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

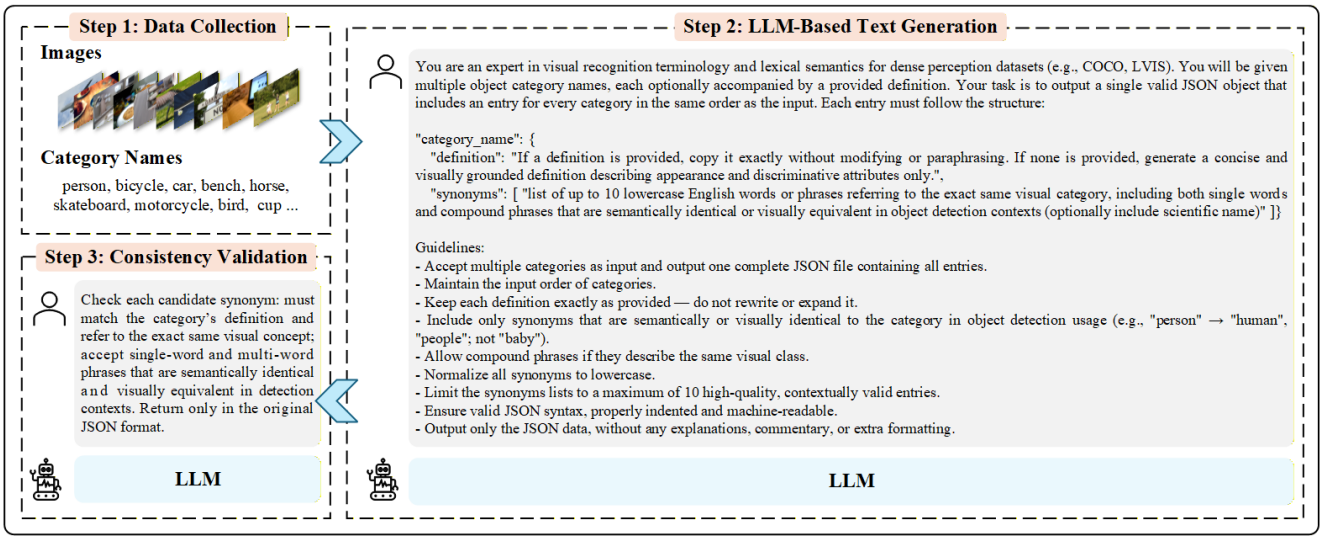

Figure 7. Overview of the SEViC construction pipeline, consisting of (1) Data Collection from COCO and LVIS datasets, which gathers the complete category vocabulary and initial textual resources; (2) LLM-based Text Generation, which expands each category into semantically consistent definitions and synonyms; 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

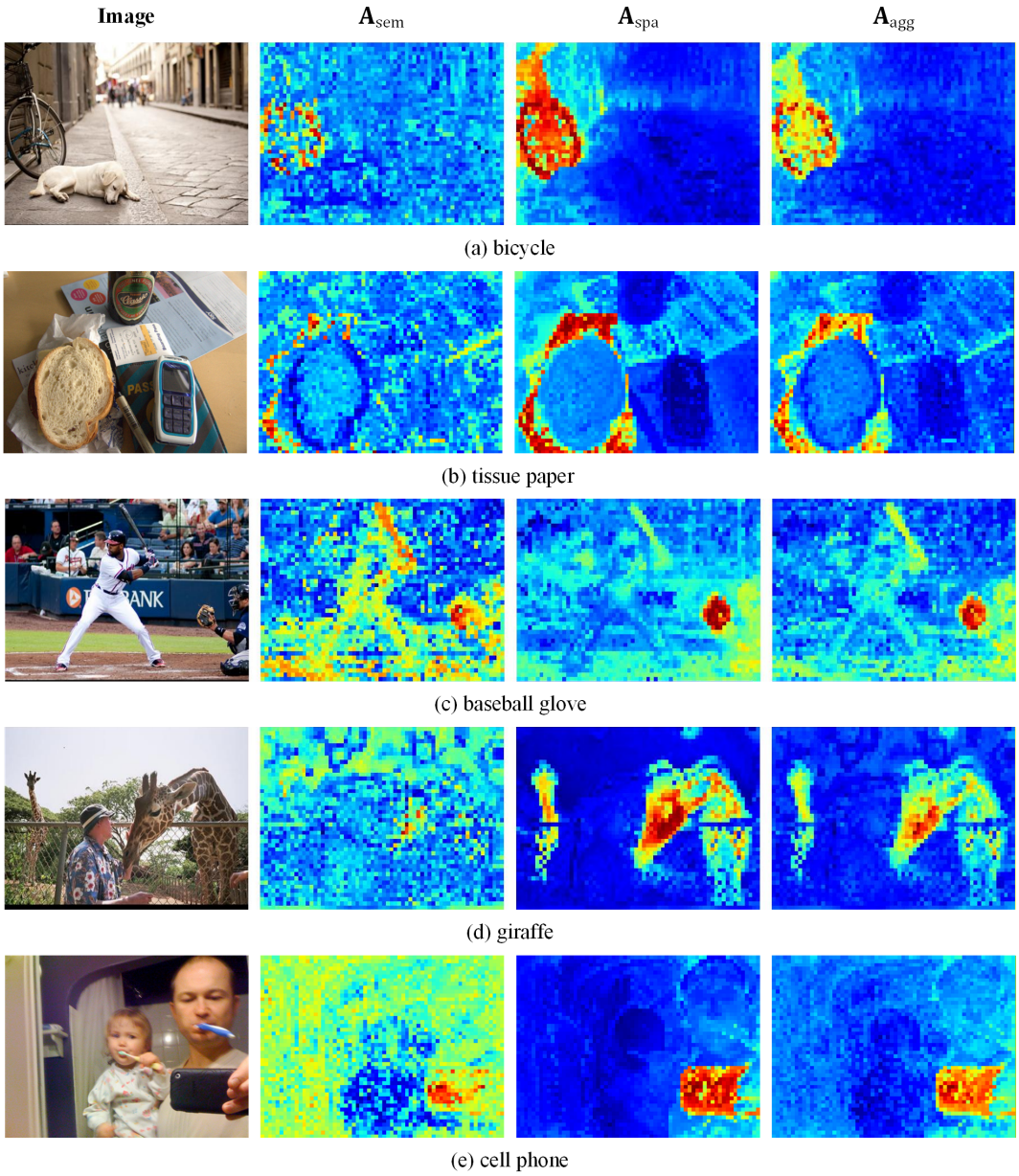

Figure 8. Visualization of five examples from the SAR module with k=7, showing semantic attention $\mathbf{A}_{\text{sem}}$, spatial correlation attention $\mathbf{A}_{\text{spa}}$, and aggregated attention $\mathbf{A}_{\text{agg}}$. Specifically, $\mathbf{A}_{\text{sem}}$ highlights text-relevant regions 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.