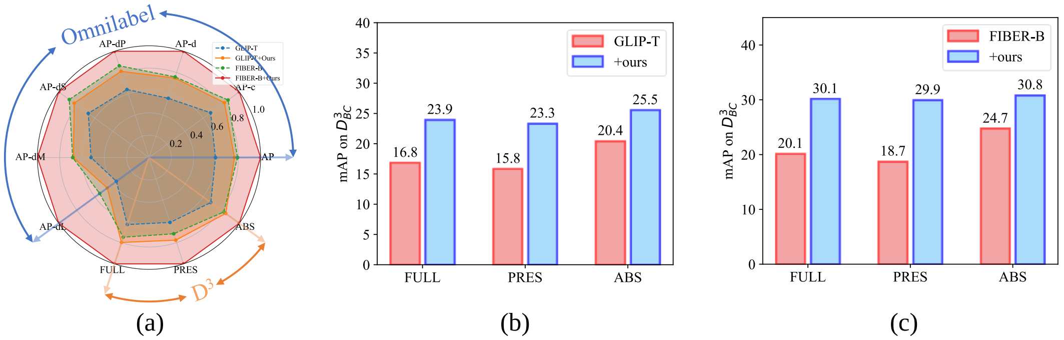

Figure 1: Performance and robustness comparison of different methods. (a) Our approach, with Contextual Consistency Learning, achieves the best overall results, reaching a normalized score of 1 in all metrics. (b,c) Benchmark backgrounds are altered to test robustness. Tested on $D^3{}_{\text{BC}}$, baseline methods degrade, while ours remains stable. See Section 4.3 for details.

performance comparable to that on the original benchmark. As shown in Figure 1, our experimental results demonstrate that addressing this issue significantly improves model performance.

To address this issue, we propose the CCL framework that enforces invariance of object features across different scenes, as shown in Figure 2. However, existing datasets exhibit a notable limitation: they lack comprehensive data pairs that depict the same object in diverse contextual settings. This data gap is crucial because CCL requires models to encounter and learn from variations of the same object in different environments or scenarios. Without such diverse representations, models struggle to generalize under varying real-world conditions. To overcome this limitation, we introduce CBDG, which first increases the number of categories and then leverages SAM Kirillov et al. (2023) and the Stable Diffusion model Rombach et al. (2022) to generate data pairs across different scenes while ensuring consistent foreground objects, thus improving both category variation and background diversity in training data.

Our experimental results demonstrate significant improvements on two challenging benchmarks, D^3 Xie et al. (2023) and OmniLabel Schulter et al. (2023), achieving +16.3 AP on OmniLabel and +14.9 AP on D^3. The proposed CBDG and CCLoss are complementary components that collectively form a robust training paradigm. Specifically, the CBDG improves feature learning through diverse scene-object compositions, while the CCLoss ensures robust feature representation across varying backgrounds. Furthermore, our approach is fundamentally model-agnostic, enabling seamless integration into a wide range of existing architectures, such as Dou et al. (2022); Li et al. (2022b), with consistent performance gains across different frameworks.

In summary, the contributions are as follows.

- This study identifies an issue where object features are highly susceptible to environmental changes, leading to potential overfitting and poor generalization to unseen scenarios.
- To ensure feature robustness to context changes, we propose CCL, which enforces object consistency across backgrounds via CBDG and CCLoss.
- Our method is simple, efficient, and model-agnostic, imposing no additional inference overhead while consistently delivering performance improvements across diverse datasets and models. Moreover, despite working with a much smaller subset of the original dataset, we achieve state-of-the-art results on two descriptive open-vocabulary detection benchmarks.

2 Related work

Vision language localization tasks. Open-vocabulary object detection (OVOD) aims to enable models to recognize novel objects or unseen categories during inference Gu et al. (2021); Minderer et al. (2023); Zareian et al. (2021); Du et al. (2022) , extending beyond traditional categorical detection. However, this ability is typically limited to detecting object categories based on labels, rather than understanding long descriptions. In contrast, referring expression comprehension (REC)

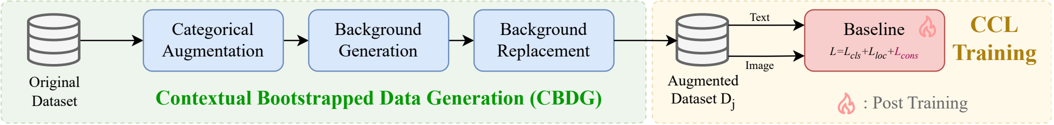

Figure 2: Overview of our approach. CBDG generates D_j via Categorical Augmentation, Background Generation and Background Replacement. CCL training uses D_j with CCLoss added to total loss.

involves understanding and localizing objects in an image based on natural language descriptions that refer to specific instances of objects Yu et al. (2016); Wu et al. (2020); Mao et al. (2016), making it inherently more flexible and context-aware. While OVOD and REC both address the challenge of understanding objects in images, we focus on simultaneously handling novel categories and complex natural language descriptions. We opt for described object detection (DOD) Xie et al. (2023) and OmniLabel Schulter et al. (2023) as robust solutions to these challenges, as they incorporate both the recognition of novel categories and the understanding of intricate descriptions.

Diffusion models for scenario generation. Stable Diffusion Rombach et al. (2022) marks a shift in text-to-image synthesis by operating in a compressed latent space using iterative denoising. Unlike GANs Goodfellow et al. (2014); Mirza & Osindero (2014) or VAEs Kingma et al. (2013); Van Den Oord et al. (2017) that generate images in pixel space, it uses a VAE to encode images into low-dimensional latents, allowing efficient training and high-resolution output. Guided by a pre-trained CLIP text encoder, the model aligns generated images with complex textual descriptions, from concrete objects to abstract scenes.

Recent diffusion-based methods have shown strong performance in image inpainting, enabling object and scene editing via text or spatial inputs. GLIDE Nichol et al. (2021) enables text-guided object replacement while preserving scene consistency, and GLIGEN Li et al. (2023b) extends this by incorporating bounding boxes for more precise control over object placement. For background replacement, IAM Yu et al. (2023) integrates segmentation with diffusion to regenerate regions based on textual prompts. Despite their effectiveness, these methods often suffer from boundary artifacts due to over-smoothing during denoising. We evaluate GLIDE, IAM, and Stable Diffusion Rombach et al. (2022) in CBDG and ultimately choose Stable Diffusion for background generation.

Cross-modal object detection models. With the advancement of multimodal vision language models, such as CLIP Radford et al. (2021) and ALIGN Jia et al. (2021), the development of methods that integrate vision and language to address visual recognition tasks has emerged as a prominent trend. GLIP Li et al. (2022b), based on CLIP Radford et al. (2021), leverages free-form language supervision during training and frames object detection as visual localization, constructing a foundation for semantically enriched pre-trained models. Building on this, FIBER Dou et al. (2022) employs a two-stage training approach, transitioning from coarse-grained to fine-grained, enhancing the adaptability of the pre-trained model to a broad spectrum of downstream tasks at both image-level and region-level. In our work, we use GLIP Li et al. (2022b) and FIBER Dou et al. (2022) as baseline models and incorporate our CCL method to validate the experimental results.

3 Method

3.1 Overview

We introduce CCL, a novel framework designed to address the challenge of maintaining detection and grounding consistency when models encounter diverse and unseen object categories across varying contextual backgrounds. To achieve this goal, we address two fundamental aspects of the problem: the lack of appropriate training data and the need for effective consistency-preserving mechanisms.

In Section 3.2, we describe our CBDG pipeline, which leverages advanced segmentation and generative models to create a rich and varied dataset. This data preparation process is specifically designed to support our consistency learning objectives. Following this, in Section 3.3, we detail our CCLoss formulation, which ensures that the model learns to maintain object identity across different backgrounds.

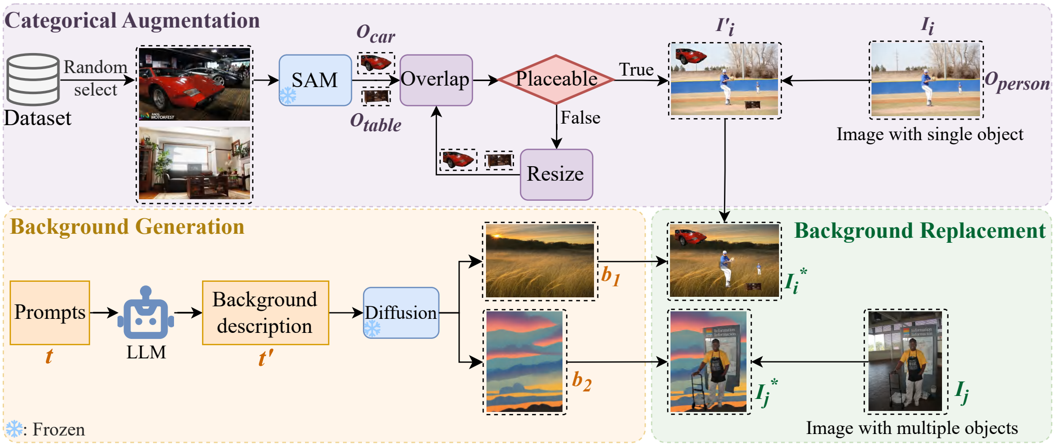

Figure 3: CBDG Pipeline. We use ChatGPT to generate background prompts for a diffusion model, enabling diverse background synthesis. For single-class images, CBDG augments object categories before background replacement. For multi-class images, CBDG replaces only the background.

3.2 Contextual Bootstrapped Data Generation

Current open-set visual grounding methods struggle to maintain robustness across diverse real-world scenarios, particularly when objects appear in unfamiliar contextual settings. This limitation stems from a fundamental data scarcity: existing datasets rarely capture the full spectrum of object-background interactions, leading to biased model performance. To overcome this, we propose a multistage data augmentation framework that synthesizes diverse and realistic object-context compositions by combining SAM-based object manipulation with text-guided background generation, as shown in Figure 3. Our method constructs a compositionally diverse joint dataset D_j that mitigates common inpainting artifacts and improves model generalization.

Categorical augmentation. In our approach, we use the Flickr30k Entities visual grounding dataset Plummer et al. (2015) alongside a subset of the Objects365 object detection dataset Shao et al. (2019) to create a combined training dataset. The selected subset of Objects365 tends to contain images dominated by a few categories, with multiple instances of that category present. Details are discussed in Supplementary Section D.3 and Section D.5.

For images with a single object, we aim to enhance the diversity of object categories within each image by introducing objects from different categories while maintaining spatial and contextual coherence. To achieve this, we leverage the SAM model Kirillov et al. (2023) to extract precise objects O_i and the corresponding position (x_o, y_o) for each image I, where i means the category ID to which the object belongs. The object masks allow us to identify individual objects and their spatial locations. Based on these masks, we randomly select objects $O_{i\notin C}$ from other images within the same subset but belonging to different categories, where C represents the category set. Then these objects are positioned in the current image at carefully chosen locations. Specifically, we define $P = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$ the potential placement position set for the new object, N is the number of candidate locations. From these positions, we randomly select $(x, y) \in P\backslash(x_o, y_o)$ that does not overlap with the existing objects in the image, ensuring a clean and non-interfering insertion of the new object. This process of placement can be formalized as:

$$\texttt{Augmentation} : (x_{o_k}, y_{o_k}) \in P\backslash(x_o, y_o), o_k \in O_{i\notin C}, \tag{1}$$

where k represents the category of selected object, (x_{o_k}, y_{o_k}) denotes the position randomly chosen from the set of candidate locations according to the above rule. After categorical augmentation, the original image I becomes I'.

In scenarios where no suitable empty position is available, such as when the current image contains large objects or a large number of dispersed objects, which results in limited available space, we adopt a resizing strategy. In these cases, we reduce the size of the new object to $1/\alpha$ of its original size and attempt to place it again, where α is a scaling factor. This process is repeated until an empty

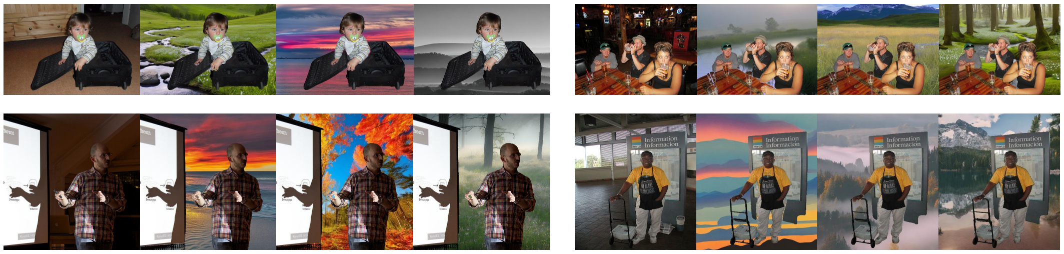

Figure 4: Four groups of images are shown, each composed of four sub-images: the leftmost sub-image in every group is the original, while the remaining three display background replacements.

placement area is found or the number of resizing attempts exceeds a threshold N_R. If resizing attempts fail to find a suitable location, we abandon the current image and instead select another image to enhance the diversity of object categories, thus ensuring a broader range of category representation in the final dataset.

Background generation. With more object categories added, the foreground dataset now includes images with varied labels and their corresponding bounding boxes. To reduce model overfitting and improve generalization, we next generate diverse background images, placing the same objects in different scenes. Image inpainting methods Nichol et al. (2021); Li et al. (2023b); Yu et al. (2023) often struggle with blurred edges and backgrounds that still reflect foreground features, making realistic scene changes difficult (see Supplementary Section B.6). To avoid these issues, we use a simpler alternative that better separates foreground from background.

Instead of relying on limited original image content, we generate new and simple backgrounds directly. Using a Large Language Model (LLM) Brown et al. (2020), denoted as $\mathcal{G}$, we create text prompts in three categories: *Seasonal, Sky, and Natural Landscape*, to ensure variety and relevance. Details of these prompts are provided in Supplementary Section D.1. These prompts are input into Stable Diffusion $\mathcal{D}$ Rombach et al. (2022), which generates matching background images. This method allows us to build a diverse, context-aware background dataset without the limitations of inpainting:

$$\texttt{Generation}: b = \mathcal{D}(t'), t' = \mathcal{G}(t), \tag{2}$$

where $t \in \{$*Seasonal, Sky, Natural Landscape*$\}$, t' is the background description generated by ChatGPT, and b represents the background image generated by stable Diffusion, which constitutes the background dataset D_{bg}. We further analyze the diversity of generated scenes in Supplementary Section B.4 and Section D.4.

Background replacement. At this stage, we have both the generated background dataset D_{bg} and the foreground dataset D_{fg} with various object categories. To create new image variations, we randomly select background images for each foreground image. Using the bounding boxes, we extract foreground objects with the SAM model $\mathcal{S}$ Kirillov et al. (2023). The post-processing techniques are detailed in Supplementary Section D.2. These objects and their spatial layout are kept unchanged. After isolating the foreground, we replace the original background with a selected one, generating multiple new images per original. The foreground stays the same, while the backgrounds vary, producing diverse scenes with consistent object content. The replacement process is defined as:

$$\texttt{Replacement}: I^* = \mathcal{S}(I', bbox) \oplus b, b \in D_{bg}, \tag{3}$$

where $bbox$ represents the bounding boxes of objects in the image I' after categorical augmentation, $\oplus$ denotes the composition of foreground and background, I^* represents the image with replaced background.

CBDG enables us to significantly augment the dataset with diverse background settings while maintaining the integrity of the foreground objects, providing a more robust foundation for training our model. As shown in Figure 4, after CBDG, for each original image, several additional images are generated with replaced backgrounds. This results in a total of K images per original, all sharing the same foreground objects, but differing in their backgrounds. K represents the batch size used during training. Alternative data generation schemes are also compared, see Supplementary Section B.3. The augmented images are then utilized for the subsequent consistency constraints in our approach.

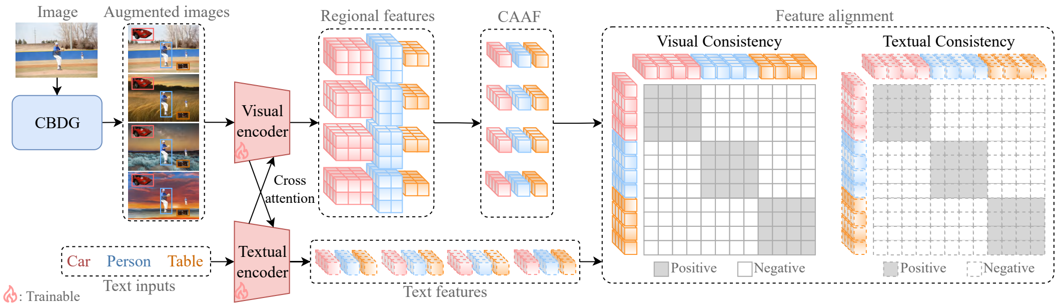

Figure 5: CCL Framework. Visual and textual features are encoded, with regional features pooled into CAAF. Consistency loss is applied within each modality.

3.3 Contextual Consistency Loss

Given that after CBDG, we now have access to a dataset D_j in which each group of images shares the same foreground object but varies in background. We introduce the Contextual Consistency Loss (CCLoss), a novel training objective designed to enforce representation invariance for the same object category across varying contextual environments. As illustrated in Figure 5, our method uses CCLoss to maintain the consistency of foreground object representations across different backgrounds. By constructing training batches that contain instances of the same object under different contextual settings, CCLoss encourages the model to focus on semantically meaningful foreground features rather than background-dependent or spurious cues. This section elaborates on the underlying model architecture, the detailed formulation of consistency loss, and its integration into the overall training objective.

Model architecture. We employ a language-based object detector Li et al. (2022b); Dou et al. (2022) as the backbone for feature extraction and object detection, taking advantage of its strong capability in bridging vision and language representations. Specifically, images and textual descriptions are first encoded to obtain their respective feature embeddings, ensuring a comprehensive understanding of both modalities. These extracted features are subsequently processed through a Feature Pyramid Network (FPN), which effectively refines and integrates multiscale representations, thereby enhancing detection performance across various object sizes and contexts. To further improve localization accuracy, the refined image features are then passed on to DynamicHead, a dedicated module designed to predict a set of candidate regions where objects are most likely to be located. This hierarchical and adaptive processing pipeline ensures robust and efficient object detection.

Consistency loss. During the training phase, we organize each batch by grouping images that share identical foreground objects but exhibit diverse background settings. This arrangement enables the computation of the CCLoss function, which serves as a critical mechanism for training the model to preserve invariant representations of foreground objects across varying contextual environments.

The total loss function, as depicted in Eq. 4, comprises three fundamental components: localization loss, classification loss, and contextual consistency loss ($\mathcal{L}_{\text{cons}}$). Each of these components contributes to optimizing the model performance in different aspects: precise object localization, accurate category classification, and robust feature representation that maintains foreground consistency irrespective of background variations. The first two components of the loss function are detailed in GLIP Li et al. (2022b). The integration of these loss terms ensures a balanced optimization process that addresses discriminative and invariant feature learning.

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{loc}} + \mathcal{L}_{\text{cons}}, \tag{4}$$

Eq. 5 provides the formulation of CCLoss. CCLoss combines the text and image modality losses with weighting factors. λ_{T} and λ_{I} are weighting parameters to balance the loss contributions in the text and the image modality.

$$\mathcal{L}_{\text{cons}} = \lambda_{\text{T}} \cdot \mathcal{L}_{\text{T}} + \lambda_{\text{I}} \cdot \mathcal{L}_{\text{I}}, \tag{5}$$

For image features obtained from the image encoder, we first perform a pooling operation on them to obtain the Context-Aware Aggregated Feature (CAAF), denoted as f, followed by applying a consistency loss among the CAAF. Given a batch with C categories and K images, the contrastive loss for the vision modality is defined as:

$$\mathcal{L}_{\mathrm{I}} = -\frac{1}{CK}\sum_{c=1}^{C}\sum_{k=1}^{K}\log\frac{\exp\bigl(\mathrm{sim}(\mathbf{f}_{ck},\mathbf{f}_{c})/\tau\bigr)}{\sum_{c'=1}^{C}\sum_{k'=1}^{K}\exp\bigl(\mathrm{sim}(\mathbf{f}_{ck},\mathbf{f}_{c'k'})/\tau\bigr)}, \tag{6}$$

where $\mathbf{f}_{ck}$ is the k-th image feature of the c-th category. $\mathbf{f}_{c'k'}$ is the k'-th image feature of the c'-th category. $\mathbf{f}_c$ is the centroid of the image features for the c-th category, calculated as the mean of the K image features. $\mathrm{sim}(\cdot,\cdot)$ is cosine similarity. τ is the temperature parameter.

Similarly, for text features, we implement a contrastive learning objective that promotes feature clustering within the same category while enforcing separation among different categories. However, the application of this text contrastive loss is contingent upon the baseline architecture: When using FIBER as the baseline, where cross-modal interactions between image and text encoders are enabled, we fully utilize this loss term. In contrast, when employing GLIP as the baseline, which processes image and text modalities independently, we effectively disable this component by setting its weight λ_{T} to zero. Given a batch with C categories and K images, the contrastive loss for the text modality is defined as:

$$\mathcal{L}_{\mathrm{T}} = -\frac{1}{CK}\sum_{c=1}^{C}\sum_{k=1}^{K}\log\frac{\exp\bigl(\mathrm{sim}(\mathbf{t}_{ck},\mathbf{t}_{c})/\tau\bigr)}{\sum_{c'=1}^{C}\sum_{k'=1}^{K}\exp\bigl(\mathrm{sim}(\mathbf{t}_{ck},\mathbf{t}_{c'k'})/\tau\bigr)}, \tag{7}$$

where $\mathbf{t}_{ck}$ is the k-th text feature of the c-th category. $\mathbf{t}_{c'k'}$ is the k'-th text feature of the c'-th category. $\mathbf{t}_c$ is the centroid of the text features for the c-th category, calculated as the mean of the K text features. The design of our CCLoss follows a progressive evolution, with the detailed process provided in Supplementary Section B.7.

4 EXPERIMENTS

4.1 EXPERIMENTAL DESIGN

Training setup. To evaluate the generalizability of our proposed method, we use two baseline models, GLIP Li et al. (2022b) and FIBER Dou et al. (2022). These models serve as benchmarks for comparison. The datasets used to train the baseline models are 1) Objects365 (O365) Shao et al. (2019) and 2) GoldG, including Flickr30K Plummer et al. (2015), VG Caption Krishna et al. (2017), and GQA Hudson & Manning (2019), which together contain 0.8 million images, providing a diverse and large-scale training set.

In contrast, for our method, we work with a smaller subset of the original dataset, with only 0.25 million images as the initial joint dataset for CBDG. Specifically, we incorporate the Flickr30k Entities Plummer et al. (2015) dataset along with only 0.22M images of the Objects365 dataset Shao et al. (2019), which is much smaller than the full dataset used for the baselines.

We generate three main categories of background images in CBDG: seasonal, sky, and natural landscape. In total, we have 13,185 unique descriptions, resulting in 144,654 generated images. The breakdown of categories and the corresponding number of images is as follows: seasonal (3387 descriptions, 48,156 images), sky (3399 descriptions, 48,210 images), and natural landscape (3399 descriptions, 48,288 images).

For training, we use publicly available pre-trained model checkpoints of both GLIP Li et al. (2022b) and FIBER Dou et al. (2022). These pre-trained weights serve as the starting point for fine-tuning. We fine-tune the model for one epoch on our dataset D_j. After this fine-tuning process, we obtain the final results, which demonstrate the effectiveness of our method when applied to a smaller and more constrained dataset. The implementation details and computational cost can be found in Supplementary Section A. We report the choice and tuning of hyperparameters in Supplementary Section B.1.

Benchmark selection. We choose OmniLabel Schulter et al. (2023) and D^3 Xie et al. (2023) as benchmark evaluation methods, both of which use Average Precision (AP) as the evaluation

Table 1: Performance of our method compared with SOTA methods.

	OmniLabel							D^3		
Method	AP	AP-c	AP-d	AP-dP	AP-dS	AP-dM	AP-dL	FULL	PRES	ABS
Detic Zhou et al. (2022)	8.0	15.6	5.4	8.0	5.7	5.4	6.2	-	-	-
OFA-DOD Xie et al. (2023)	-	-	-	-	-	-	-	21.6	23.7	15.4
RelationLLM-L Xie et al. (2025)	-	-	-	-	-	-	-	24.3	24.6	23.4
GN-GLIP Zhao et al. (2024)	22.2	27.2	18.8	29.0	-	-	-	21.4	20.6	23.7
GN-FIBER Zhao et al. (2024)	28.1	32.1	25.1	36.5	-	-	-	26.0	25.2	28.1
ROD-MLLM Yin et al. (2025)	-	-	25.3	30.9	31.8	24.5	21.0	29.7	30.0	28.7
Real-Model Chen et al. (2025)	-	-	36.5	**52.1**	**54.4**	33.2	25.5	34.1	34.4	33.2
GLIP-T Li et al. (2022b)	19.3	23.6	16.4	25.8	29.4	14.8	8.2	19.1	18.3	21.5
+ours	32.2	36.1	28.8	39.8	43.3	26.5	17.6	30.0	29.2	32.3
FIBER-B Dou et al. (2022)	25.7	30.3	22.3	34.8	38.6	19.5	12.4	22.7	21.5	26.0
+ours	**42.0**	**44.1**	**39.2**	50.8	53.7	**38.2**	**32.3**	**37.6**	**37.2**	**38.8**

metric. The reason we select these two benchmarks is that they not only provide object category labels but also include a rich diversity of textual descriptions, which place a greater emphasis on the model's ability to understand and interpret language. This aspect makes these benchmarks particularly valuable for evaluating the model's performance in tasks involving both visual and linguistic information. Compared to other REC Yu et al. (2016); Wu et al. (2020); Mao et al. (2016) and OVOD Gupta et al. (2019); Chen et al. (2015); Krasin et al. (2017) benchmarks, D^3 Xie et al. (2023) and OmniLabel Schulter et al. (2023) offer a broader evaluation of object detection capabilities. These benchmarks include negative samples and more precisely defined bounding boxes corresponding to textual descriptions, which can refer to zero, one, or multiple objects in the image. This makes the tasks more challenging and forces the model to effectively localize and recognize objects based on a range of different descriptions and contexts, offering a more comprehensive test of its generalization and performance in diverse scenarios.

4.2 Comparison with SOTA Methods

Table 1 presents a comparison between our method and the current SOTA methods on the OmniLabel Schulter et al. (2023) and D^3 Xie et al. (2023) benchmarks. The first column lists various model methods, followed by seven columns representing the seven AP metrics on OmniLabel. These metrics include: plain categories (AP-c) and free-form descriptions (AP-d). AP-dP evaluates only positive descriptions. AP-dS/M/L assess descriptions of varying lengths (up to 3 words, 4-8 words, and more than 8 words). The last three columns represent the AP metrics on D^3: FULL, PRES, and ABS, which evaluate all descriptions, only presence descriptions, and only absence descriptions, respectively.

We use GLIP-T Li et al. (2022b) and FIBER-B Dou et al. (2022) as baselines and fine-tune them on our method. With the integration of our proposed CCL method, significant improvements are observed across multiple benchmarks. Specifically, when applied to the FIBER baseline, the method achieves a notable increase of +16.3 AP on the OmniLabel benchmark and +14.9 AP on the D^3 benchmark. Similarly, when implemented with the GLIP baseline, our method demonstrates consistent performance gains, achieving +12.9 AP on the OmniLabel benchmark and +10.9 AP on the D^3 benchmark. These results underscore the effectiveness of our approach in improving contextual understanding and consistency across diverse datasets. We further evaluate our method on phrase grounding tasks to demonstrate broader applicability (see Supplementary Section B.5).

4.3 Robustness Evaluation under Background Variations

To quantitatively assess the robustness of open-vocabulary detection (OVD) models under environmental and background variations, we introduce a new experiment setup derived from the D^3 dataset. For each of the 10,578 original images in D^3, we generate three additional variants by replacing the background using the CBDG method proposed in this work. These new background images are generated independently of the training data, ensuring no overlap or information leakage. The result-

Table 2: Performance comparison on $D^3{}_{\text{BC}}$ benchmark across different models and settings.

Method	$D^3{}_{\text{BC}}$		
	FULL	PRES	ABS
GLIP-T	16.8	15.8	20.4
+ours	29.6	28.9	31.9
FIBER-B	20.1	18.7	24.7
+ours	33.1	32.8	34.0

Table 3: Ablation study of contextual bootstrapped data generation and CCLoss.

Method	OmniLabel			D^3		
	AP	AP-c	AP-d	FULL	PRES	ABS
GLIP-T	19.3	23.6	16.4	19.1	18.3	21.5
+data	24.8	29.2	21.8	23.2	22.5	25.3
+ours	32.2	36.1	28.8	30.0	29.2	32.3
FIBER-B	25.7	30.3	22.3	22.7	21.5	26.0
+data	32.7	35.8	29.6	29.1	28.3	31.2
+ours	42.0	44.1	39.2	37.6	37.2	38.8

ing dataset, termed $D^3{}_{\text{BC}}$, consists of the original images and their background-altered counterparts, totaling 42,312 samples. We evaluate two representative baseline models, GLIP-T and FIBER-B, on both D^3 and $D^3{}_{\text{BC}}$, and further examine their performance when enhanced with our proposed CCL method. This yields four experimental settings. As summarized in Table 2, both baselines exhibit substantial performance degradation on $D^3{}_{\text{BC}}$, revealing their susceptibility to background shifts. However, models incorporating our CCL approach demonstrate significantly improved robustness with much smaller performance drops. These results highlight the effectiveness of CCL in improving model resilience to environmental variations. Moreover, this experiment suggests that our method maintains robustness not only under background shifts but also across different domains.

4.4 Ablation Study on CBDG and CCLoss

Given that our method is fundamentally grounded in consistency and incorporates a certain degree of data generation, we perform a series of ablation experiments to evaluate the contribution of each individual component. In particular, we conduct two distinct experimental setups to assess the impact of both CBDG and CCLoss. The first experiment introduces CBDG to the baseline model, followed by fine-tuning the model for one epoch on D_j. To ensure a fair comparison, we keep the training parameters consistent with those used in the baseline experiment. The second experiment represents our complete experimental setup, adding CBDG and CCLoss to the baseline model and fine-tuning the model for one epoch. As shown in Table 3, both CBDG and CCLoss play an essential role in enhancing the model's performance. CBDG increases the diversity of training data, improving the model's robustness across varying conditions. Meanwhile, the CCLoss reinforces object consistency across different contexts, ensuring that the model can reliably detect and localize objects regardless of their surrounding environment. The combined effects of these two components contribute significantly to the observed performance improvements. We further analyze the impact of dataset scale in Supplementary Section B.2.

5 Conclusion

Summary. We propose Contextual Consistency Learning (CCL) to tackle inconsistent object feature representation in descriptive open-vocabulary object detection. CCL combines Contextual Bootstrapped Data Generation (CBDG) and Contextual Consistency Loss (CCLoss). CBDG uses SAM and Stable Diffusion to generate diverse scene-object compositions, while CCLoss enforces feature invariance across backgrounds. Despite using significantly less data, CCL improves model performance. It is model-agnostic, incurs no inference overhead, and integrates easily into existing architectures. Our work underscores the importance of intra-modal consistency for robust object detection in dynamic environments, paving the way for future extensions to broader vision-language tasks and large-scale models.

Limitation & Future work. Due to the inherent limitations of SAM, segmentation errors or under-segmentation may occur when extracting foreground objects in our CBDG. Although our post-processing techniques effectively mitigate these issues and achieve SOTA performance, further research is needed to completely eliminate such problems and further enhance performance.

Ethics Statement

This work does not involve human subjects, private or sensitive data, or applications that may cause direct harm. All datasets used in this study are publicly available and widely adopted in the research community. Our method focuses on improving robustness in object detection by generating synthetic background variations, which does not introduce new ethical concerns. We have taken care to avoid reinforcing social biases, and the proposed framework is intended solely for academic research.

Reproducibility Statement

We have made significant efforts to ensure the reproducibility of our results. All implementation details, including network architectures, training schedules, and hyperparameters, are described in the main paper and supplement. The datasets used are publicly available, and we provide detailed descriptions of preprocessing steps in the supplementary materials. Furthermore, the theoretical formulation of our loss function is fully detailed in Supplementary Section B.7, with proofs and additional derivations provided in the supplement. As mentioned in the abstract, data, code and models will be made publicly available.

Use of Large Language Models (LLMs)

We use large language models (LLMs) solely to assist with the polishing of English writing, such as improving grammar, clarity, and readability. In addition, using a Large Language Model (LLM) Brown et al. (2020), we generate prompts that are subsequently fed into Stable Diffusion Rombach et al. (2022) to synthesize background images for our experiments. No part of the research design, experimental implementation, data analysis, or result interpretation relied on LLMs. All scientific contributions, ideas, and experiments are conceived and conducted entirely by the authors.

References

Malik Javed Akhtar, Rabbia Mahum, Faisal Shafique Butt, Rashid Amin, Ahmed M El-Sherbeeny, Seongkwan Mark Lee, and Sarang Shaikh. A robust framework for object detection in a traffic surveillance system. *Electronics*, 11(21):3425, 2022.

Codruta O Ancuti, Cosmin Ancuti, and Radu Timofte. Nh-haze: An image dehazing benchmark with non-homogeneous hazy and haze-free images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 444–445, 2020.

Muhammad Awais, Weiming Zhuang, Lingjuan Lyu, and Sung-Ho Bae. Frod: Robust object detection for free. *CoRR*, 2023.

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

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Yuming Chen, Jiangyan Feng, Haodong Zhang, Lijun Gong, Feng Zhu, Rui Zhao, Qibin Hou, Ming-Ming Cheng, and Yibing Song. Re-aligning language to visual objects with an agentic workflow. *arXiv preprint arXiv:2503.23508*, 2025.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Yuexiong Ding, Ming Zhang, Jia Pan, Jinxing Hu, and Xiaowei Luo. Robust object detection in extreme construction conditions. *Automation in Construction*, 165:105487, 2024.

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

Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14084–14093, 2022.

Qi Fan, Mattia Segu, Yu-Wing Tai, Fisher Yu, Chi-Keung Tang, Bernt Schiele, and Dengxin Dai. Towards robust object detection invariant to real-world domain shifts. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*. OpenReview, 2023.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Stephen Gould, Richard Fulton, and Daphne Koller. Decomposing a scene into geometric and semantically consistent regions. In *2009 IEEE 12th international conference on computer vision*, pp. 1–8. IEEE, 2009.

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

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

Drew A Hudson and Christopher D Manning. 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*, pp. 6700–6709, 2019.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Sheng Jin, Xueying Jiang, Jiaxing Huang, Lewei Lu, and Shijian Lu. Llms meet vlms: Boost open vocabulary object detection with fine-grained descriptors. *arXiv preprint arXiv:2402.04630*, 2024.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1780–1790, 2021.

Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2(3):18, 2017.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-vlm: Open-vocabulary object detection upon frozen vision and language models. *arXiv preprint arXiv:2209.15639*, 2022.

Yejun Lee and Jaejun Yoo. Improving contrail detection via diffusion-based data augmentation framework. *age*, 8000:12000, 2025.

Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE transactions on image processing*, 28(1): 492–505, 2018.

Jizhizi Li, Jing Zhang, Stephen J Maybank, and Dacheng Tao. Bridging composite and real: towards end-to-end deep image matting. *International Journal of Computer Vision*, 130(2):246–266, 2022a.

Liunian Li, Zi-Yi Dou, Nanyun Peng, and Kai-Wei Chang. Desco: Learning object recognition with rich language descriptions. *Advances in Neural Information Processing Systems*, 36:37511–37526, 2023a.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10965–10975, 2022b.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22511–22521, 2023b.

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

Chuang Lin, Yi Jiang, Lizhen Qu, Zehuan Yuan, and Jianfei Cai. Generative region-language pretraining for open-ended object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13958–13968, 2024.

Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023.

Chengxin Liu, Kewei Wang, Hao Lu, Zhiguo Cao, and Ziming Zhang. Robust object detection with inaccurate bounding boxes. In *European Conference on Computer Vision*, pp. 53–69. Springer, 2022.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Nesma Talaat Abbas Mahmoud, Indrek Virro, AGM Zaman, Tormi Lillerand, Wai Tik Chan, Olga Liivapuu, Kallol Roy, and Jüri Olt. Robust object detection under smooth perturbations in precision agriculture. *AgriEngineering*, 6(4):4570–4584, 2024.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 11–20, 2016.

Ze-Yu Mi and Yu-Bin Yang. Add: Attribution-driven data augmentation framework for boosting image super-resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23101–23110, 2025.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European conference on computer vision*, pp. 728–755. Springer, 2022.

Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36:72983–73007, 2023.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649, 2015.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

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

Samuel Schulter, Yumin Suh, Konstantinos M Dafnis, Zhixing Zhang, Shiyu Zhao, Dimitris Metaxas, et al. Omnilabel: A challenging benchmark for language-based object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11953–11962, 2023.

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

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.

Sanjay Subramanian, William Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension. *arXiv preprint arXiv:2204.05991*, 2022.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Anh-Khoa Nguyen Vu, Quoc-Truong Truong, Vinh-Tiep Nguyen, Thanh Duc Ngo, Thanh-Toan Do, and Tam V Nguyen. Multi-perspective data augmentation for few-shot object detection. *arXiv preprint arXiv:2502.18195*, 2025.

Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. Phrasecut: Language-based image segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10216–10225, 2020.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3485–3492. IEEE, 2010.

Chi Xie, Zhao Zhang, Yixuan Wu, Feng Zhu, Rui Zhao, and Shuang Liang. Described object detection: Liberating object detection with flexible expressions. *Advances in Neural Information Processing Systems*, 36:79095–79107, 2023.

Chi Xie, Shuang Liang, Jie Li, Zhao Zhang, Feng Zhu, Rui Zhao, and Yichen Wei. Relationlmm: Large multimodal model as open and versatile visual relationship generalist. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Shiyu Xuan, Qingpei Guo, Ming Yang, and Shiliang Zhang. Pink: Unveiling the power of referential comprehension for multi-modal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13838–13848, 2024.

Heng Yin, Yuqiang Ren, Ke Yan, Shouhong Ding, and Yongtao Hao. Rod-mllm: Towards more reliable object detection in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14358–14368, 2025.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 69–85. Springer, 2016.

Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6023–6032, 2019.

Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. Contextual object detection with multimodal large language models. *International Journal of Computer Vision*, pp. 1–19, 2024.

Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14393–14402, 2021.

Yufei Zhan, Yousong Zhu, Zhiyang Chen, Fan Yang, Ming Tang, and Jinqiao Wang. Griffon: Spelling out all object locations at any granularity with large language models. In *European Conference on Computer Vision*, pp. 405–422. Springer, 2024.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Shiyu Zhao, Zhixing Zhang, Samuel Schulter, Long Zhao, BG Vijay Kumar, Anastasis Stathopoulos, Manmohan Chandraker, and Dimitris N Metaxas. Exploiting unlabeled data with vision and language models for object detection. In *European conference on computer vision*, pp. 159–175. Springer, 2022.

Shiyu Zhao, Long Zhao, Yumin Suh, Dimitris N Metaxas, Manmohan Chandraker, Samuel Schulter, et al. Generating enhanced negatives for training language-based object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13592–13602, 2024.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European conference on computer vision*, pp. 350–368. Springer, 2022.

Yongshuo Zong, Qin Zhang, Dongsheng An, Zhihua Li, Xiang Xu, Linghan Xu, Zhuowen Tu, Yifan Xing, and Onkar Dabeer. Ground-v: Teaching vlms to ground complex instructions in pixels. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24635–24645, 2025.