arXiv:2607.11008v1 [cs.CV] 13 Jul 2026

SynCLIP: Synonym-Coherent Language-Image Pretraining for Robust Open-Vocabulary Dense Perception

Mingjie Xie[1] Guangjun He[2*] Dongli Xu[3] Youtian Lin[4] Hongjue Li[1]
Pengming Feng[2] Jian Guan[5] Yue Deng[1,6]
[1]Beihang University [2]State Key Laboratory of Space Information System and Integrated Application
[3]Independent Researcher [4]Nanjing University [5]Harbin Engineering University [6]Beijing Zhongguancun Academy

Abstract

Open-vocabulary dense perception (OVDP) aims to localize objects unseen during training by leveraging textual knowledge. Despite the remarkable progress of recent CLIP-based approaches, we identify a critical limitation: synonym-induced grounding inconsistency, where semantically equivalent expressions yield disparate spatial attention patterns. This inconsistency undermines the robustness and performance of existing methods in real-world OVDP applications. To address this issue, we propose SynCLIP, a Synonym-Coherent Language-Image Pretraining framework that enhances synonym-robust grounding for OVDP. SynCLIP introduces a Semantic-consistent Spatial Attention alignment (SSA) module to enhance spatial attention consistency by minimizing discrepancies between attention maps of original and synonymous expressions. Furthermore, a Spatial Attention Refinement (SAR) module selectively strengthens the most semantically relevant spatial regions within aligned maps for more precise and stable grounding. To support synonym-coherent pretraining, we also construct a Synonym-Enriched Visual Corpus (SEViC), which augments each category with multiple synonyms and textual definitions. Extensive experiments on multiple benchmarks demonstrate that SynCLIP substantially improves grounding consistency under diverse linguistic variants and achieves state-of-the-art performance among CLIP-based OVDP methods. Code is available at https://github.com/Justlovesmile/SynCLIP.

1. Introduction

Open-vocabulary dense perception (OVDP) aims to recognize and localize objects from novel categories that are not predefined in the training set by leveraging textual knowledge [46]. Unlike traditional dense perception tasks such as object detection and segmentation [7, 10], which operate within a fixed label space and thus struggle to generalize to real-world environments with an open set of object categories [31, 35], OVDP employs textual expressions to flexibly represent labels. This language-driven supervision enables OVDP models to generalize beyond the closed-set categories when handling unseen categories. Consequently, OVDP has attracted growing attention for deployment in real-world applications such as robotics and autonomous driving [42, 45].

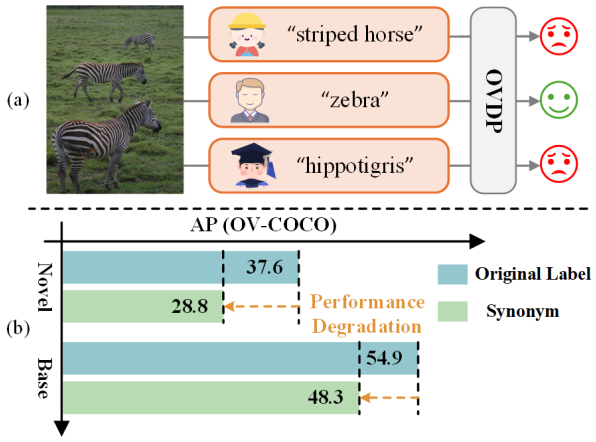


Figure 1. Illustration of synonym-induced grounding inconsistency. (a) shows inconsistent dense perception across diverse synonymous expressions. (b) presents performance degradation caused by synonym-induced grounding inconsistency. Here, 'Novel' and 'Base' denote the performance on unseen and seen categories, respectively.

To enable open-vocabulary capability, recent studies [4, 38, 43] commonly leverage vision-language models (VLMs) pretrained on large-scale image–text pairs, such as CLIP [17, 24, 26]. These methods transfer the global vision–language alignment learned by VLMs to region-level representations, allowing visual regions to be associated with category-related textual expressions and thereby facilitating the recognition and localization of unseen categories. For example, a series of approaches [30, 33, 34] enhance

*Corresponding author.

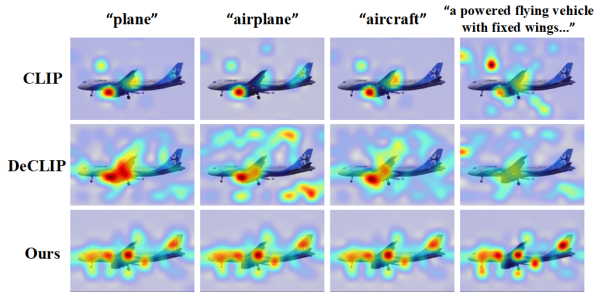


Figure 2. Comparison of spatial attention distributions generated by different models for synonymous expressions, *e.g.*, synonyms and definition. Our method yields more consistent and semantically aligned attention, indicating synonym-robust grounding.

region-level vision–language alignment by employing techniques such as pseudo-region self-distillation or guidance from vision foundation models (VFMs). These strategies have led to notable improvements in OVDP tasks, including open-vocabulary detection (OVD) and segmentation (OVS).

Despite these advances, existing CLIP-based OVDP methods overlook a critical issue, *i.e., synonym-induced grounding inconsistency*. When diverse synonyms or semantically equivalent expressions are used to describe the same object, they often yield inconsistent spatial attention distributions. Such variations reduce model robustness and cause performance degradation in real-world OVDP applications. As shown in Figure 1, simply replacing the original label (*e.g.*, *"zebra"*) with synonymous expressions (*e.g.*, *"striped horse"* and *"hippotigris"*) results in a substantial drop in Average Precision (AP) for F-ViT [33] on the OV-COCO benchmark [20]. This observation reveals the severity of synonym-induced grounding inconsistency and underscores the necessity for a synonym-coherent approach for real-world applications.

Figure 2 further illustrates this issue through qualitative comparisons of several CLIP-based methods [24, 30]. For the same image, different synonymous terms (*i.e.*, *"plane"*, *"airplane"* and *"aircraft"*) and semantically equivalent expressions (*i.e.*, *"a powered flying vehicle with fixed wings ..."*) yield markedly different spatial attention distributions, despite referring to the same visual object. Consequently, the model is guided toward different visual regions within the same image, as shown in rows one and two of Figure 2. Although CLIP [24] demonstrates strong global representation capability, it lacks effective local region representation and may direct attention to irrelevant areas. Region-level alignment approaches, such as DeCLIP [30], partially mitigate this limitation by reinforcing region-level vision–language alignment. However, they still fail to ensure consistent attention across synonymous expressions, ultimately leading to inconsistent localization.

To address this issue, we propose Synonym-Coherent Language-Image Pretraining (SynCLIP), a novel framework designed to achieve synonym-robust grounding for OVDP tasks. SynCLIP consists of two key components: a Semantic-consistent Spatial Attention alignment (SSA) module and a Spatial Attention Refinement (SAR) module, which jointly enforce consistent and precise spatial attention across synonymous expressions. Specifically, SSA enforces spatial attention consistency by minimizing the discrepancies between attention maps generated from original and semantically enriched synonymous expressions through a semantic alignment loss. This alignment effectively mitigates attention shifts caused by linguistic variation, thereby improving the robustness of dense perception. To further refine spatial grounding, SAR adopts a semantic token selection mechanism to identify the most semantically relevant spatial regions from enriched attention maps. It then leverages the spatial contextual reasoning capability of VFMs [23] to produce refined context-aware attention maps, which are subsequently aggregated to enhance attention accuracy and localization precision. In addition, we construct a Synonym-Enriched Visual Corpus (SEViC) that augments each category with multiple synonyms and textual definitions, serving as the foundation for synonym-coherent pretraining in SynCLIP.

The effectiveness of SynCLIP is validated on multiple OVDP benchmarks. Experimental results demonstrate that SynCLIP significantly improves grounding consistency across diverse synonymous expressions and achieves state-of-the-art performance among CLIP-based methods. The main contributions of this paper are summarized as follows:

- We are the first to identify and analyze **synonym-induced grounding inconsistency** in CLIP-based methods, and reveal its negative impact on OVDP tasks.
- We propose SynCLIP, a novel framework that enhances **spatial attention consistency** across **semantically equivalent expressions**, thereby enhancing grounding robustness in real-world OVDP scenarios.
- We construct SEViC, a synonym-enriched image–text corpus that enhances category supervision with synonyms and definitions, providing **a foundation for synonym-coherent language–image pretraining**.
- Extensive experiments on multiple benchmarks verify the effectiveness of SynCLIP, achieving state-of-the-art performance and improved grounding consistency.

2. Related Work

2.1. Vision-Language Pretraining

Vision-language pretraining has emerged as a powerful paradigm for advancing visual understanding, particularly in open-vocabulary scenarios [46]. Among existing methods, contrastive vision-language models, such as CLIP [24, 26], ALIGN [11] and Florence [40], have shown strong generalization and visual-language alignment by learning

from web-scale image-text pairs. These models align image and text embeddings in a shared feature space through contrastive learning and have achieved remarkable success in tasks such as image-text retrieval and classification [5, 15]. However, most of them perform alignment at the image level, relying on global representations without explicitly modeling the correspondence between local regions and textual expressions [43]. This design limits their effectiveness in OVDP tasks, where precise spatial grounding and text-aligned region-level representation are essential.

To bridge this gap, several recent efforts have explored region-level vision-language alignment by adapting or fine-tuning CLIP to better model local interactions. For instance, CLIPSelf [33] aligns region-level features with their corresponding cropped image-level representations, without requiring explicit region-text annotations. CLIM [34] treats randomly sampled patches as pseudo regions and aligns them with corresponding text in a mosaicked image, enabling spatially localized vision-language learning without box annotations. In contrast, DeCLIP [30] transfers spatial correlation patterns from VFMs for contextual consistency. These approaches have collectively improved the region-level alignment capabilities of CLIP-based models, thereby advancing their performance in OVDP tasks.

Despite these efforts, existing methods still fail to ensure consistent spatial attention across synonymous expressions. This limitation motivates our work, which explicitly addresses the need for synonym-robust grounding ability in real-world applications.

2.2. Open-Vocabulary Dense Perception

OVDP has attracted increasing attention for its ability to localize arbitrary visual concepts defined by textual categories. It primarily includes two sub-tasks, *i.e.*, OVD and OVS. In OVD, methods like ViLD [6], VLDet [19] and GLIP [16] utilize CLIP-derived text embeddings to represent categories and align them with region proposals. These approaches typically treat text as classifiers or attention queries. Similarly, in OVS, methods like ZSSeg [36] and OVSeg [18] usually follow a two-stage pipeline that first generates class-agnostic segmentation masks and then classifies each mask using CLIP-derived text embeddings. More recent approaches, such as SAN [37] and CAT-Seg [2], directly leverage CLIP embeddings to predict segmentation masks in an end-to-end manner. While these approaches successfully extend CLIP to OVDP tasks, they overlook the critical issue of synonym-induced grounding inconsistency, which significantly reduces the robustness of existing methods in real-world applications.

3. Proposed Method

In this section, we introduce SynCLIP, a novel framework designed to enhance synonym-robust grounding ability for OVDP tasks. As illustrated in Figure 3, SynCLIP consists of two main components, *i.e.*, semantic-consistent spatial attention alignment (SSA) and spatial attention refinement (SAR), which explicitly enforce consistent and precise spatial attention across diverse synonymous expressions.

3.1. Preliminaries

CLIP has become a foundational paradigm in vision-language research, which consists of a visual encoder $\mathcal{F}_v(\cdot)$ and a text encoder $\mathcal{F}_t(\cdot)$. The visual encoder is typically implemented using either convolutional neural networks (CNNs) [9] or vision transformers (ViTs) [3], while the text encoder usually adopts a standard transformer architecture to process textual inputs. In this study, we adopt the ViT-based variant of CLIP. Given an image-text pair, CLIP encodes the image $I \in \mathbb{R}^{3\times h\times w}$ and text T into visual embedding $\mathbf{z} = \mathcal{F}_v(I)$ and text embedding $\mathbf{t} = \mathcal{F}_t(T)$, respectively. Here, $\mathbf{z}$ is the dense feature of the image, which is obtained by slightly modifying the last attention block of $\mathcal{F}_v(\cdot)$ following prior work [44].

Specifically, let $\mathbf{x} = \{\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_m\}$ denote the input tokens to the last attention block of visual encoder $\mathcal{F}_v(\cdot)$, where $\mathbf{x} \in \mathbb{R}^{(m+1)\times d}$ and $\mathbf{x}_0$ is the [CLS] token that captures the global representation of the image. The remaining tokens, each representing a $p \times p$ image patch, correspond to spatial locations, where $m = wh/p^2$ is the total number of image patches. Thus, $\mathbf{z}$ is calculated as follows:

$$\mathbf{q} = \text{Proj}_q(\mathbf{x}), \mathbf{k} = \text{Proj}_k(\mathbf{x}), \mathbf{v} = \text{Proj}_v(\mathbf{x}), \tag{1}$$

$$\mathbf{y} = \mathbf{x} + \text{Proj}(\mathbf{Attn}_{csa}\mathbf{v}), \tag{2}$$

$$\mathbf{z} = \mathbf{y} + \text{FFN}(\mathbf{y}), \tag{3}$$

where $\mathbf{q}$, $\mathbf{k}$ and $\mathbf{v}$ represent the query, key and value embedding, respectively. $\text{Proj}(\cdot)$ and $\text{FFN}(\cdot)$ denote the projection layers and feed-forward network, respectively. $\mathbf{Attn}_{csa}$ represents the correlative self-attention process [28], which is formulated as follows:

$$\begin{aligned}\mathbf{Attn}_{csa} = &\ \text{SoftMax}(\mathbf{q}\mathbf{q}^\top/\tau) \\ &+ \text{SoftMax}(\mathbf{k}\mathbf{k}^\top/\tau),\end{aligned} \tag{4}$$

where the temperature coefficient τ is set to $\sqrt{d}$ by default. After obtaining the visual embedding $\mathbf{z}$, we discard the [CLS] token to attain dense feature representation $\mathbf{X}_{\text{dense}} \in \mathbb{R}^{m\times d}$. For ease of description, the following sections use $\mathbf{X}_{\text{dense}} = \mathcal{F}_v(I)$ to represent the above process.

3.2. Semantic-consistent Spatial Attention Alignment

To mitigate synonym-induced grounding inconsistency, we propose SSA, which promotes consistent spatial grounding by aligning the attention maps of original labels with those derived from semantically enriched expressions.

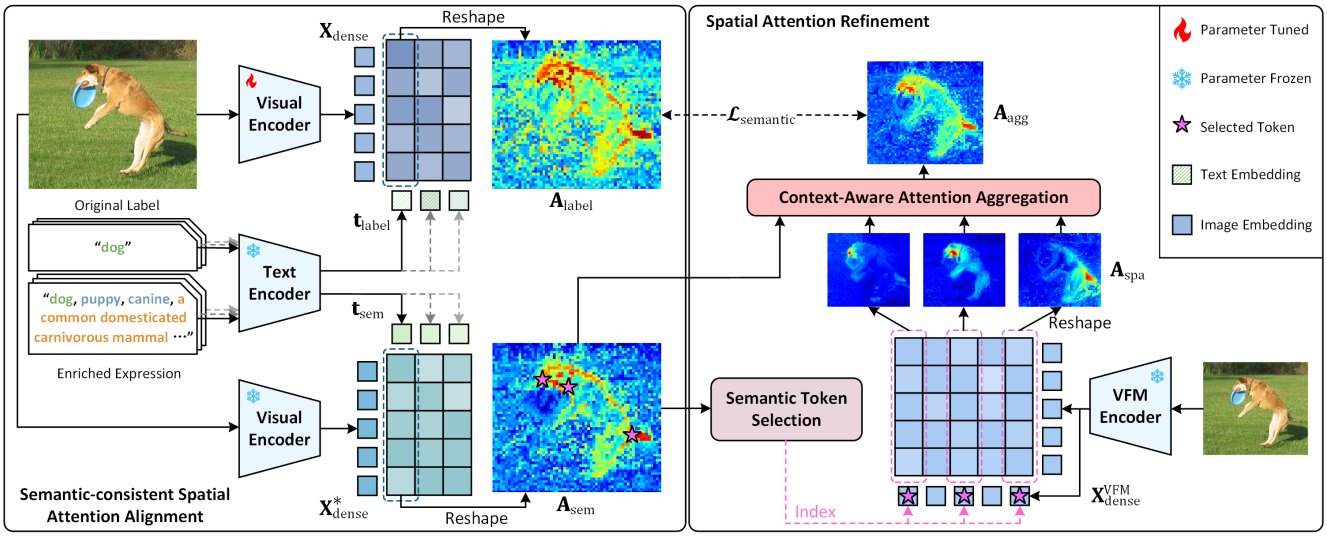


Figure 3. Illustration of the SynCLIP framework, which consists of two key components, *i.e.*, semantic-consistent spatial attention alignment (SSA) and spatial attention refinement (SAR) modules. SSA enforces consistent grounding by aligning spatial attention between original labels and semantically enriched expressions, thereby reducing synonym-induced inconsistency. SAR refines attention precision by selecting the most semantically relevant tokens and aggregating context-aware spatial correlations derived from vision foundation models (VFMs). This synergistic design enables SynCLIP to achieve robust and precise grounding across diverse synonymous expressions.

As SSA relies on sufficient textual variations to support this alignment, we first construct a synonym-enriched visual corpus (SEViC), which augments category-level supervision with diverse synonymous expressions, including synonyms and definitions. SEViC serves as a foundational resource for learning robust language–image associations under expression diversity. Detailed construction of SEViC is described in Section 4.1.1.

Based on SEViC, SSA leverages a set of semantically enriched expressions as reference guidance. Specifically, given an input image I and a set of original labels $\mathbb{T}_{\text{label}} = \{T_i^{\text{label}}\}_{i=1}^{n}$ (*e.g.*, *"dog"*), SSA first constructs the corresponding semantically enriched synonymous expressions $\mathbb{T}_{\text{sem}} = \{T_i^{\text{sem}}\}_{i=1}^{n}$ by incorporating synonyms and definitions (*e.g.*, *"dog, puppy, canine, a common domesticated carnivorous mammal ..."*). These enriched expressions provide more comprehensive semantic grounding cues, whose attention maps tend to localize object regions more accurately and consistently. Therefore, they serve as reliable references for aligning the spatial attention of original labels. After that, both $\mathbb{T}_{\text{label}}$ and $\mathbb{T}_{\text{sem}}$ are encoded by the frozen text encoder $\mathcal{F}_t(\cdot)$ of CLIP, yielding sentence-level text embeddings $\mathbf{t}_{\text{label}} \in \mathbb{R}^{n\times d}$ and $\mathbf{t}_{\text{sem}} \in \mathbb{R}^{n\times d}$, where n denotes the number of expressions and d is the embedding dimension.

To model semantically consistent grounding, SSA employs a teacher-student dual visual encoder architecture to extract dense visual representations. Given an input image I, the student model $\mathcal{F}_v(\cdot)$ and its frozen counterpart $\mathcal{F}_v^*(\cdot)$, serving as the teacher model, generate dense feature representations $\mathbf{X}_{\text{dense}} = \mathcal{F}_v(I) \in \mathbb{R}^{m\times d}$ and $\mathbf{X}^*_{\text{dense}} = \mathcal{F}_v^*(I) \in \mathbb{R}^{m\times d}$, respectively. SSA then computes attention maps between visual features and both original and semantically enriched text embeddings, establishing a foundation for spatial attention alignment.

$$\mathbf{A}_{\text{label}} = \frac{\mathbf{t}_{\text{label}}\mathbf{X}_{\text{dense}}^{\top}}{||\mathbf{t}_{\text{label}}|| \cdot ||\mathbf{X}_{\text{dense}}||}, \tag{5}$$

$$\mathbf{A}_{\text{sem}} = \frac{\mathbf{t}_{\text{sem}}\mathbf{X}^*_{\text{dense}}{}^{\top}}{||\mathbf{t}_{\text{sem}}|| \cdot ||\mathbf{X}^*_{\text{dense}}||}, \tag{6}$$

where $\mathbf{A} \in \mathbb{R}^{n\times m}$ denotes the spatial attention maps between the text and image embeddings. To enforce semantic consistency across diverse expressions referring to the same concept, SSA introduces the semantic alignment loss:

$$\mathcal{L}_{\text{semantic}} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}||\mathbf{A}_{\text{label}}^{i,j} - \mathbf{A}_{\text{sem}}^{i,j}||_2. \tag{7}$$

By minimizing the discrepancy between the attention distributions derived from original and enriched expressions, SSA encourages the model to produce consistent spatial attention across synonymous or semantically equivalent expressions, thereby enhancing the stability and coherence of spatial grounding regardless of expression diversity.

3.3. Spatial Attention Refinement

To further improve spatial grounding precision, we propose SAR, which consists of two stages, *i.e.*, semantic token selection to identify the most semantically relevant spatial locations, and context-aware attention aggregation to refine

spatial attention through contextual reasoning. SAR leverages the spatial representation capability of vision foundation models (VFMs) [23] to enhance both spatial accuracy and semantic coherence, leading to more reliable localization across synonymous expressions.

3.3.1. Semantic Token Selection

Given the attention maps of semantically enriched expressions $\mathbf{A}_{\text{sem}}$, which capture the correlation between dense feature representation $\mathbf{X}^{*}_{\text{dense}}$ and the enriched text embedding $\mathbf{t}_{\text{sem}}$, SAR first identifies the most semantically relevant spatial locations via a semantic token selection mechanism. Specifically, the top-k tokens with the highest attention scores are selected for each enriched expression:

$$\mathcal{K} = \text{TopK}(\mathbf{A}_{\text{sem}}, k), \tag{8}$$

where $\mathcal{K} \in \mathbb{R}^{n\times k}$ stores the indices of selected tokens across n text embeddings. A pretrained VFM model, denoted as $\mathcal{F}_v^{\text{VFM}}(\cdot)$, extracts the dense visual feature $\mathbf{X}^{\text{VFM}}_{\text{dense}} \in \mathbb{R}^{m\times d}$ that encodes rich spatial context. Here, $\mathbf{X}^{\text{VFM}}_{\text{dense}} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m\}$ represents the image tokens of VFM feature, which has the same shape of $\mathbf{X}^{*}_{\text{dense}}$.

Based on the indices in $\mathcal{K}$, SAR computes the spatial correlation matrix $\mathbf{S}$ by evaluating cosine similarity between each selected token and all image tokens:

$$\mathbf{s}_{i,j} = \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{\max(||\mathbf{x}_i|| \cdot ||\mathbf{x}_j||, \epsilon)}, \tag{9}$$

where ϵ is set to 10^{-8} by default to avoid division by zero at extrema. This results in a set of k spatial correlation matrices per text embedding, *i.e.*, $\mathbf{S} = \{\mathbf{s}_{i,j} | i \in \mathcal{K}, j \in [1, m]\} \in \mathbb{R}^{n\times k\times m}$, each of which represents the spatial correlation from a key semantic location to the entire image.

3.3.2. Context-Aware Attention Aggregation

To obtain a refined and semantically coherent attention map, the spatial correlation matrices $\mathbf{S}$ are aggregated into a final attention map $\mathbf{A}_{\text{spa}} = \text{Aggregator}(\mathbf{S}) \in \mathbb{R}^{n\times m}$, where $\text{Aggregator}(\cdot)$ computes the mean across the k selected tokens. This simple averaging strategy proves empirically effective, as it highlights spatial regions that are consistently correlated with the most informative semantic cues. Figure 4 visualizes this process, showing how semantic attention, spatial correlation, and aggregated attention jointly contribute to SAR.

Finally, to balance semantic relevance and spatial precision, SAR fuses the semantic attention $\mathbf{A}_{\text{sem}}$ and the spatial correlation attention $\mathbf{A}_{\text{spa}}$ as:

$$\mathbf{A}_{\text{agg}} = \alpha\mathbf{A}_{\text{spa}} + \beta\mathbf{A}_{\text{sem}}, \tag{10}$$

where α and β control the trade-off between spatial accuracy and semantic consistency, and the effect of their values is further analyzed in the Appendix. The refined attention $\mathbf{A}_{\text{agg}}$ supervises grounding accuracy and consistency

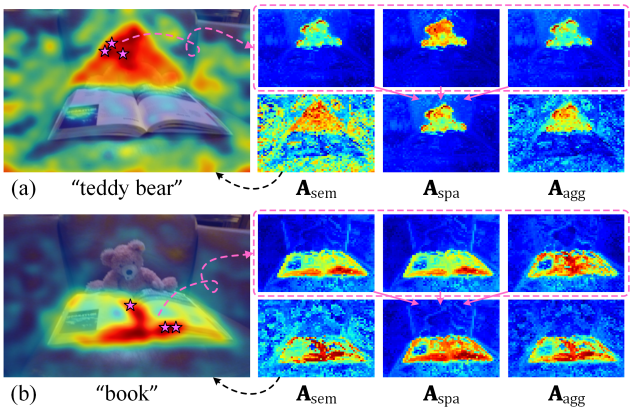


Figure 4. Illustration of the SAR process with k=3. From left to right, we visualize the semantic attention $\mathbf{A}_{\text{sem}}$, the spatial correlation attention $\mathbf{A}_{\text{spa}}$, and the aggregated attention $\mathbf{A}_{\text{agg}}$. The aggregated map preserves both semantic relevance and text-guided spatial localization.

by minimizing the optimized semantic alignment loss:

$$\mathcal{L}^{+}_{\text{semantic}} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} \|\mathbf{A}^{i,j}_{\text{label}} - \mathbf{A}^{i,j}_{\text{agg}}\|_2. \tag{11}$$

Through this refinement and aggregation mechanism, SAR enhances spatial fidelity of attention maps while preserving the semantics conveyed by enriched textual expressions, thereby ensuring semantic coherence and accurate spatial grounding across diverse synonymous expressions for robust performance in real-world OVDP scenarios.

4. Experiments

4.1. Datasets and Metrics

4.1.1. Synonym-Enriched Visual Corpus

To enable synonym-coherent language–image pretraining for robust OVDP, we construct a synonym-enriched visual corpus (SEViC). SEViC augments category-level supervision with semantically consistent expressions, including synonyms and textual definitions, encouraging consistent grounding across synonymous expressions.

To ensure generalizability, SEViC is built upon the training images of COCO2017 dataset [20], a widely used benchmark offering diverse visual contexts. Since LVIS [8] extends COCO2017 with finer-grained and long-tailed annotations over the same image set, we adopt category names from both datasets to form a comprehensive vocabulary of 1,232 unique categories, covering both frequent and rare concepts. In addition, two meta-categories, *"object"* and *"background"*, are further added to represent special global semantics. To enrich linguistic supervision, we incorporate the LVIS-provided synonyms and definitions where

#	Class	Definition	Synonyms
1	boat	*a small vessel for traveling on water ...*	*vessel, watercraft, small boat, ...*
2	giraffe	*tall animal having a very long neck and legs ...*	*giraffa camelopardalis, long-necked animal, ...*
3	frisbee	*a light, plastic disk propelled with a flip ...*	*flying disc, throw disc, wham-o disc, ...*

Table 1. Examples of synonym-enriched textual expressions in SEViC. For each category, SEViC includes a concise definition and multiple semantically consistent synonyms, which are used to enrich linguistic supervision during synonym-coherent pretraining.

available, and employ large language models (LLMs), *e.g.*, DeepSeek [21], to automatically generate multiple synonyms and concise definitions, as illustrated in Table 1. All generated expressions are validated via LLM-based consistency checks to ensure semantic fidelity. Each image is finally paired with its associated categories and verified expressions, forming a synonym-enhanced image–text corpus comprising 118,287 images, 1,234 categories, and 11,558 semantically enriched expressions.

SEViC serves as the foundation for pretraining SynCLIP, enabling the model to learn synonym-robust dense perception, thereby mitigating the synonym-induced grounding inconsistency observed in CLIP-based models.

4.1.2. Downstream Finetuning Datasets

We conduct evaluations on two widely used benchmarks, *i.e.*, OV-COCO [20] and OV-LVIS [8]. For OV-COCO, following the setup of previous work [34], we divide the 80 classes in the COCO dataset [20] into 48 base classes and 17 novel classes. The training set comprises 107,761 images containing annotations for base classes only, while evaluation is performed on a test set of 4,836 images with annotations for both base and novel classes. Detection performance is measured using the box Average Precision (AP) at IoU threshold 0.5, denoted as AP_{50}, with a specific focus on $\text{AP}_{50}^{\text{novel}}$, which reflects the model's ability to generalize to novel classes.

Regarding OV-LVIS, we retain 866 common and frequent classes in the LVIS dataset [8] as base classes and 337 rare classes as novel classes, following the setup in ViLD [6]. Similarly, the training annotations are limited to base classes. We report mean Average Precision (mAP) of masks averaged over IoU thresholds from 0.5 to 0.95. The performance on rare (*i.e.*, novel) classes, denoted as $\text{mAP}_r^{\text{mask}}$, is the primary evaluation metric in experiments.

4.2. Implementation Details

SynCLIP is implemented based on the DeCLIP framework [30], upon which we integrate the proposed SSA and SAR modules. The default weight of semantic alignment loss is set to 0.05. We use DINOv2 [23] as the VFM encoder. The balance coefficients for attention aggregation, *i.e.*, α and β, are both set to 0.5 by default. Training is conducted using 4 GPUs with a batch size of 8 per GPU. We train our SynCLIP for 6 epochs using the AdamW [22] optimizer with a learning rate of 2×10^{-5} and a weight decay of 0.1.

For downstream tasks, we adopt F-ViT [33] as the baseline, forming our method, *i.e.*, F-ViT+SynCLIP. All other hyper-parameters follow the default settings of the respective baselines. Following a previous study [30, 33], we ensure fair comparisons using the same training data as the corresponding baselines.

4.3. Main Results

We evaluate SynCLIP on two representative benchmarks [8, 20] to verify both its general performance and robustness to linguistic variations. Experiments are conducted in two stages. We first conduct benchmark comparisons with state-of-the-art methods to demonstrate the effectiveness of synonym-coherent pretraining. We then perform a synonym-robustness evaluation under synonym substitution, further validating SynCLIP's ability to maintain consistent grounding across diverse textual expressions.

4.3.1. Benchmark Comparison

To assess whether introducing synonym-coherent alignment benefits general OVDP performance, we evaluate SynCLIP under standard protocols. The core insight is that integrating synonym-consistent supervision during pretraining helps the model learn stable and semantically invariant vision–language attention. By reducing inconsistencies across synonymous expressions, SynCLIP acquires stronger category semantics, leading to improved novel-class recognition and generalization.

Results on OV-COCO: As shown in Table 2a, SynCLIP significantly enhances the performance over F-ViT on the OV-COCO benchmark, achieving 43.6 and 49.8 $\text{AP}_{50}^{\text{novel}}$ with ViT-B/16 and ViT-L/14 backbones, respectively. Beyond raw accuracy, the improvement highlights a more semantically stable vision–language representation, where different lexical forms referring to the same concept yield consistent region–text alignment. This stability directly translates into stronger recognition of unseen classes, confirming that SynCLIP's synonym-coherent pretraining narrows the linguistic gap between base and novel categories, enabling SynCLIP to outperform previous methods on the OV-COCO.

Results on OV-LVIS: Table 2b presents results on the more challenging OV-LVIS benchmark. FViT+SynCLIP achieves 27.8 and 37.2 $\text{mAP}_r^{\text{mask}}$ with ViT-B/16 and ViT-

Method	Venue	Backbone	AP_{50}^{novel}
ViLD [6]	ICLR2022	RN50	27.6
F-VLM [14]	ICLR2023	RN50	28.0
OV-DETR [41]	ECCV2022	RN50	29.4
BARON-KD [32]	CVPR2023	RN50	34.0
RegionCLIP [43]	CVPR2022	RN50x4	39.3
CORA$^+$ [35]	CVPR2023	RN50x4	43.1
OV-DQUO [29]	AAAI2025	RN50x4	45.6
RO-ViT [13]	CVPR2023	ViT-L/16	33.0
CFM-ViT [12]	ICCV2023	ViT-L/16	34.1
F-ViT+CLIPSelf [33]	ICLR2024	ViT-B/16	37.6
F-ViT+CLIPSelf [33]	ICLR2024	ViT-L/14	44.3
F-ViT+DeCLIP [30]	CVPR2025	ViT-B/16	41.1
F-ViT+DeCLIP [30]	CVPR2025	ViT-L/14	46.2
F-ViT+SynCLIP (Ours)	CVPR2026	ViT-B/16	**43.6**
F-ViT+SynCLIP (Ours)	CVPR2026	ViT-L/14	**49.8**

(a) Results on OV-COCO benchmark

Method	Venue	Backbone	mAP_r^{mask}
ViLD [6]	ICLR2022	RN50	16.3
OV-DETR [41]	ECCV2022	RN50	17.4
RegionCLIP [43]	CVPR2022	RN50x4	22.0
BARON-KD [32]	CVPR2023	RN50	22.6
OV-SAM [39]	CVPR2024	RN50x16	24.0
CORA$^+$ [35]	CVPR2023	RN50x4	28.1
F-VLM [14]	ICLR2023	RN50x64	32.8
PromptOVD [25]	Arxiv2023	ViT-B/16	23.1
RO-ViT [13]	CVPR2023	ViT-L/16	32.4
F-ViT+CLIPSelf [33]	ICLR2024	ViT-B/16	25.3
F-ViT+CLIPSelf [33]	ICLR2024	ViT-L/14	34.9
F-ViT+DeCLIP [30]	CVPR2025	ViT-B/16	26.8
F-ViT+DeCLIP [30]	CVPR2025	ViT-L/14	**37.2**
F-ViT+SynCLIP (Ours)	CVPR2026	ViT-B/16	**27.8**
F-ViT+SynCLIP (Ours)	CVPR2026	ViT-L/14	**37.2**

(b) Results on OV-LVIS benchmark

Table 2. Comparison with state-of-the-art methods. 'RN50' represents ResNet50 [9], where 'RN50x4' is scaled up by a factor of 4 according to the scaling rules [27]. 'B' and 'L' in ViT-based methods stand for base and large model sizes. '/16' and '/14' indicate the downsample ratio of input images. The best results among methods using comparable backbones are highlighted in bold.

L/14 backbones, respectively. The former surpasses all compared methods with similar backbones, while the latter is comparable to the best existing result under the large-backbone setting. It is worth noting that OV-LVIS poses a fundamentally more challenging scenario than OV-COCO, as it involves fine-grained categories with long-tailed distributions and extensive lexical overlap. Under such conditions, models tend to overfit to frequent names and struggle to generalize across semantically similar yet lexically diverse concepts. SynCLIP maintains competitive performance in this setting, indicating its ability to transfer category semantics beyond surface lexical cues.

Across both OV-COCO and OV-LVIS, the consistent improvements can be attributed to SynCLIP's synonym-coherent pretraining, which provides more stable and semantically aligned region–text attention. SSA reduces attention variation caused by lexical differences, and SAR further refines spatial grounding with context-aware cues. These two components jointly strengthen the underlying vision–language alignment, allowing SynCLIP to capture clearer category semantics and generalize more effectively to novel categories. This explains why SynCLIP is able to outperform existing SOTA methods under standard open-vocabulary settings.

4.3.2. Synonym-Robust Dense Perception

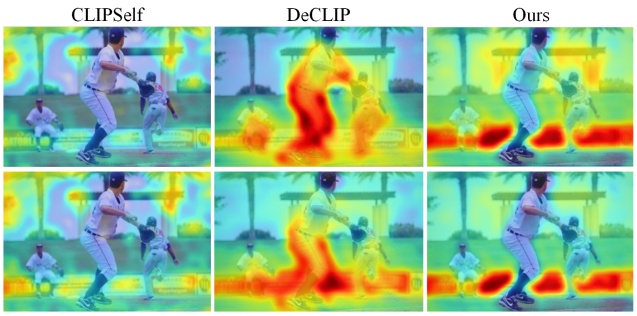
While the benchmark results demonstrate overall improvement, a key motivation of our SynCLIP is to enhance robustness when the original category names are replaced with synonymous expressions. To assess this, we evaluate grounding performance under a synonym substitution setting, following the same protocol used for comparing against state-of-the-art OVDP methods.

Method	syn.	AP_{50}^{novel}	AP_{50}^{base}
F-ViT+CLIPSelf [33]	✘	37.6	54.9
F-ViT+CLIPSelf [33]	✔	28.8 (-8.8)	48.3 (-6.6)
F-ViT+DeCLIP† [30]	✘	41.0	50.9
F-ViT+DeCLIP [30]	✔	31.5 (-9.5)	48.7 (-2.2)
F-ViT+SynCLIP (Ours)	✘	43.6	51.8
F-ViT+SynCLIP (Ours)	✔	**39.2 (-4.4)**	**51.2 (-0.6)**

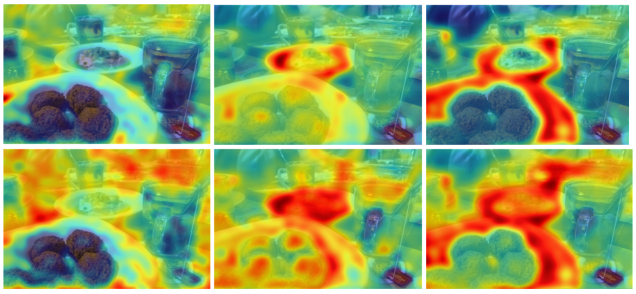
Table 3. Robustness evaluation under synonym substitution on OV-COCO. AP_{50}^{novel} measures grounding performance on novel classes, and AP_{50}^{base} on base classes. 'syn.' indicates evaluation where category names are replaced by synonymous expressions, revealing each model's sensitivity to linguistic variation. † denotes reproduced results.

Table 3 shows that existing approaches exhibit notable drops when synonyms are used: CLIPSelf decreases by 8.8 and DeCLIP [30] by 9.5 AP_{50}^{novel}. SynCLIP also experiences a reduction, but the magnitude is smaller with only 4.4, while base-class performance remains largely unchanged. This suggests that our synonym-coherent pretraining helps reduce the attention shifts introduced by linguistic variations and leads to more stable grounding across semantically equivalent expressions. Figure 5 provides complementary qualitative evidence, showing that SynCLIP maintains more coherent attention regions under different synonymous expressions.

Together, these results align with the grounding inconsis-



(a) Top: *"banner"* vs. Bottom: *"advertising banner"*



(b) Top: *"bowl"* vs. Bottom: *"deep dish"*

Figure 5. Qualitative comparison of attention maps under synonymous expressions. (a) *"banner"* vs. *"advertising banner"* and (b) *"bowl"* vs. *"deep dish"*. Compared with CLIPSelf and DeCLIP, SynCLIP maintains more stable and coherent localization across different but semantically equivalent expressions, providing qualitative evidence for its improved robustness to linguistic variations and the effectiveness of synonym-coherent pretraining.

SSA	SAR	AP_{50}^{novel}	AP_{50}^{base}	AP_{50}^{all}
✘	✘	41.0	50.9	48.3
✔	✘	41.1	51.8	48.2
✘	✔	42.2	52.2	49.6
✔	✔	**43.6**	51.8	49.6

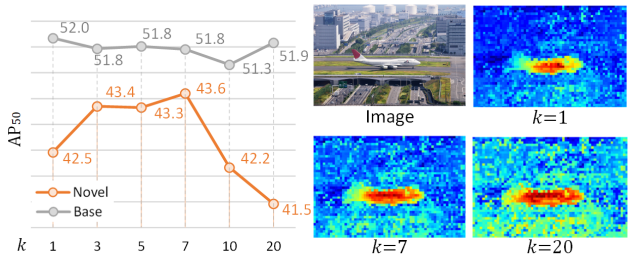
Table 4. Ablation study of components in SynCLIP with F-ViT on the OV-COCO benchmark. Both SSA and SAR contribute to performance gains, and their combination yields the best novel-category results, demonstrating their complementarity.

tency observed in our motivation and indicate that SynCLIP offers improved robustness to textual variability compared with prior OVDP methods.

4.4. Ablation Study

4.4.1. Component Analysis

Table 4 shows the effectiveness of the main components in SynCLIP with F-ViT through quantitative results on the OV-COCO. Starting from the baseline, we observe an AP_{50}^{novel} of 41.0 in the first row. Introducing SSA alone yields a marginal improvement to 41.1 AP_{50}^{novel}, confirming that its standalone effect is limited due to the redundant information introduced by semantically enriched expression, which hinders precise spatial attention. In contrast, applying SAR alone significantly boosts AP_{50}^{novel} to 42.2, highlighting the importance of refining spatial attention with semantic tokens. When combining both components, SynCLIP achieves the best performance with 43.6 AP_{50}^{novel}, demonstrating the complementary benefits of SSA and SAR in enhancing overall robustness and performance.



(a) Performance trends (b) Spatial attention visualization

Figure 6. Effect of the number of semantic tokens, *i.e.*, k. (a) shows performance trends on the OV-COCO benchmark with k increasing. (b) shows the qualitative comparison of spatial correlation attention, *i.e.*, $\mathbf{A}_{\text{spa}}$, with various k values.

4.4.2. Effect of Semantic Token Selection

To analyze the effectiveness of the semantic token selection mechanism in SAR, we investigate how varying the number of semantic tokens, *i.e.*, k, influences performance on the OV-COCO benchmark. As shown in Figure 6 (a), increasing k from 1 to 7 progressively enhances AP_{50}^{novel} from 42.5 to a peak of 43.6. However, further increasing k to 10 and 20 causes the performance to drop to 42.2 and 41.5, respectively. This demonstrates that selecting a moderate number of semantic tokens (*e.g.*, $k = 7$) achieves an optimal trade-off between informativeness and noise suppression, as shown in Figure 6 (b). Excessive token introduces irrelevant cues, which degrade spatial grounding accuracy.

5. Conclusion

This paper identifies synonym-induced grounding inconsistency as a critical limitation in CLIP-based methods, where synonymous expressions cause divergent spatial attention and hinder robust dense perception. To address this, we propose SynCLIP, which enforces semantic coherence by aligning attention maps across synonymous expressions and enhances spatial precision through semantic token selection and context-aware attention refinement. In addition, we construct SEViC that enriches category-level supervision with synonymous expressions, providing a foundation for pretraining SynCLIP. Extensive evaluations on multiple benchmarks demonstrate that SynCLIP significantly improves grounding consistency and achieves state-of-the-art performance, paving the way for more reliable and robust real-world OVDP applications.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[2] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. CAT-Seg: Cost aggregation for open-vocabulary semantic segmentation. In *CVPR*, pages 4113–4123, 2024. 3

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3

[4] Shenghao Fu, Junkai Yan, Qize Yang, Xihan Wei, Xiaohua Xie, and Wei-Shi Zheng. A hierarchical semantic distillation framework for open-vocabulary object detection. *IEEE TMM*, 27:8835–8846, 2025. 1

[5] Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, and Jianfeng Gao. Vision-language pre-training: Basics, recent advances, and future trends. *Found. Trends Comput. Graph. Vis.*, 14:163–352, 2022. 3

[6] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022. 3, 6, 7

[7] Jian Guan, Mingjie Xie, Youtian Lin, Guangjun He, and Pengming Feng. EARL: An elliptical distribution aided adaptive rotation label assignment for oriented object detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.*, 61:1–15, 2023. 1

[8] Agrim Gupta, Piotr Dollar, Ross Girshick, et al. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*, pages 5356–5364, 2019. 5, 6, 1

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3, 7

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *CVPR*, pages 2961–2969, 2017. 1

[11] 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 *ICML*, pages 4904–4916. PMLR, 2021. 2

[12] Dahun Kim, Anelia Angelova, Weicheng Kuo, et al. Contrastive feature masking open-vocabulary vision transformer. In *ICCV*, pages 15602–15612, 2023. 7

[13] Dahun Kim, Anelia Angelova, Weicheng Kuo, et al. Region-aware pretraining for open-vocabulary object detection with vision transformers. In *CVPR*, pages 11144–11154, 2023. 7

[14] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-VLM: Open-vocabulary object detection upon frozen vision and language models. In *ICLR*, 2023. 7

[15] Haiyan Lan, Shujun Li, Mingjie Xie, Xuanjia Zhao, Hongning Liu, Pengming Feng, Dongli Xu, Guangjun He, and Jian Guan. Band prompting aided SAR and multi-spectral data fusion framework for local climate zone classification. In *ICASSP*, pages 1–5, 2025. 3

[16] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *CVPR*, pages 10965–10975, 2022. 3

[17] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *CVPR*, pages 23390–23400, 2023. 1

[18] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted CLIP. In *CVPR*, pages 7061–7070, 2023. 3

[19] Chuang Lin, Peize Sun, Yi Jiang, Ping Luo, Lizhen Qu, Gholamreza Haffari, Zehuan Yuan, and Jianfei Cai. Learning object-language alignments for open-vocabulary object detection. In *ICLR*, 2023. 3

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 2, 5, 6, 1

[21] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. DeepSeek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. 6, 1

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 6

[23] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024. 2, 5, 6

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 1, 2

[25] Hwanjun Song and Jihwan Bang. Prompt-guided transformers for end-to-end open-vocabulary object detection. *arXiv preprint arXiv:2303.14386*, 2023. 7

[26] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. EVA-CLIP: Improved training techniques for CLIP at scale. *arXiv preprint arXiv:2303.15389*, 2023. 1, 2

[27] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *ICML*, pages 6105–6114. PMLR, 2019. 7

[28] Feng Wang, Jieru Mei, and Alan Yuille. SCLIP: Rethinking self-attention for dense vision-language inference. In *ECCV*, pages 315–332. Springer, 2024. 3

[29] Junjie Wang, Bin Chen, Bin Kang, Yulin Li, Weizhi Xian, Yichi Chen, and Yong Xu. OV-DQUO: Open-vocabulary

detr with denoising text query training and open-world unknown objects supervision. In *AAAI*, pages 7762–7770, 2025. 7

[30] Junjie Wang, Bin Chen, Yulin Li, Bin Kang, Yichi Chen, and Zhuotao Tian. DeCLIP: Decoupled learning for open-vocabulary dense perception. In *CVPR*, pages 14824–14834, 2025. 1, 2, 3, 6, 7

[31] Jianzong Wu, Xiangtai Li, Shilin Xu, Haobo Yuan, Henghui Ding, Yibo Yang, Xia Li, Jiangning Zhang, Yunhai Tong, Xudong Jiang, et al. Towards open vocabulary learning: A survey. *IEEE TPAMI*, 46(7):5092–5113, 2024. 1

[32] Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection. In *CVPR*, pages 15254–15264, 2023. 7

[33] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. CLIPSelf: Vision transformer distills itself for open-vocabulary dense prediction. In *ICLR*, 2024. 1, 2, 3, 6, 7

[34] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Wentao Liu, and Chen Change Loy. CLIM: Contrastive language-image mosaic for region representation. In *AAAI*, pages 6117–6125, 2024. 1, 3, 6

[35] Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. CORA: Adapting CLIP for open-vocabulary detection with region prompting and anchor pre-matching. In *CVPR*, pages 7031–7040, 2023. 1, 7

[36] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *ECCV*, pages 736–753. Springer, 2022. 3

[37] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *CVPR*, pages 2945–2954, 2023. 3

[38] Yuhang Yang, Jinhong Deng, Wen Li, and Lixin Duan. ResCLIP: Residual attention for training-free dense vision-language inference. In *CVPR*, pages 29968–29978, 2025. 1

[39] Haobo Yuan, Xiangtai Li, Chong Zhou, Yining Li, Kai Chen, and Chen Change Loy. Open-vocabulary SAM: Segment and recognize twenty-thousand classes interactively. In *ECCV*, pages 419–437. Springer, 2024. 7

[40] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 2

[41] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In *ECCV*, pages 106–122. Springer, 2022. 7

[42] Yuhang Zheng, Xiangyu Chen, Yupeng Zheng, Songen Gu, Runyi Yang, Bu Jin, Pengfei Li, Chengliang Zhong, Zengmao Wang, Lina Liu, et al. GaussianGrasper: 3d language gaussian splatting for open-vocabulary robotic grasping. *IEEE Robot. Autom. Lett.*, 9(9):7827–7834, 2024. 1

[43] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. RegionCLIP: Region-based language-image pretraining. In *CVPR*, pages 16793–16803, 2022. 1, 3, 7

[44] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from CLIP. In *ECCV*, pages 696–712. Springer, 2022. 3

[45] Yijie Zhou, Xianhui Cheng, Qiming Zhang, Lei Wang, Wenchao Ding, Xiangyang Xue, Chunbo Luo, and Jian Pu. ALGPT: Multi-agent cooperative framework for open-vocabulary multi-modal auto-annotating in autonomous driving. *IEEE Trans. Intell. Veh.*, pages 1–15, 2024. 1

[46] Chaoyang Zhu and Long Chen. A survey on open-vocabulary detection and segmentation: Past, present, and future. *IEEE TPAMI*, 46(12):8954–8975, 2024. 1, 2

SynCLIP: Synonym-Coherent Language-Image Pretraining for Robust Open-Vocabulary Dense Perception

Supplementary Material

Overview

This supplementary material provides additional details and results that complement the main paper. Section 6 describes the construction pipeline of SEViC. Section 7 presents additional ablation studies of key components in SynCLIP. Section 8 provides extended qualitative analyses, and Section 9 reports the efficiency analysis.

6. Details of SEViC Construction

This section provides a detailed description of the construction pipeline of our synonym-enriched visual corpus (SEViC), complementing the overview in the main paper. As illustrated in Figure 7, the pipeline consists of three major stages, *i.e.*, Data Collection, which gathers the full category vocabulary and initial textual resources; LLM-based Text Generation, which expands each category into semantically consistent expressions; and Consistency Validation, which filters and verifies all generated expressions to ensure semantic fidelity. Together, these stages yield a noise-reduced, semantically aligned corpus that supports synonym-coherent pretraining.

6.1. Data Collection

We begin by collecting all category names from COCO2017 [20] and LVIS [8], which share the same image set but differ in granularity and coverage. This provides a unified vocabulary of 1,232 unique object category names and 118,287 images. When available, we also extract their accompanying LVIS-provided definitions and synonyms. These human-curated textual descriptions serve as the initial reference for later LLM-based enrichment. Two meta-categories, *"object"* and *"background"*, are additionally included to represent global semantics.

6.2. LLM-Based Text Generation

To enrich category-level linguistic supervision, we employ a large language model (LLM), *i.e.*, DeepSeek [21], to generate structured semantic metadata. For each category name, optionally with definitions provided by LVIS, the LLM generates a JSON-formatted semantic entry including: (1) a concise definition that is either preserved verbatim when provided or generated when absent, and (2) a list of synonyms, containing up to ten words or compound phrases, maintaining strict semantic and visual equivalence for dense perception. All generated entries follow a machine-readable JSON format for seamless integration into open-vocabulary pretraining.

Synonym	Definition	AP_{50}^{novel}	AP_{50}^{base}
✘	✘	41.1	51.7
✔	✘	42.1	51.5
✘	✔	42.3	51.7
✔	✔	**43.6**	51.8

Table 5. Ablation of textual variant types used in SSA on OV-COCO. The results show that synonyms and definitions provide complementary semantic enrichment, and using both yields the strongest improvement, confirming that SSA benefits from jointly leveraging lexical diversity and explicit semantic grounding.

6.3. Consistency Validation

To eliminate ambiguous or inconsistent expressions introduced by LLMs, we perform a consistency validation using another LLM, *i.e.*, ChatGPT [1]. Each candidate expression is strictly verified to precisely match the category's canonical definition and to refer to exactly the same visual object instances as the original label, rejecting overly broad or context-dependent terms. This validation process eliminates noise and ensures that all retained expressions are semantically equivalent, thereby providing high-quality supervision for pretraining.

After the above stages, each collected image is linked to its associated categories along with validated definitions and synonyms, resulting in a high-quality synonym-enriched visual corpus that serves as a foundation for pretraining SynCLIP.

7. Additional Ablation Studies

This section provides additional analyses on two key components of SynCLIP, *i.e.*, the types of semantically enriched textual variants used in SSA for semantic–consistent attention alignment, and the aggregation weights in SAR that balance semantic relevance and spatial precision. These studies further validate the design choices introduced in the main paper and clarify how each component contributes to robust synonym-coherent dense perception.

7.1. Effect of Textual Variant Types

As shown in Table 5, removing both synonyms and definitions yields the weakest performance, indicating that SSA requires semantically diverse cues to guide attention alignment. Using synonyms or definitions alone already provides

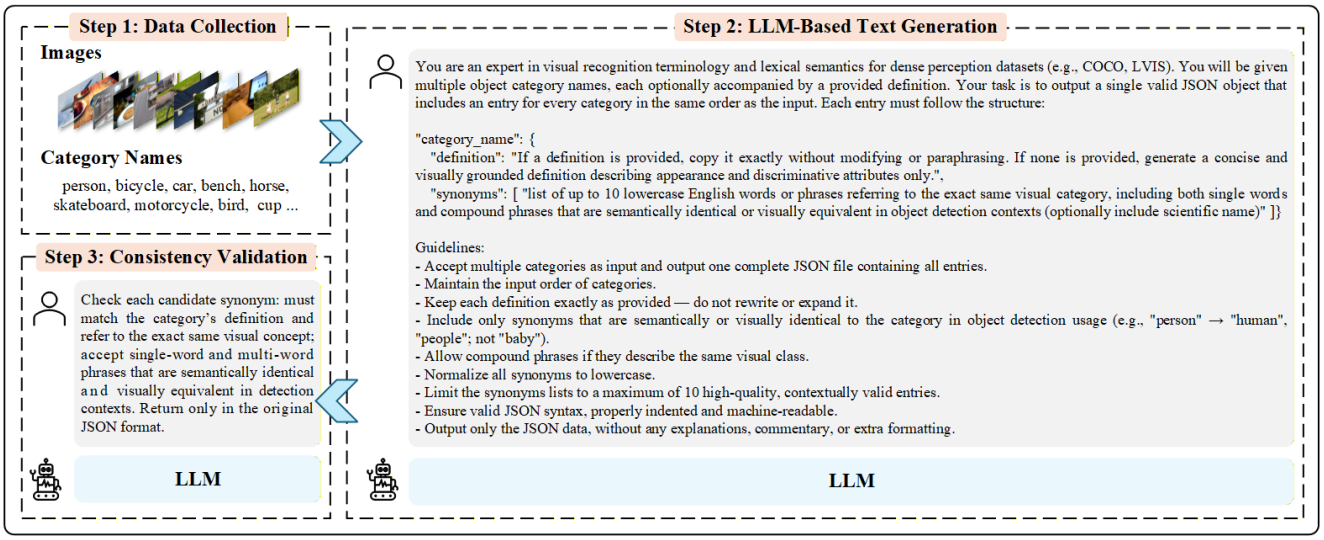


Figure 7. Overview of the SEViC construction pipeline, consisting of (1) Data Collection from COCO and LVIS datasets, which gathers the complete category vocabulary and initial textual resources; (2) LLM-based Text Generation, which expands each category into semantically consistent definitions and synonyms; and (3) Consistency Validation, which filters and verifies all generated expressions through LLM-driven checks to ensure semantic fidelity. These steps jointly produce a semantically coherent, synonym-enriched visual corpus.

clear gains: synonyms introduce lexical variability while retaining the concept, and definitions offer explicit semantic grounding that often captures appearance or functionality. Their combination delivers the highest improvement of 43.6 AP_{50}^{novel}, showing that lexical diversity and semantic specificity contribute complementary benefits. These results confirm that SSA is most effective when aligned to a semantically enriched reference.

7.2. Effect of Attention Aggregation

The SAR module aggregates spatial and semantic attentions via coefficients α and β. Table 6 shows that overly large β leads to more concept-driven activations with degraded localization, while overly large α produces spatially precise but semantically weaker maps that generalize poorly to novel categories. Optimal performance on OV-COCO emerges at a balanced configuration with $\alpha = \beta = 0.5$, demonstrating that robust dense perception requires both spatial structural cues and semantically discriminative signals. This validates the SAR design and highlights the necessity of harmonizing spatial precision with semantic consistency.

8. Additional Qualitative Analysis

This section presents additional qualitative analyses, including visualizations of attention maps from the SAR module and prediction results on two standard dense perception benchmarks, offering a more comprehensive illustration of the effectiveness and superiority of our method.

α	β	AP_{50}^{novel}	AP_{50}^{base}	AP_{50}^{all}
0.9	0.1	41.4	51.0	48.5
0.8	0.2	41.8	51.3	48.8
0.7	0.3	41.4	51.1	48.6
0.6	0.4	42.4	51.4	49.1
0.5	0.5	**43.6**	51.8	49.6
0.4	0.6	43.1	52.1	51.8
0.3	0.7	43.5	52.2	49.9
0.2	0.8	42.5	52.1	49.6
0.1	0.9	41.7	51.7	49.1

Table 6. Ablation of spatial (α) and semantic (β) attention aggregation weights in SAR on OV-COCO. The results demonstrate that balanced fusion produces the best performance, validating the design of SAR as an effective way to harmonize spatial precision with semantic consistency.

8.1. Visualization of Spatial Attention Refinement

Figure 8 provides more qualitative examples of attention maps from the SAR module in SynCLIP, including semantic attention $\mathbf{A}_{sem}$, spatial correlation attention $\mathbf{A}_{spa}$, and aggregated attention $\mathbf{A}_{agg}$. Consistent with the analysis in the main paper, $\mathbf{A}_{sem}$ correctly focuses on concept-relevant areas but may suffer from broad or noisy responses due to lexical variations across enriched expressions. $\mathbf{A}_{spa}$, computed via semantic token selection and visual foundation model (VFM)-driven spatial contextual reasoning [23], yields concentrated activations around the true object extent. Thus, $\mathbf{A}_{agg}$ aggregates $\mathbf{A}_{sem}$ and $\mathbf{A}_{spa}$, preserving both semantic

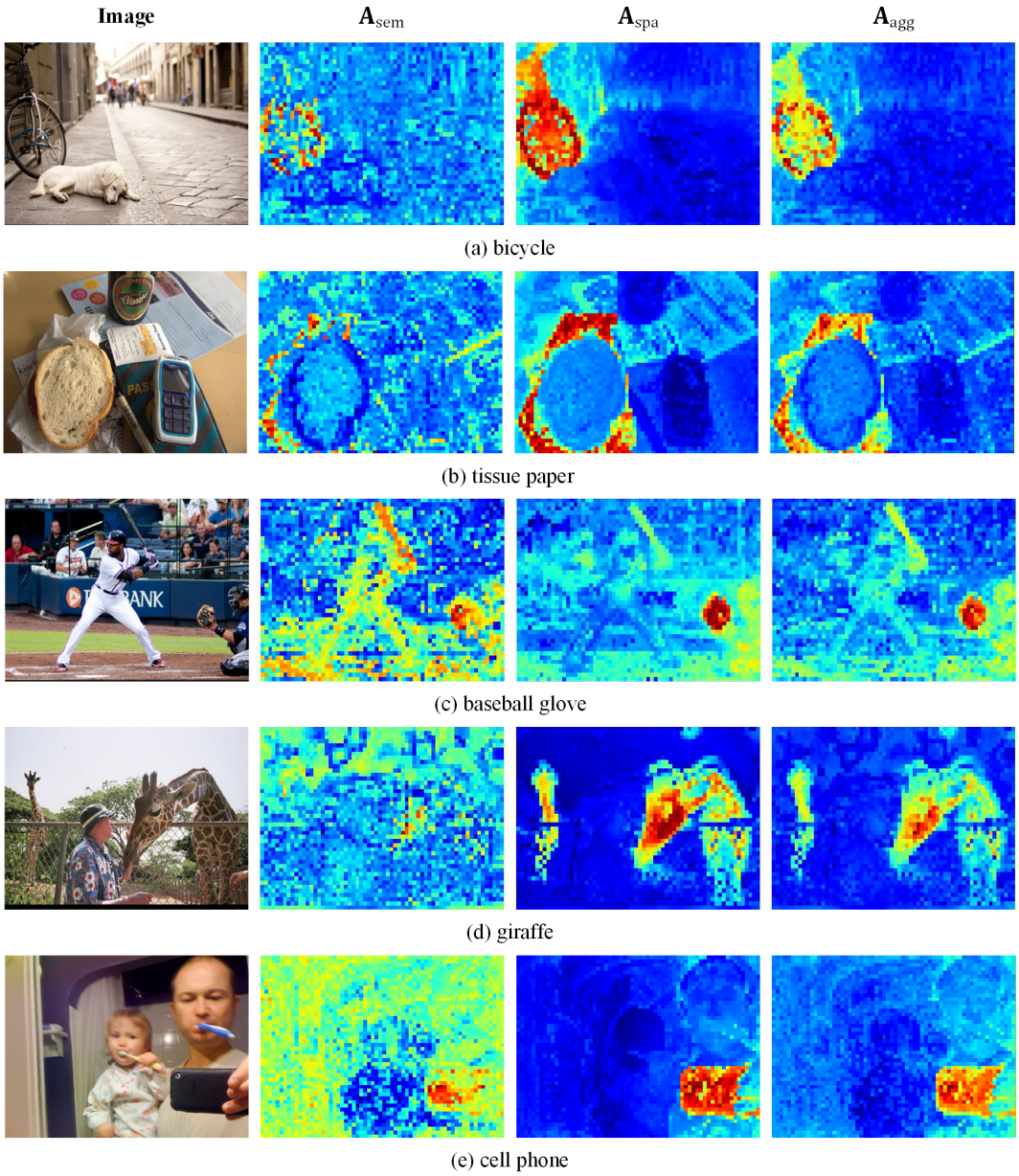


Figure 8. Visualization of five examples from the SAR module with k=7, showing semantic attention $\mathbf{A}_{\text{sem}}$, spatial correlation attention $\mathbf{A}_{\text{spa}}$, and aggregated attention $\mathbf{A}_{\text{agg}}$. Specifically, $\mathbf{A}_{\text{sem}}$ highlights text-relevant regions but often contains diffuse or noisy activations. $\mathbf{A}_{\text{spa}}$ obtained via semantic token selection provides more accurate localization of target regions, while $\mathbf{A}_{\text{agg}}$, obtained by fusing $\mathbf{A}_{\text{sem}}$ and $\mathbf{A}_{\text{spa}}$, preserves both semantic relevance and text-guided spatial precision.

relevance and text-guided spatial precision. These visualizations highlight the effectiveness of semantic token selection and spatial correlation aggregation in refining text-conditioned spatial attention.

8.2. Visualization of Prediction Results

Figure 9 presents the qualitative results on OV-COCO and OV-LVIS datasets, where yellow and blue boxes correspond to novel and base categories, respectively. It can be observed that our method, *i.e.*, FViT+SynCLIP, successfully detects both base and novel objects with high localization accuracy. These examples demonstrate the model's capability to generalize effectively to novel categories, complementing the quantitative results reported in the main paper.

(a) Qualitative results on OV-COCO

(b) Qualitative results on OV-LVIS

Figure 9. Qualitative results on (a) OV-COCO and (b) OV-LVIS datasets. Each example includes both ground-truth annotations and predictions from our FViT+SynCLIP model. Yellow boxes denote novel categories, while blue boxes indicate base categories. The model accurately localizes diverse objects and generalizes reliably to novel categories, demonstrating the effectiveness of the proposed method in dense perception.

9. Efficiency Analysis

To provide a comprehensive view of the computational characteristics of SynCLIP, we report the training time, model parameters, FLOPs and AP_{50}^{novel} on OV-COCO in Table 7. All measurements are conducted under the same hardware setup using four NVIDIA A100 GPUs with 40GB memory and an input resolution of 560. Compared with

Method	Backbone	Time(min)	Params(M)	FLOPs(G)	AP_{50}^{novel}
DeCLIP	ViT-B/16	19.3	86.3	33.7	41.1
SynCLIP	ViT-B/16	32.5	86.3	33.7	43.6
DeCLIP	ViT-L/14	103.4	304.1	155.6	46.2
SynCLIP	ViT-L/14	144.7	304.1	155.6	49.8

Table 7. Comparison of training time, model size, FLOPs and AP_{50}^{novel} between SynCLIP and DeCLIP. Although SynCLIP incurs a small pretraining-time overhead, it keeps parameters and FLOPs unchanged at inference while producing substantially higher AP_{50}^{novel}, demonstrating that the benefits arise from synonym-consistent pretraining rather than from larger model capacity or increased inference cost.

DeCLIP, SynCLIP introduces a modest increase in training time, *i.e.*, one epoch takes 32.5 minutes vs. 19.3 minutes on ViT-B/16 and 144.7 minutes vs. 103.4 minutes on ViT-L/14. This overhead stems from the semantic alignment and attention refinement modules used only during pretraining, which require forwarding frozen auxiliary networks and performing extra attention computations.

Importantly, these modules are not used during inference. As a result, SynCLIP maintains the same number of parameters and FLOPs as DeCLIP, with no increase in inference-time latency. The observed performance gains therefore stem not from increased model capacity, but from the more structured and semantically consistent pretraining objective, which enhances robustness to downstream dense perception tasks.