Learning Semantic Consistency for Open-Vocabulary Dense Perception

Li Ding[1] **Junjie Wang**[1] **Jingjun Yang**[1,2] **Jiaze Wang**[3]
Libo Qin[1] **Li Jiang**[4,2] **Zhuotao Tian**[1,2†]

[1]Harbin Institute of Technology (Shenzhen) [2]Shenzhen Loop Area Institute
[3]FitX Technology (Hong Kong) Limited [4]The Chinese University of Hong Kong (Shenzhen)

Abstract

Open-vocabulary dense perception requires grounding language-specified concepts in local regions for recognition, localization, and segmentation beyond closed category sets. CLIP learns image-level vision-language alignment, whereas dense prediction requires spatially precise and semantically reliable local features. Although recent dense CLIP adaptation methods reduce this gap through region-level semantic transfer and spatial correlation guidance, maintaining stable local semantics under contextual variation remains challenging. Specifically, the global contextual modeling in CLIP may introduce undesirable dependencies into local features. Changes in background, scale, viewpoint, or scene layout may cause identical object regions to yield inconsistent dense representations, leading to unstable correspondences and semantic drift. Moreover, spatial correlation guidance from visual foundation models captures visual relatedness between patches, but not whether coherent regions correspond to the queried textual concept, leaving concept-level dense alignment under-constrained. To address these issues, we introduce Semantic Spatial Consistency Learning **(SSCL)**, an adaptation method with two complementary objectives: 1) Context-Invariant Spatiotemporal Consistency (CISC) enforces semantic consistency for the same anchor region under varying contexts to mitigate context-induced feature drift; 2) Structure-Guided Semantic Refinement (SGSR) leverages spatial affinities from a frozen visual foundation model to refine CLIP text-patch responses, producing text-conditioned dense supervision with improved spatial coherence and concept-level alignment. Experiments on video instance segmentation, region classification, semantic segmentation, and object detection show consistent gains over the baseline methods, demonstrating the effectiveness and generalization capability of the proposed SSCL. Code is available at https://github.com/LiWithDream/sscl.

1 Introduction

Open-vocabulary dense perception aims to identify, localize, and segment visual concepts specified via natural language without confining predictions to a fixed set of predefined categories. Leveraging large-scale image-text pretraining, CLIP [1] has become a foundation model for tasks including open-vocabulary semantic segmentation [2–5], object detection [6–8], and video instance segmentation [9–11]. Nevertheless, CLIP primarily optimizes global image-text alignment at the image level, posing substantial challenges for dense prediction tasks, which demand reliable feature representations in local spatial regions rather than only from a holistic view.

Recent efforts on dense CLIP adaptation [12, 13, 8, 14, 15] have made significant strides toward addressing this issue. Specifically, RegionCLIP [13] improves region-level visual representations

†Corresponding author (tianzhuotao@hit.edu.cn).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

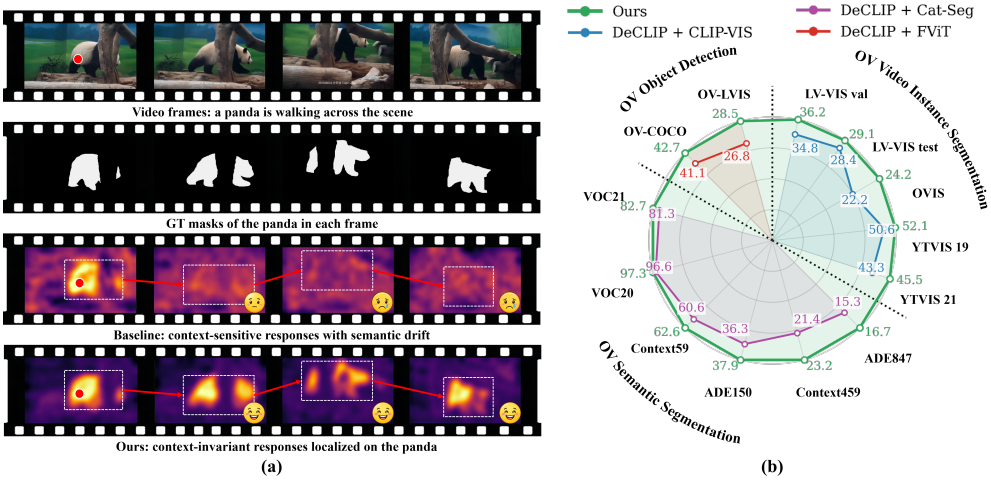

Figure 1: **Qualitative and quantitative overview of SSCL. (a)** Given an anchor point (●) initialized on the target in the first frame, baseline dense CLIP responses become diffuse and drift under context changes, whereas SSCL maintains spatially concentrated and coherent responses. **(b)** Performance summary on open-vocabulary dense perception benchmarks, including video instance segmentation, region classification, semantic segmentation, and object detection. SSCL consistently improves over dense CLIP adaptation baselines under comparable evaluation settings.

by explicitly aligning image regions with textual tokens. CLIPSelf [8] utilizes image-level representations from cropped regions as teachers for corresponding representations pooled from dense features, thereby enhancing local region recognition without requiring explicit region-text supervision. CLIP-DINOiser [14] and DeCLIP [15] further refine dense CLIP representations by integrating region-level self-distillation with spatial correlation guidance from pretrained visual foundation models (VFMs) [16–19]. The core challenge in adapting CLIP for dense perception tasks lies in enhancing spatial awareness while preserving generalization capability.

Key Observation. Despite CLIP's global contextual modeling being effective for image-level recognition, it unintentionally introduces problematic dependencies into local feature representations. In particular, representations of patch tokens become sensitive to variations in background, object scale, viewpoint, and scene layout. As illustrated in Figure 1(a), identical object regions can yield inconsistent dense representations under varying visual contexts, leading to unstable correspondences and semantic drift. This issue is particularly acute in open-vocabulary video perception tasks [9–11], where frame-to-frame contextual fluctuations significantly degrade performance stability.

Furthermore, although existing studies [14, 15] demonstrate that visual affinities derived from VFMs serve as effective learning guidance, such affinities are computed solely based on visual inputs without explicit conditioning on textual concepts. Consequently, their utility in generating reliable concept-specific dense responses is inherently constrained. These observations collectively highlight a critical research question: *How can we obtain robust feature representations that integrate concept-aware visual guidance, thus maintaining semantic consistency and stability despite contextual variations across diverse visual scenarios?*

Our Approach. In this work, we propose Semantic Spatial Consistency Learning **(SSCL)**, a framework designed to enhance open-vocabulary dense perception through two complementary objectives. The first objective, Context-Invariant Spatiotemporal Consistency (CISC), promotes robust local representations by enforcing semantic stability across varying contexts, mitigating feature drift caused by changes in background or scene composition. The second objective, Structure-Guided Semantic Refinement (SGSR), utilizes region affinities from a frozen VFM to refine the text-patch responses of CLIP, thereby boosting spatial coherence and enhancing concept-level awareness in semantic predictions. Unlike existing methods [14, 15], which employ VFMs primarily as visual regularizers, SGSR explicitly integrates them to improve spatial coherence and conceptual alignment of text-conditioned semantic outputs.

Experiments on tasks including open-vocabulary video instance segmentation, region classification, semantic segmentation, and object detection consistently demonstrate superior performance compared to the existing methods, as illustrated in Figure 1(b). These performance gains remain robust across varying backbone scales and downstream tasks, highlighting the generalizability of the proposed method. To summarize, our contributions are as follows.

- As existing dense CLIP adaptation methods struggle to maintain stable local semantics under contextual variations, we identify context-induced feature drift as a key obstacle to reliable open-vocabulary dense perception.
- To mitigate this issue, we introduce SSCL, a unified adaptation framework integrating CISC and SGSR to enhance robustness against contextual variations and improve concept-aware dense alignment.
- Extensive experiments across four tasks demonstrate that SSCL consistently outperforms the existing methods, highlighting its strong transferability and effectiveness.

2 Background and Motivation

2.1 Preliminaries

Open-Vocabulary (OV) dense perception extends dense prediction from a fixed set of annotated categories to a broader semantic space defined by natural language. CLIP-based models, with their strong vision-language alignment, naturally support this OV capability. In this section, we introduce relevant preliminaries and highlight two empirical observations that motivate our approach.

CLIP for OV Semantics. CLIP [1] learns a shared embedding space for images and text from large-scale image-text data. During inference, visual features can be directly matched with textual prompts, enabling category specification via natural language rather than a fixed classifier. While this makes CLIP particularly suitable for OV tasks, its training primarily utilizes image-level supervision, lacking adaptation for patch-level features to maintain spatial stability, local discriminativeness, or region-awareness. Consequently, these limitations hinder its efficacy in dense perception tasks.

Dense CLIP Adaptation. Existing dense CLIP adaptation methods, such as CLIPSelf [8] and DeCLIP [15], commonly adopt a region-level self-distillation pipeline. Specifically, given an image $\mathbf{I}$, a trainable student CLIP image encoder produces a dense feature map $\mathbf{X} = f_\theta(\mathbf{I})$, where $\mathbf{X}$ is an $H \times W$ grid of d-dimensional patch features. For each sampled region r, RoI pooling aggregates the corresponding dense features into a region representation,

$$\bar{\mathbf{x}}_r = \mathrm{RoIPool}(\mathbf{X}, r). \tag{1}$$

In parallel, a frozen CLIP teacher encoder processes the image crop corresponding to each region, producing $\mathbf{z}_r = f_{\mathrm{T}}(\mathbf{I}_r)$, where $\mathbf{I}_r$ denotes the cropped region and $\mathbf{z}_r \in \mathbb{R}^d$ is the teacher's [CLS] representation. This forms an asymmetric self-distillation pipeline, where crop-level CLIP targets supervise RoI-pooled dense student representations. Self-distillation is then performed by aligning the student's pooled regional feature with the corresponding teacher representation:

$$\mathcal{L}_{\mathrm{sd}} = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} d\left(\frac{\bar{\mathbf{x}}_r}{\|\bar{\mathbf{x}}_r\|_2}, \frac{\mathbf{z}_r}{\|\mathbf{z}_r\|_2} \right). \tag{2}$$

This region-level self-distillation enables the student encoder to acquire localized dense features aligned with OV semantics. In our approach, we adopt this framework as the baseline training strategy and further enhance it with supplementary objectives designed to strengthen contextual robustness and improve region-level text-visual alignment.

2.2 Key Observations

While existing dense CLIP adaptation methods have significantly advanced OV dense perception [12, 8, 13, 15, 14], they still face two critical challenges: limited robustness against spatiotemporal and contextual variations, and unreliable dense text-visual alignment at the region level. In the following,

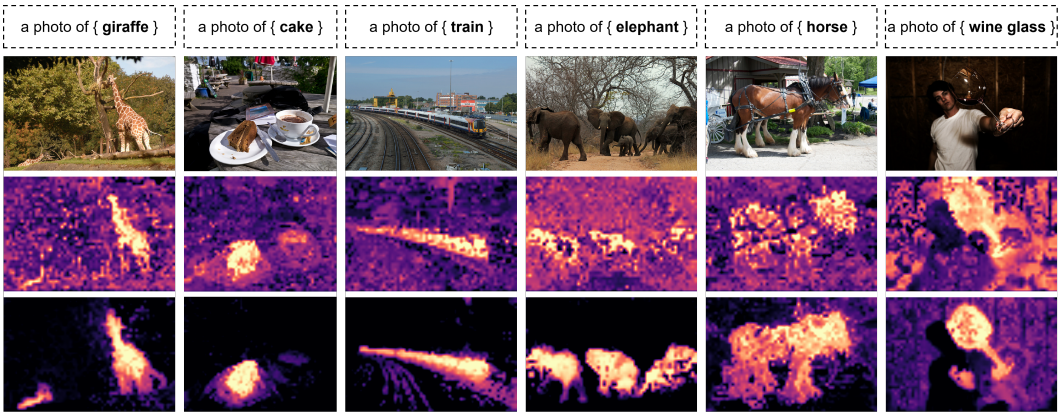

Figure 2: **Visualization of under-constrained text-conditioned region alignment.** From top to bottom, we show the queried text prompts, input images, raw CLIP text-patch responses, and SGSR-refined responses. The raw text-patch responses provide concept-conditioned semantic cues, but they are often spatially noisy, fragmented, or mixed with visually related background regions. **SGSR** refines these responses by propagating the text-conditioned cues through VFM-derived region affinities, producing responses that are more spatially coherent and better aligned with the queried object regions.

we present two empirical observations that elucidate these limitations and provide the motivation for our proposed approach.

Observation 1: Dense CLIP features are sensitive to contextual changes. Applying dense-adapted CLIP to video perception tasks reveals significant semantic instability across frames. Although the same object persists between adjacent frames, extensive variations in position, scale, viewpoint, and surrounding background frequently occur. Ideally, local representations should remain semantically consistent despite these visual perturbations. However, as depicted in Figure 1, initializing a query on an object in the initial frame results in progressively unstable similarity maps in subsequent frames. Specifically, peak activation gradually drifts away from the original object, accompanied by a noticeable decline in overall activation. This indicates that dense CLIP features are sensitive to contextual influences, compromising their semantic robustness across varying visual conditions.

We attribute this instability largely to the global self-attention mechanism employed by Vision Transformers (ViT) [20]. In ViT, each patch token attends globally to all other tokens within an image, thus its representation is influenced not only by immediate visual details but also by broader contextual information such as background elements, nearby objects, and the overall scene composition. While global contextual modeling is advantageous for holistic image-level semantic recognition, it undermines the robustness necessary for stable dense perception tasks. Consequently, identical local regions yield inconsistent representations when embedded in different visual contexts.

Motivated by this observation, we propose the Context-Invariant Spatiotemporal Consistency strategy in Section 3.2, explicitly designed to enforce context-invariant consistency for local regions across diverse visual observations.

Observation 2: Text-conditioned region alignment remains under-constrained. The previous observation shows that dense CLIP adaptation should preserve stable local semantics under changing visual contexts. In OV dense perception, it is also crucial that these local semantics reliably align with the queried textual concepts. Typically, this alignment is achieved by matching CLIP text embeddings to dense patch features, generating text-patch response maps corresponding to various concepts. However, given that CLIP is primarily pretrained with image-level supervision, these response maps often provide only coarse semantic hints rather than precise, region-specific predictions. As illustrated in Figure 2, raw CLIP text-patch responses may only partially highlight discriminative object regions, inadvertently extend to visually related background areas, or fail altogether to cover the entire object.

Although recent dense CLIP adaptation methods enhance local regularity and spatial coherence of CLIP patch features by integrating spatial correlations derived from VFMs [14, 15], they do not inherently indicate whether visually coherent regions align semantically with the queried textual con-

cepts. Consequently, we introduce the Structure-Guided Semantic Refinement module in Section 3.3. Instead of employing VFM-derived affinities purely as a visual regularizer, we utilize them explicitly to refine CLIP text-patch responses, ensuring concept-aware visual guidance.

3 Method

3.1 Overview

As illustrated in Figure 3, our framework augments the base dense CLIP adaptation pipeline with two complementary components. CISC regularizes the dense CLIP representation by enforcing representation consistency across context-varying pseudo-temporal views, thereby improving the semantic stability of local features under changing context. Then, SGSR further improves region-level dense text-visual alignment by refining CLIP text-conditioned responses with VFM-derived region affinities, yielding cleaner and more region-consistent supervision during the training process. Together, these two strategies complement the self-distillation objective by enhancing both contextual robustness and region-level cross-modal alignment. The following subsections detail each component and the resulting joint training objective.

3.2 Context-Invariant Spatiotemporal Consistency

To enhance feature stability against contextual variations, we propose the CISC strategy. The core concept involves enforcing representation consistency for identical anchor regions across multiple observations with varying surrounding contexts, thus encouraging the model to maintain robust local semantics despite contextual shifts.

Anchor-Centric Pseudo-Temporal Modeling. As shown in Figure 3(a), the anchor view denotes a tight crop around a local image region, while the pseudo-temporal views are larger crops that contain the same region with different surrounding contexts. CISC aims to make the region representation consistent across these views. Although real videos provide such observations, they also introduce non-rigid deformation, occlusion, motion blur, camera motion, and correspondence errors, which can make patch-level matching noisy. We therefore synthesize pseudo-temporal observations from static images, so that the anchor content is fixed while the surrounding context is controlled.

Specifically, given an image $\mathbf{I}$ and an anchor region r, we sample k patch-aligned crop-and-resize transforms $\{\mathcal{T}_k\}_{i=1}^{k}$. Each transform keeps the anchor inside the crop while varying the crop center and scale, producing pseudo-temporal observations $\mathbf{I}^{(i)} = \mathcal{T}_i(\mathbf{I})$ with different surrounding contexts. Since the transforms are patch-aligned and known, the anchor can be exactly traced across views without external tracking or temporal annotations.

Consistency Objective. During training, we use local image regions as anchors. For each anchor r, we construct a tight, patch-aligned anchor view $\mathbf{I}_r$, which preserves the anchor content while suppressing view-specific surrounding context. This view serves as the anchor-local teacher target. The frozen teacher f_{T} encodes the canonical anchor view, while the student f_θ encodes each pseudo-temporal view:

$$\mathbf{X}_r^{\mathrm{T}} = f_{\mathrm{T}}(\mathbf{I}_r) \in \mathbb{R}^{H_r \times W_r \times d}, \qquad \mathbf{X}^{(i)} = f_\theta(\mathbf{I}^{(i)}) \in \mathbb{R}^{H_i \times W_i \times d}. \tag{3}$$

This follows the teacher-student decomposition of the base self-distillation pipeline, but changes the supervision from image- or RoI-level alignment to spatially resolved anchor alignment. Because $\mathcal{T}_i$ is patch-aligned and known, we trace the anchor on the student feature grid, crop the corresponding feature region, and resize it to the teacher grid:

$$\mathbf{X}_r^{(i)} = \mathrm{Resize}\left(\mathbf{X}^{(i)}\big|_{\mathcal{T}_i(r)},\, H_r, W_r\right) \in \mathbb{R}^{H_r \times W_r \times d}, \tag{4}$$

where $\mathcal{T}_i(r)$ denotes the anchor region mapped onto the feature grid of the i-th pseudo-temporal view. We then apply a point-to-point InfoNCE loss over the aligned anchor grid $\mathcal{G}_r$:

$$\mathcal{L}_{\mathrm{CISC}} = -\frac{1}{k|\mathcal{G}_r|} \sum_{i=1}^{k} \sum_{p \in \mathcal{G}_r} \log \frac{\exp\left(\left\langle \bar{\mathbf{x}}_r^{(i)}(p), \bar{\mathbf{x}}_r^{\mathrm{T}}(p) \right\rangle / \tau\right)}{\sum_{q \in \mathcal{G}_r} \exp\left(\left\langle \bar{\mathbf{x}}_r^{(i)}(p), \bar{\mathbf{x}}_r^{\mathrm{T}}(q) \right\rangle / \tau\right)}. \tag{5}$$

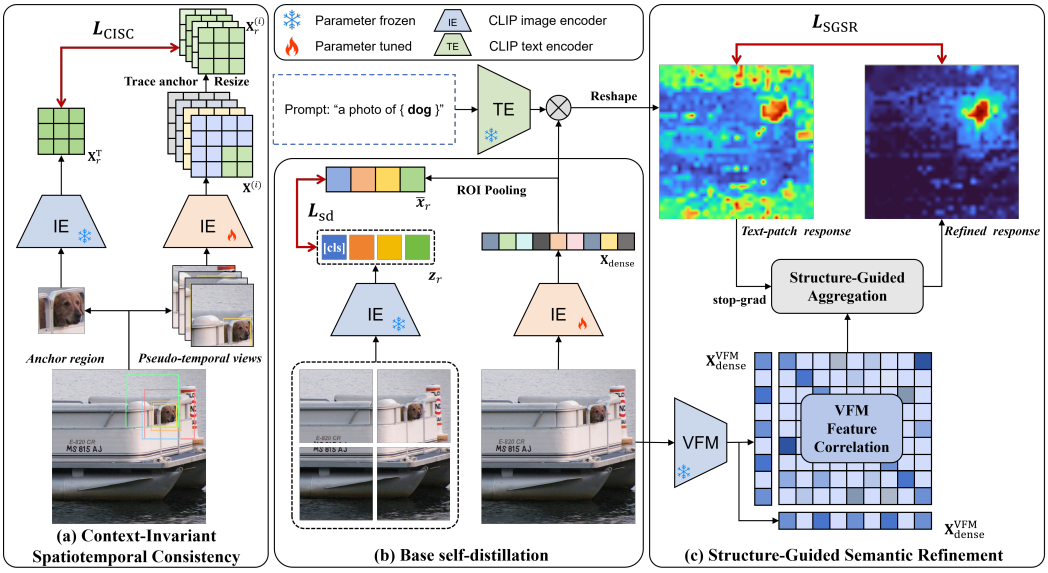

Figure 3: **Overview of SSCL.** SSCL extends self-distillation with two complementary objectives for open-vocabulary dense perception. CISC learns context-invariant anchor representations by aligning spatially corresponding dense features of the same anchor across controlled pseudo-temporal views. SGSR refines noisy CLIP text-patch responses with VFM-induced feature correlations, providing structure-aware semantic supervision for more coherent region-level text-visual alignment.

Here, $\bar{\mathbf{x}}$ denotes ℓ_2-normalized features, p is matched to the same anchor-grid location in the teacher feature, and other locations q serve as spatial negatives. This objective encourages context-invariant anchor representations while preserving spatial discriminability within the anchor.

3.3 Structure-Guided Semantic Refinement

The anchor-consistency objective stabilizes local visual features under context variation. For open-vocabulary dense prediction, these features must further produce coherent responses to queried textual concepts. As discussed in Section 2.2, CLIP text-patch similarities are query-aware but often spatially fragmented, whereas VFM affinities capture region structure but are query-agnostic. As shown in Figure 3(c), SGSR combines these complementary signals by using CLIP responses as semantic seeds and propagating them through VFM-induced structural affinities. In this way, the refined supervision preserves the queried text semantics while improving region-level coherence.

During training, for an image with image-level category set $\mathcal{C}_\mathbf{I}$, we apply SGSR to each selected category $c \in \mathcal{C}_\mathbf{I}$. This category selection only uses image-level labels and does not introduce mask-, box-, or pixel-level supervision.

Structure-Aware Semantic Refinement. For a category c, we instantiate a text prompt τ_c using the fixed template `a photo of {category}`, and encode it with the CLIP text encoder as $\mathbf{t}_c = g_{\text{text}}(\tau_c)$. Given the student dense feature map $\mathbf{X}$, let $\mathbf{x}_p \in \mathbb{R}^d$ denote its feature vector at patch location $p \in \mathcal{G}$. The text-patch response for category c is computed as:

$$\mathbf{R}_c(p) = \left\langle \frac{\mathbf{x}_p}{\|\mathbf{x}_p\|_2}, \frac{\mathbf{t}_c}{\|\mathbf{t}_c\|_2} \right\rangle. \tag{6}$$

To obtain region-level visual structure, we feed the same image into a frozen VFM and extract dense tokens $\mathbf{X}^{\text{VFM}}_{\text{dense}} = h_{\text{VFM}}(\mathbf{I}) \in \mathbb{R}^{H_v \times W_v \times d_v}$. After aligning the VFM token grid with the CLIP patch grid, let $\mathbf{x}^{\text{VFM}}_p \in \mathbb{R}^{d_v}$ be the VFM feature at location $p \in \mathcal{G}$. We define the affinity score as:

$$\mathbf{S}_{pq} = \left\langle \frac{\mathbf{x}^{\text{VFM}}_p}{\|\mathbf{x}^{\text{VFM}}_p\|_2}, \frac{\mathbf{x}^{\text{VFM}}_q}{\|\mathbf{x}^{\text{VFM}}_q\|_2} \right\rangle. \tag{7}$$

Then, we convert the text-patch response into a spatial distribution and obtain the refined response map $\widetilde{\mathbf{R}}_c$ through structure-guided aggregation:

$$\alpha_c(p) = \frac{\exp(\mathbf{R}_c(p))}{\sum_{p' \in \mathcal{G}} \exp(\mathbf{R}_c(p'))}, \qquad \widetilde{\mathbf{R}}_c(p) = \sum_{q \in \mathcal{G}} \mathbf{S}_{pq}\, \alpha_c(q). \tag{8}$$

Here, $\widetilde{\mathbf{R}}_c(p)$ measures how strongly patch p is structurally connected, under the VFM representation, to patches that CLIP associates with category c. This refinement transfers query-specific semantic evidence across visually coherent regions, yielding region-consistent responses while reducing background leakage and fragmented activations.

Refined Semantic Target. Given the refined response $\widetilde{\mathbf{R}}_c$, we regularize the original CLIP text-patch response $\mathbf{R}_c$ toward its detached structure-refined counterpart. For an image with valid category set $\mathcal{Y}$ and valid spatial locations $\mathcal{V}$, the SGSR loss is defined as

$$\mathcal{L}_{\text{SGSR}} = \frac{1}{|\mathcal{Y}|} \sum_{c \in \mathcal{Y}} \frac{1}{|\mathcal{V}|} \sum_{p \in \mathcal{V}} \rho\left(\mathbf{R}_c(p), \text{sg}\left(\widetilde{\mathbf{R}}_c(p)\right)\right), \tag{9}$$

where $\rho(\cdot, \cdot)$ is the Smooth L1 loss and $\text{sg}(\cdot)$ denotes stop-gradient. The detached target ensures that gradients flow only through the student response $\mathbf{R}_c$.

3.4 Overall Training Objective

The overall training objective integrates the self-distillation loss in Eq. (2), the CISC loss in Eq. (5), and the SGSR loss in Eq. (9), formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{sd}} + \lambda_1 \mathcal{L}_{\text{CISC}} + \lambda_2 \mathcal{L}_{\text{SGSR}} \tag{10}$$

where $\lambda_1 = 0.25$ and $\lambda_2 = 1$ are balancing coefficients. Specifically, $\mathcal{L}_{\text{sd}}$ transfers open-vocabulary semantics, $\mathcal{L}_{\text{CISC}}$ maintains feature consistency across varying contexts, and $\mathcal{L}_{\text{SGSR}}$ refines coarse semantic responses using structural cues. Joint optimization of these losses yields dense representations with enhanced semantic alignment, spatial coherence, and robustness against contextual variations.

During training, the CLIP teacher and VFM remain frozen, updating only the student image encoder. It is worth noting that the proposed methods are exclusively applied during training, introducing no additional computational overhead at inference. Besides, the above-mentioned hyperparameters remain consistent across all benchmarks and models, demonstrating the strong generalization capability of the proposed designs.

4 Experiments

4.1 Benchmarks and Evaluation

We evaluate the adapted dense CLIP representation on four OV dense perception tasks: video instance segmentation, region classification, semantic segmentation, and object detection. Unless otherwise stated, we follow the official evaluation protocol and report the primary metric of each benchmark. Detailed descriptions of the datasets, evaluation protocols, model configuration, and implementation details are provided in the Appendix due to space limitations.

4.2 Main Results

OV Video Instance Segmentation. For video instance segmentation, we adopt CLIP-VIS [9] and replace its visual backbone with the adapted dense encoder. Table 1 reports results on LV-VIS [21], OVIS [22], and YouTube-VIS 2019/2021 [23], including LV-VIS novel-category AP_N when available, and compares against recent open-vocabulary VIS baselines such as OV2Seg [10] and OVFormer [11], where our method achieves state-of-the-art performance across all benchmarks.

OV Region Classification. For region classification, we evaluate the adapted dense encoder on COCO-Panoptic [29] validation by pooling features within annotated regions and classifying them via similarity to text embeddings. Table 2 reports Acc@1 and Acc@5 for boxes, thing masks, and stuff masks, comparing against recent dense CLIP adaptation baselines such as CLIPSelf [8] and DeCLIP [15], where our method attains state-of-the-art accuracy on all metrics.

Table 1: **OV Video Instance Segmentation Results.** We compare methods on LV-VIS, OVIS, and YouTube-VIS 2019/2021, reporting overall AP on all benchmarks and LV-VIS novel-category AP_N when available.

Method	Training Data	Backbone	LV-VIS val		LV-VIS test		OVIS	YTVIS19	YTVIS21
			AP	AP_N	AP	AP_N	AP	AP	AP
Detic [24]+SORT [25]	LVIS	SwinB	12.8	6.6	9.4	4.7	11.7	23.8	21.6
Detic [24]+OWTB [26]	LVIS	SwinB	14.5	11.8	13.6	5.5	30.0	9.7	11.4
Detic [24]+XMem [27]	LVIS	SwinB	16.3	10.6	13.1	7.7	—	—	—
OV2Seg [10]	LVIS	SwinB	21.1	16.3	16.4	11.5	17.5	37.6	33.9
OVFormer [11]	LVIS	SwinB+ViT-B	—	—	—	—	21.3	44.3	37.6
CLIP-VIS [9]	LVIS	ConvNeXt-B	32.2	40.2	25.3	30.6	18.5	42.1	37.9
CLIP-VIS [9]+DeCLIP [28]	LVIS	ViT-B/16	34.8	44.3	28.4	34.0	22.2	50.6	43.3
CLIP-VIS [9]+DeCLIP [28]	LVIS	ViT-L/14	37.7	45.9	30.9	36.9	29.3	54.8	49.1
CLIP-VIS+Ours	LVIS	ViT-B/16	**36.2**	**44.9**	**29.1**	**35.0**	**24.2**	**52.1**	**45.5**
CLIP-VIS+Ours	LVIS	ViT-L/14	**39.5**	**46.9**	**32.3**	**39.1**	**31.0**	**55.4**	**51.0**

Table 2: **OV Region Classification Results.** We compare methods on COCO-Panoptic validation, reporting Acc@1 and Acc@5 on boxes, thing masks, and stuff masks.

Method	Backbone	Boxes		Thing Masks		Stuff Masks	
		Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5
EVA-CLIP [30]	ViT-B/16	18.2	33.2	20.6	36.5	18.4	43.5
EVA-CLIP [30]	ViT-L/14	56.7	78.0	59.0	79.8	20.8	41.9
CLIPSelf [8]	ViT-B/16	72.1	91.3	74.4	91.8	46.8	80.2
CLIPSelf [8]	ViT-L/14	77.1	93.3	78.7	93.7	44.4	78.3
DeCLIP [15]	ViT-B/16	75.3	93.0	77.1	91.4	52.7	82.1
DeCLIP [15]	ViT-L/14	80.4	95.2	80.2	94.5	54.6	84.2
Ours	ViT-B/16	**77.1**	**94.0**	**78.3**	**93.6**	**57.9**	**86.4**
Ours	ViT-L/14	**81.2**	**96.6**	**82.8**	**95.9**	**58.3**	**86.6**

OV Semantic Segmentation. For semantic segmentation, we evaluate the adapted dense encoder with CAT-Seg [4] and H-CLIP [5] on ADE20K [31], PASCAL-Context [32], and VOC [33], replacing only the image encoder and otherwise following their original settings. Table 3 summarizes the results, showing that our method achieves state-of-the-art performance across all reported benchmarks.

Table 3: **OV Semantic Segmentation Results.** We compare methods across ADE20K, PASCAL-Context, and VOC benchmarks, reporting performance on both expanded and standard label spaces when available and listing the training set used by each method.

Method	Backbone	Training Set	ADE847	Context459	ADE150	Context59	VOC20	VOC21
ZegFormer [34]	ViT-B/16	COCO-Stuff	5.6	10.4	18.0	45.5	89.5	65.5
ZSseg [35]	ViT-B/16	COCO-Stuff	7.0	—	20.5	47.7	88.4	—
OVSeg [3]	ViT-L/14	COCO-Stuff	9.0	12.4	29.6	55.7	94.5	—
SAN [36]	ViT-L/14	COCO-Stuff	13.7	17.1	33.3	60.2	95.5	—
ODISE [37]	ViT-L/14	COCO-Panoptic	11.1	14.5	29.9	57.3	—	84.6
MAFT [38]	ConvNeXt-L	COCO-Stuff	13.1	17.0	34.4	57.5	93.0	—
FC-CLIP [39]	ConvNeXt-L	COCO-Panoptic	14.8	18.2	34.1	58.4	95.4	81.8
FrozenSeg [40]	ConvNeXt-L	COCO-Panoptic	14.8	19.7	34.4	—	—	82.5
CAT-Seg [4]	ViT-B/16	COCO-Stuff	12.0	19.0	31.8	57.5	94.6	77.3
CAT-Seg [4]	ViT-L/14	COCO-Stuff	16.0	23.8	37.9	63.3	97.0	82.5
H-CLIP [5]	ViT-B/16	COCO-Stuff	12.5	19.4	32.4	57.9	95.2	78.2
H-CLIP [5]	ViT-L/14	COCO-Stuff	16.5	24.2	38.4	64.1	97.7	83.2
CAT-Seg [4]+DeCLIP [15]	ViT-B/16	COCO-Stuff	15.3	21.4	36.3	60.6	96.6	81.3
CAT-Seg [4]+DeCLIP [15]	ViT-L/14	COCO-Stuff	17.6	25.9	40.7	63.9	97.7	83.9
CAT-Seg+Ours	ViT-B/16	COCO-Stuff	16.3	22.6	37.6	62.1	97.2	82.2
CAT-Seg+Ours	ViT-L/14	COCO-Stuff	18.9	26.7	41.9	64.5	97.6	85.3
H-CLIP+Ours	ViT-B/16	COCO-Stuff	**16.7**	**23.2**	**37.9**	**62.6**	**97.3**	**82.7**
H-CLIP+Ours	ViT-L/14	COCO-Stuff	**19.1**	**27.3**	**42.5**	**64.7**	**97.8**	**85.6**

Table 4: **OV Object Detection Results.** We compare methods on OV-COCO and OV-LVIS, reporting novel-category AP_{50} on OV-COCO and rare-category AP on OV-LVIS.

(a) OV-COCO benchmark

Method	Backbone	AP_{50}^{novel}
ViLD [6]	RN50	27.6
Detic [24]	RN50	27.8
F-VLM [7]	RN50	28.0
OV-DETR [43]	RN50	29.4
BARON-KD [44]	RN50	34.0
CORA [45]	RN50x4	41.7
CORA+ [45]	RN50x4	43.1
PromptOVD [46]	ViT-B/16	30.6
RO-ViT [47]	ViT-L/16	33.0
CFM-ViT [48]	ViT-L/16	34.1
F-ViT+CLIPSelf [8]	ViT-B/16	37.6
F-ViT+CLIPSelf [8]	ViT-L/14	44.3
F-ViT+DeCLIP [15]	ViT-B/16	41.1
F-ViT+DeCLIP [15]	ViT-L/14	46.2
F-ViT+Ours	ViT-B/16	**42.7**
F-ViT+Ours	ViT-L/14	**47.1**

(b) OV-LVIS benchmark

Method	Backbone	AP_r
ViLD [6]	RN50	16.6
OV-DETR [43]	RN50	17.4
BARON-KD [44]	RN50	22.6
CORA+ [45]	RN50x4	28.1
F-VLM [7]	RN50x64	32.8
PromptOVD [46]	ViT-B/16	23.1
OW-ViT [49]	ViT-L/14	25.6
RO-ViT [47]	ViT-L/16	32.4
CFM-ViT [48]	ViT-L/16	33.9
RO-ViT [47]	ViT-H/16	34.1
F-ViT+CLIPSelf [8]	ViT-B/16	25.3
F-ViT+CLIPSelf [8]	ViT-L/14	34.9
F-ViT+DeCLIP [15]	ViT-B/16	26.8
F-ViT+DeCLIP [15]	ViT-L/14	37.2
F-ViT+Ours	ViT-B/16	**28.5**
F-ViT+Ours	ViT-L/14	**38.6**

OV Object Detection. For object detection, we adopt F-ViT, replace its visual backbone with the adapted dense encoder, and otherwise follow the original benchmark settings on OV-COCO [41] and OV-LVIS [42]. Table 4 reports the primary metric of each benchmark, where our method reaches state-of-the-art results on both benchmarks.

4.3 Ablation Study

We conduct ablations with EVA-CLIP ViT-B/16 and DINOv2-B under the same training setting as the main experiments. To verify the contribution of the two proposed objectives, Table 5 compares the base self-distillation objective with variants that add CISC alone, SGSR alone, and both objectives. CISC and SGSR each improve the base dense CLIP adaptation baseline, and their combination yields the best overall performance, indicating that contextual robustness and structure-aware semantic refinement provide complementary benefits. Additional ablations on the CISC loss formulation, SGSR design choices, and loss-weight sensitivity are provided in the appendix.

Table 5: **Complementarity ablation on CISC and SGSR.** Region classification is evaluated on COCO-Panoptic, and semantic segmentation uses H-CLIP on ADE150 and Context-59.

Setting	Region Classification (Acc@1)			Semantic Segmentation (mIoU)	
	Boxes	Thing	Stuff	ADE150	Context59
Base self-distillation	73.1	70.8	47.6	34.5	59.0
Base + CISC	76.0	77.4	57.8	36.3	61.4
Base + SGSR	74.5	72.4	50.9	35.0	60.4
Full model	**77.1**	**78.3**	**57.9**	**37.9**	**62.6**

5 Concluding Remarks

Summary. In this work, we presented SSCL, a dense CLIP adaptation framework for open-vocabulary dense perception that improves both contextual robustness and concept-aware spatial alignment. SSCL introduces CISC to reduce context-induced semantic drift and SGSR to refine CLIP text-patch responses with VFM-derived region affinities. Experiments across video instance segmentation, region classification, semantic segmentation, and object detection show consistent improvements over dense CLIP adaptation baselines under comparable settings. These results suggest that jointly addressing context sensitivity and text-conditioned region alignment is an effective direction for transferring image-level vision-language models to open-vocabulary dense perception.

Limitation and Broader Impacts. Although CISC encourages context-invariant local representations through anchor-centric pseudo-temporal views, such synthesized views may still provide only

an approximation of the contextual changes observed in real videos. Future work will explore richer context modeling to further strengthen robustness under dynamic visual conditions. As a general representation adaptation method for visual understanding, SSCL may help reduce dense annotation requirements, and we do not anticipate obvious negative societal impacts.

Acknowledgments and Disclosure of Funding

This work was supported by the Shenzhen Science and Technology Program (ZDCY20250901113000001, KJZD20240903102901003) and by the Guangdong Basic and Applied Basic Research Foundation (2025A1515011546).

References

[1] 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(ICML)*, 2021.

[2] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022.

[3] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition(CVPR)*, 2023.

[4] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2024.

[5] Zelin Peng, Zhengqin Xu, Zhilin Zeng, Yu Huang, Yaoming Wang, and Wei Shen. Parameter-efficient fine-tuning in hyperspherical space for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition(CVPR)*, 2025.

[6] 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.

[7] 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.

[8] 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 *International Conference on Learning Representations (ICLR)*, 2024.

[9] Wenqi Zhu, Jiale Cao, Jin Xie, Shuangming Yang, and Yanwei Pang. Clip-vis: Adapting clip for open-vocabulary video instance segmentation. *IEEE Transactions on Circuits and Systems for Video Technology(ICSVT)*, 2024.

[10] Haochen Wang, Cilin Yan, Shuai Wang, Xiaolong Jiang, Xu Tang, Yao Hu, Weidi Xie, and Efstratios Gavves. Towards open-vocabulary video instance segmentation. In *proceedings of the IEEE/CVF international conference on computer vision(ICCV)*, 2023.

[11] Hao Fang, Peng Wu, Yawei Li, Xinxin Zhang, and Xiankai Lu. Unified embedding alignment for open-vocabulary video instance segmentation. In *European Conference on Computer Vision(ECCV)*, 2024.

[12] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition(CVPR)*, 2022.

[13] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition(CVPR)*, 2022.

[14] Monika Wysoczańska, Oriane Siméoni, Michaël Ramamonjisoa, Andrei Bursuc, Tomasz Trzciński, and Patrick Pérez. Clip-dinoiser: Teaching clip a few dino tricks for open-vocabulary semantic segmentation. In *European Conference on Computer Vision(ECCV)*, 2024.

[15] Junjie Wang, Bin Chen, Yulin Li, Bin Kang, Yichi Chen, and Zhuotao Tian. DeCLIP: Decoupled Learning for Open-Vocabulary Dense Perception. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2025.

[16] 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(ICCV)*, 2023.

[17] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[18] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision(ICCV)*, 2021.

[19] 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. *arXiv preprint arXiv:2304.07193*, 2023.

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[21] Haochen Wang, Cilin Yan, Shuai Wang, Xiaolong Jiang, Xu Tang, Yao Hu, Weidi Xie, and Efstratios Gavves. Towards open-vocabulary video instance segmentation. In *proceedings of the IEEE/CVF international conference on computer vision(ICCV)*, 2023.

[22] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip HS Torr, and Song Bai. Occluded video instance segmentation: A benchmark. *International Journal of Computer Vision(IJCV)*, 2022.

[23] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision(ICCV)*, 2019.

[24] 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(ECCV)*, 2022.

[25] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *IEEE international conference on image processin(ICIP)*, 2016.

[26] Yang Liu, Idil Esen Zulfikar, Jonathon Luiten, Achal Dave, Deva Ramanan, Bastian Leibe, Aljoša Ošep, and Laura Leal-Taixé. Opening up open world tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition(CVPR)*, 2022.

[27] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European conference on computer vision(ECCV)*, 2022.

[28] Junjie Wang, Keyu Chen, Yulin Li, Bin Chen, Hengshuang Zhao, Xiaojuan Qi, and Zhuotao Tian. Generalized decoupled learning for enhancing open-vocabulary dense perception. *arXiv preprint arXiv:2508.11256*, 2025.

[29] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition(CVPR)*, 2019.

[30] 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.

[31] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition(CVPR)*, 2017.

[32] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition(CVPR)*, 2014.

[33] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision(IJCV)*, 2010.

[34] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition(CVPR)*, 2022.

[35] 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 *European conference on computer vision(ECCV)*, 2022.

[36] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition(CVPR)*, 2023.

[37] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition(CVPR)*, 2023.

[38] Siyu Jiao, Yunchao Wei, Yaowei Wang, Yao Zhao, and Humphrey Shi. Learning mask-aware clip representations for zero-shot segmentation. *Advances in Neural Information Processing Systems(NeurIPS)*, 2023.

[39] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *Advances in Neural Information Processing Systems(NeurIPS)*, 2023.

[40] Xi Chen, Haosen Yang, Sheng Jin, Xiatian Zhu, and Hongxun Yao. Frozenseg: Harmonizing frozen foundation models for open-vocabulary segmentation. *arXiv preprint arXiv:2409.03525*, 2024.

[41] 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 *European conference on computer vision(ECCV)*, 2014.

[42] 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(CVPR)*, 2019.

[43] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In *European conference on computer vision(ECCV)*, 2022.

[44] Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition(CVPR)*, 2023.

[45] Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition(CVPR)*, 2023.

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

[47] Dahun Kim, Anelia Angelova, and Weicheng Kuo. Region-aware pretraining for open-vocabulary object detection with vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition(CVPR)*, 2023.

[48] Dahun Kim, Anelia Angelova, and Weicheng Kuo. Contrastive feature masking open-vocabulary vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV)*, 2023.

[49] 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(ECCV)*, 2022.

[50] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

[51] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[52] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[53] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *European Conference on Computer Vision(ECCV)*, 2016.

[54] Shuyang Sun, Runjia Li, Philip Torr, Xiuye Gu, and Siyang Li. CLIP as RNN: Segment countless visual concepts without training endeavor. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition(CVPR)*, 2024.

[55] Jiho Choi, Seonho Lee, Seungho Lee, Minhyun Lee, and Hyunjung Shim. Understanding multi-granularity for open-vocabulary part segmentation. In *Advances in Neural Information Processing Systems(NeurIPS)*, 2024.

[56] Meng Wei, Xiaoyu Yue, Wenwei Zhang, Shu Kong, Xihui Liu, and Jiangmiao Pang. OV-PARTS: Towards open-vocabulary part segmentation. In *Advances in Neural Information Processing Systems(NeurIPS)*, 2023.

[57] Claudia Cuttano, Gabriele Trivigno, Christoph Reich, Daniel Cremers, Carlo Masone, and Stefan Roth. INSID3: Training-free in-context segmentation with DINOv3. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition(CVPR)*, 2026.

[58] Yajie Liu, Jinjin Zhang, Qingjie Liu, and Di Huang. Reconstructing CLIP for open-vocabulary dense perception. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition(CVPR)*, 2026.

[59] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. *arXiv preprint arXiv:2309.16671*, 2023.

[60] Naomi Kombol, Ivan Martinović, Siniša Šegvić, and Giorgos Tolias. SPAR: Single-pass any-resolution ViT for open-vocabulary segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition(CVPR)*, 2026.

[61] Nikolay Kormushev, Josip Šarić, and Matej Kristan. Mitigating objectness bias and region-to-text misalignment for open-vocabulary panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition(CVPR)*, 2026.

[62] Jiahao Li, Yang Lu, Yachao Zhang, Fangyong Wang, Yuan Xie, and Yanyun Qu. Direct segmentation without logits optimization for training-free open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition(CVPR)*, 2026.

[63] Dengke Zhang, Fagui Liu, and Quan Tang. CorrCLIP: Reconstructing patch correlations in CLIP for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV)*, 2025.

A Further Analysis of the Core Observations

This section analyzes two observations that motivate SSCL: context-induced drift in dense local features and noisy concept-grounded text-patch responses. CISC mainly improves the stability of local features across changing context, while SGSR mainly improves the spatial coherence of text-conditioned dense supervision. The discussion below explains why these two objectives address different residual failure modes of dense CLIP adaptation.

A.1 Context-Induced Drift and Under-Constrained Dense Alignment

The motivating observation behind SSCL is that region-level self-distillation, while already strong, still leaves two crucial properties under-constrained. First, patch features inside a region can remain sensitive to surrounding context because ViT tokens interact globally through self-attention. As a result, the same semantic region may drift when its background, framing, or neighboring objects change. Second, the induced text-patch responses can remain spatially noisy because image-text pretraining does not directly supervise concept-conditioned localization at the patch level. The two auxiliary objectives introduced in the main paper are therefore designed to target these leftover degrees of freedom rather than to replace the base distillation mechanism.

A.2 Interpreting Context-Invariant Consistency

CISC uses the same frozen teacher reference for all pseudo-temporal views of an anchor. For a student token $\bar{\mathbf{x}}_r^{(k)}(p)$, the positive is the teacher token $\bar{\mathbf{x}}_r^{T}(p)$ at the corresponding anchor-grid location. The remaining teacher locations $q \neq p$ are spatial competitors in the InfoNCE denominator. All tokens are ℓ_2-normalized, as in Eq. (5). Thus, CISC encourages each view to identify the same teacher location relative to a shared set of alternatives.

This shared-reference construction explains the intended consistency effect. A student token must remain predictive of its anchor-local teacher target when the surrounding crop changes. The objective can therefore discourage view-dependent changes that interfere with spatial matching. It does not directly minimize a distance between student views, identify separate content and nuisance components, or guarantee that all contextual information is removed.

The spatial competitors need not be semantically unrelated. Different locations within an anchor can belong to the same object or have similar teacher features, making their correspondence ambiguous. The contrastive term encourages spatial discrimination to the extent supported by the teacher representation, while base region-level self-distillation continues to supervise pooled semantics. The video evaluations test transfer from static-image views to temporal correspondence. They do not imply that CISC explicitly models motion or temporal dynamics.

Scope of Context Invariance. CISC targets view-dependent feature drift for fixed local content. It does not explicitly remove contextual or positional information from the encoder. The consistency term is optimized jointly with region-level self-distillation and SGSR. Strong invariance could nevertheless be detrimental when a prediction depends on relative position or relationships between instances. Section D.9 evaluates the complete method on referring segmentation, where expressions can require such cues. This task-level comparison does not isolate CISC's effect on relational or positional information.

A.3 SGSR as Structure-Guided Response Refinement

For category c, let $R_c(p)$ be the cosine similarity between the student feature at spatial location p and the frozen text embedding. The response is computed directly from these embeddings, without an additional prediction head. We apply a spatial softmax independently for each category to obtain $\boldsymbol{\alpha}_c$, and use the frozen VFM affinity matrix $\mathbf{S}$ to construct the target

$$\widetilde{\mathbf{R}}_c = \mathbf{S}\boldsymbol{\alpha}_c. \tag{11}$$

The CLIP response supplies category-dependent semantic evidence, while $\mathbf{S}$ supplies category-agnostic relationships between patches. The refined response is detached before it supervises the original student response, as specified in the main method.

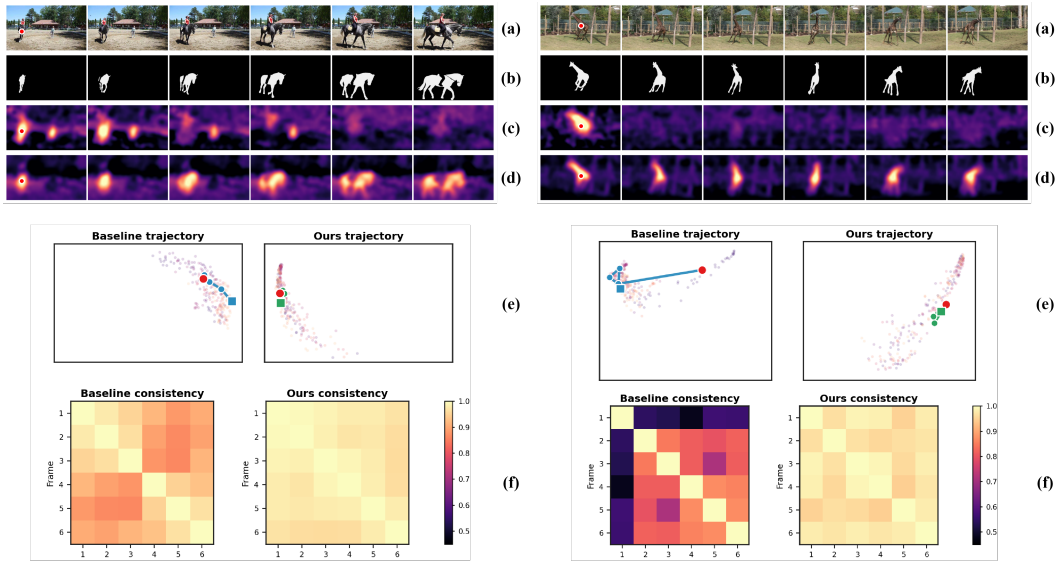

Figure 4: **Visualization of context-induced semantic drift and temporal prototype consistency.** Given an anchor point initialized on the target object in the first frame, we compute its cosine-similarity map with dense visual features in subsequent frames. **(a)** Video frames. **(b)** Ground-truth instance masks. **(c)** Baseline anchor-point similarity maps. **(d)** Anchor-point similarity maps produced by the adapted representation shown in the figure. Although the target remains visible across frames, the baseline responses become diffuse and drift toward irrelevant regions or incomplete object parts under changes in pose, scale, occlusion, and surrounding context. In contrast, our representation maintains more concentrated responses on the target region, indicating better local semantic consistency. We further compute one object prototype per frame by averaging dense features within the ground-truth mask. **(e)** The projected prototype trajectories show that baseline prototypes exhibit larger temporal drift, whereas ours remain more compact across frames. **(f)** Pairwise prototype cosine-similarity matrices summarize within-object feature stability across the displayed frames.

The response weights depend on the queried category, while the frozen VFM supplies a common affinity matrix across categories. Section D.3 evaluates the effect of normalizing the responses before aggregation.

When the VFM affinities connect patches belonging to the same object or surface, the aggregation can spread semantic evidence from discriminative parts to other coherent regions. It can also propagate errors when the affinities connect unrelated regions, so it does not guarantee correct boundaries or semantic labels. Figure 5 illustrates the observed refinement of raw CLIP responses, and Section D.7 evaluates the dependence on the chosen structural encoder.

A.4 Why the Two Objectives Are Complementary

The two losses operate in different spaces, motivating the complementarity evaluated in Table 5. CISC acts in the visual feature space and makes local features less sensitive to incidental context variation. SGSR acts in the semantic response space and makes the text-induced dense target more coherent and region-aware. Intuitively, CISC addresses feature variation across contexts, while SGSR addresses spatially fragmented text-patch responses.

Table 5 supports this complementarity: each objective improves the base self-distillation model, and their combination gives the best result on every reported metric. Section D.1 discusses the corresponding gains, while Section D.4 examines the objectives jointly through their loss weights.

B Additional Qualitative Analysis

We examine the qualitative behavior of the adapted representation using the correspondence and response-map visualizations in Figures 4 and 5, and discuss its limitations.

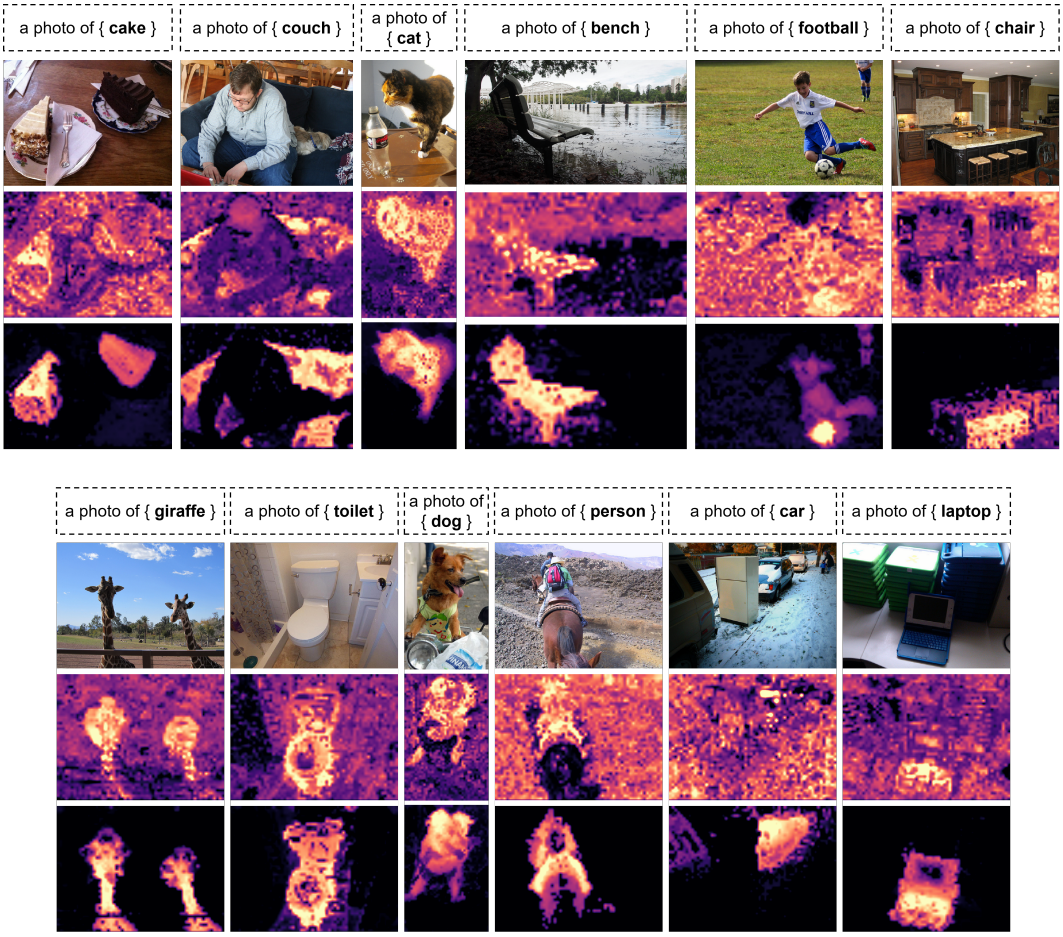

Figure 5: **Additional visualization of structure-guided response refinement.** For each queried concept, we visualize the input image, the raw CLIP text-patch activation, and the SGSR-refined activation. These examples complement the main-paper visualization by showing that the refinement behavior is consistent across categories with different scales, shapes, and contextual clutter. SGSR reduces spurious background responses and fills in missing object regions by propagating text-conditioned evidence according to visual affinities.

B.1 Temporal Consistency under Context Changes

Figure 4 examines whether an anchor feature remains associated with the same object across video frames. In the illustrated sequences, baseline similarity responses drift toward nearby regions as appearance and context change. The adapted representation maintains more concentrated responses on the target. These examples involve joint changes in pose, scale, occlusion, and surrounding context, so they do not isolate the effect of background variation.

For each frame, we average dense features inside the ground-truth instance mask to obtain an object prototype. The trajectory panels visualize these prototypes in two dimensions, while the cosine-similarity matrices compare them in the original feature space. The displayed adapted features have higher within-object temporal similarity than the baseline. This diagnostic measures stability for the tracked object, not discrimination between different objects, and the projected trajectories are qualitative illustrations.

Why Pseudo-Temporal Modeling. Training uses static images to create views with known geometric correspondence, avoiding additional video annotations or tracking-based supervision. The views are independent spatial augmentations and carry no temporal ordering. They vary framing,

scale, and surrounding context, but do not explicitly model object motion, deformation, or occlusion. The video analysis evaluates transfer of this learned representation.

B.2 Structure-Guided Semantic Refinement Maps

Figure 5 illustrates the motivation for SGSR. The raw CLIP text response is often semantically meaningful but spatially incomplete: it may fire strongly on the most discriminative object part, fragment across disconnected patches, or leak into visually related background regions. This behavior is expected because image-text pretraining optimizes image-level alignment rather than dense, concept-conditioned localization. As a result, raw text-patch responses are better viewed as noisy semantic seeds rather than full dense supervision.

In the examples in Figure 5, refinement concentrates the response on the queried objects and suppresses scattered background activations. The extent of the improvement varies with the reliability of the initial semantic response and the VFM affinities.

The reliability of refinement remains limited by the quality of the structural prior. When the VFM affinity itself connects the wrong structures—for example, when foreground and background are strongly correlated in texture or when multiple instances are nearly indistinguishable—the refinement can still propagate errors. In that sense, SGSR should be viewed as a mechanism for making the semantic signal more spatially coherent, not as a replacement for the semantic signal itself. Its best-case behavior occurs when CLIP already identifies the right concept cue and the VFM provides a reliable local geometry for spreading that cue.

B.3 Limitations and Potential Failure Modes

SSCL targets context-induced feature drift and spatially fragmented text-patch responses, but remains subject to limitations of the visual evidence and downstream model. Small or heavily occluded objects can provide insufficient information for stable instance association. Weak semantic responses and inaccurate VFM affinities can also lead to incomplete regions or boundary errors. Detection performance additionally depends on the proposals produced by the downstream detector. These potential failure modes motivate richer temporal and instance-aware supervision.

C Details of Experimental Settings

C.1 Algorithmic Summary of SSCL

Figure 3 summarizes the three training components: base region-level self-distillation in panel (b), CISC in panel (a), and SGSR in panel (c). The base objective is described in Section 2.1. The adaptation procedure is summarized below.

1. Sample a training image and construct candidate crop regions from a random $n \times n$ grid partition.
2. Select anchor regions from this candidate set and form one teacher reference crop together with multiple larger student crops that preserve the anchor inside each view.
3. Extract dense student features for every view, extract the frozen teacher feature for the reference crop, and project all anchor supports to a common patch grid.
4. Apply the base self-distillation loss between pooled student and teacher region representations, and apply CISC as a point-to-point dense correspondence loss across the projected anchor grids.
5. Construct CLIP text responses for the valid image-level categories, refine them with the frozen VFM affinity matrix, and use the refined responses as dense semantic targets for SGSR.
6. Update only the student image encoder. Keep the CLIP teacher, text encoder, and VFM frozen throughout training, and discard the auxiliary branches after adaptation.

All three objectives optimize the same student image encoder. The teacher, text encoder, and VFM remain frozen and are used only during adaptation.

C.2 Benchmark Summary

All experiments in the main paper follow the same overall transfer recipe: we first adapt a dense CLIP backbone on COCO2017, and then plug the adapted encoder into task-specific downstream frameworks without changing the benchmark definitions themselves. The four evaluation families are open-vocabulary video instance segmentation, region classification, semantic segmentation, and object detection. Their differences lie in the downstream inference head, evaluation dataset, and reported metric, while the central object of comparison remains the quality of the adapted dense visual representation.

C.3 Task-Specific Evaluation Protocols

OV Video Instance Segmentation. For video instance segmentation, we adopt CLIP-VIS [9] as the downstream framework and replace its visual backbone with the adapted dense encoder. We report the official AP metric on LV-VIS, OVIS, and YouTube-VIS 2019/2021, and additionally report novel-category AP_N on LV-VIS whenever available. These benchmarks stress different aspects of transfer: LV-VIS is the most direct test of long-tail open-vocabulary video understanding, OVIS emphasizes heavy occlusion and crowded scenes, and YouTube-VIS 2019/2021 measures robustness on the standard video instance segmentation protocol. In all cases, we keep the CLIP-VIS video head, matching strategy, and inference pipeline unchanged, so the comparison isolates the effect of the adapted visual encoder.

OV Region Classification. We evaluate the adapted dense encoder on COCO-Panoptic [29] validation by pooling features within annotated regions and classifying them via similarity to text embeddings. We report Acc@1 and Acc@5 for boxes, thing masks, and stuff masks, following the same region classification protocol as in the main paper.

OV Semantic Segmentation. For semantic segmentation, we evaluate the adapted backbone with both CAT-Seg [4] and H-CLIP [5], replacing only the image encoder and preserving the original training and inference protocols of the segmentation head. The source training set remains the one used by the corresponding baseline—in our case, COCO-Stuff for the reported CAT-Seg and H-CLIP settings—while evaluation is performed on ADE20K, PASCAL-Context, and VOC. We report mIoU on both the expanded and standard label spaces, namely ADE847/ADE150, Context459/Context59, and VOC21/VOC20 when available. This setup evaluates whether improvements in dense feature quality consistently transfer across two segmentation frameworks with different decoders while leaving their prompt construction, decoding logic, and inference settings unchanged.

OV Object Detection. For object detection, we adopt F-ViT as the default transfer baseline and follow the standard OV-COCO and OV-LVIS evaluation protocols. OV-COCO uses the standard base/novel split with 48 base and 17 novel categories and reports AP_{50}^{novel}, while OV-LVIS evaluates rare-category AP under the usual long-tailed LVIS protocol. This benchmark tests whether the adapted dense representation can benefit proposal-based open-vocabulary detection under realistic localization noise and long-tailed category distributions. We keep the detector head, proposal pathway, and task-level optimization schedule from the original F-ViT setting unchanged and substitute only the visual backbone.

Additional Transfer Tasks. The additional evaluations distinguish three transfer settings. Referring segmentation uses a fixed CaR pipeline on RefCOCO validation without downstream training. Part segmentation fine-tunes two matched CAT-Seg checkpoints on Pascal-Part-116. In-context segmentation uses training-free INSID3 prototype matching on COCO-20i and LVIS-92i. The paired comparisons and metrics are specified in Sections D.9–D.11.

Licenses for Existing Assets. We use standard public datasets, pretrained models, and downstream frameworks following their official licenses and terms of use. We credit the original creators of all datasets, models, and codebases used in our experiments through the corresponding citations. We do not redistribute raw datasets and require users to obtain them from their official sources.

C.4 Implementation Details

Model Configuration. In the main experiments, we instantiate both the student and teacher from EVA-CLIP [30], and experiment with two backbone scales: ViT-B/16 and ViT-L/14. The student is initialized from the same EVA-CLIP checkpoint as the frozen teacher, while the EVA-CLIP text encoder is kept frozen throughout training. For structure-guided refinement, we use DINOv2 [19] as the visual foundation model, pairing EVA-CLIP ViT-B/16 with DINOv2-B and EVA-CLIP ViT-L/14 with DINOv2-L. To align the VFM token grid with the CLIP patch grid, we set the input resolutions so that both branches produce the same 64×64 spatial patch-token grid. Specifically, for the ViT-B setting, EVA-CLIP and DINOv2 take 1024×1024 and 896×896 inputs, respectively. For the ViT-L setting, both branches take 896×896 inputs.

Region Sampling and Pseudo-Temporal Views. After resizing each training image, we uniformly sample n from $\{2, 3, 4, 5, 6\}$ and partition the patch lattice into an $n \times n$ grid of non-overlapping candidate cells. We uniformly select four distinct cells without replacement as anchors, without using object boxes, masks, saliency maps, or category labels for this selection. The teacher anchors are fixed before the larger student context crops are sampled. For each anchor, we construct $K = 6$ patch-aligned student views. The teacher reference uses the tight anchor crop. For an anchor with side length l_r, the six student views use square crop windows with scale multipliers uniformly sampled from $[1.2, 3.0]$ relative to l_r, and the crop center is randomly jittered by up to $\pm 0.25 l_r$ along each spatial axis while keeping the anchor fully inside the crop. All crop boundaries are quantized to integer multiples of the CLIP patch size on the resized input image before resizing, and the projected anchor region in each student view is bilinearly resized on the feature map to the teacher reference grid.

Training Setup. Unless otherwise specified, downstream evaluations start from the same adapted backbone and follow the task-specific training and inference protocols described above. Training is performed on COCO2017 for 6 epochs using AdamW with an initial learning rate of 1×10^{-5}, weight decay 0.1, and a per-GPU batch size of 4 across 4 RTX PRO 6000 GPUs, yielding a total batch size of 16. It takes approximately 1 hour per epoch for ViT-B/16 and 1.5 hours per epoch for ViT-L/14. Due to architecture- and dependency-specific compatibility requirements of the downstream frameworks, downstream training and evaluation are conducted on NVIDIA RTX A6000 GPUs. We denote the main-paper coefficients λ_1 and λ_2 by λ_{CISC} and λ_{SGSR}, respectively. For Context-Invariant Spatiotemporal Consistency, we use temperature $\tau_c = 0.2$ in the point-to-point InfoNCE loss, and set $\lambda_{\mathrm{CISC}} = 0.25$ and $\lambda_{\mathrm{SGSR}} = 1$ for the two auxiliary objectives.

C.5 Practical Overhead Discussion

SSCL incurs additional computation during dense adaptation through the multi-view student pathway, frozen crop teacher, and VFM structural branch. Only the student image encoder is optimized. The teacher, text encoder, and VFM provide fixed supervision and require neither gradients nor optimizer states.

Table 6 reports the cost of six epochs of adaptation on four RTX PRO 6000 GPUs. ViT-B/16 takes approximately 6 hours and ViT-L/14 approximately 9 hours, corresponding to 24 and 36 GPU-hours, respectively.

Table 6: **Cost of SSCL adaptation.** Wall-clock time and total GPU-hours for six epochs on four RTX PRO 6000 GPUs.

Backbone	Time (hours)	GPU-hours
ViT-B/16	6	24
ViT-L/14	9	36

After adaptation, the auxiliary branches are discarded and only the student image encoder is retained. Replacing the baseline encoder with the adapted encoder therefore preserves the downstream architecture and its inference pathway. The cost reported here is the complete adaptation cost, rather than a measured overhead relative to a separately timed self-distillation baseline. The same adapted checkpoint can be reused across downstream tasks.

C.6 Reproducibility and Transfer Notes

Text prompts and semantic targets. During pre-adaptation, SGSR uses only the valid image-level categories associated with the current training image. For each valid category, we instantiate the fixed template `a photo of {category}` and compute the corresponding CLIP text embedding. The text-patch response is the cosine similarity of normalized patch and text embeddings. Its softmax is taken over spatial locations independently for each category before VFM-guided aggregation. Restricting the category set avoids evaluating the entire vocabulary at every iteration. Because the refined response is used as a stop-gradient dense target, optimization updates only the student backbone and does not backpropagate through the text encoder, the VFM branch, or the refined target construction itself.

Token alignment and valid spatial regions. A practical issue in our setting is that EVA-CLIP and DINOv2 use different patch sizes and different image preprocessing pipelines. We therefore quantize crop boundaries to the CLIP patch grid before resizing, maintain separate normalization for the CLIP and VFM branches, and align both outputs to the same 64×64 token lattice before computing affinities or correspondences. For CISC, correspondence supervision is restricted to valid projected anchor regions rather than padded or extrapolated locations. SGSR instead refines the text-patch response over the valid full-image patch grid. This detail is important when the student view uses a much larger crop window than the teacher anchor and helps prevent artificial boundary artifacts.

Transfer principle across downstream tasks. For each paired comparison, benchmark definitions, task-specific heads, evaluation code, and post-processing follow the same downstream implementation.

D Additional Experiments

This section complements the main results with design ablations, individual and joint hyperparameter analyses, alternative VFM and VLM backbones, and additional transfer tasks. We also compare SSCL with DenseRC on region classification and CAT-Seg segmentation.

D.1 Interpreting the Main Ablation Table

The complementarity ablation in the main paper already reveals a clear task-dependent pattern. Relative to base self-distillation, adding CISC alone improves box, thing, and stuff Acc@1 by +2.9, +6.6, and +10.2, respectively, and improves segmentation mIoU by +1.8 on ADE150 and +2.4 on Context59. Adding SGSR alone yields smaller but still consistent gains of +1.4, +1.6, and +3.3 on region classification, together with +0.5 and +1.4 on the two segmentation benchmarks. When both losses are enabled, the full model reaches gains of +4.0, +7.5, and +10.3 over the base on region classification and +3.4 and +3.6 on semantic segmentation.

CISC gives larger standalone gains than SGSR on all five reported metrics. Combining the objectives gives the highest score on each metric, consistent with their complementary roles in local feature consistency and response refinement.

D.2 Ablation on CISC Design

Queries, Positives, and Negatives. For each student query $\bar{\mathbf{x}}_r^{(i)}(p)$, the teacher token $\bar{\mathbf{x}}_r^T(p)$ at the same aligned anchor-grid location is the sole positive. All other teacher tokens $\{\bar{\mathbf{x}}_r^T(q) : q \in \mathcal{G}_r,\ q \neq p\}$ are negatives. Thus, the positive-to-negative ratio per query is $1 : (|\mathcal{G}_r| - 1)$. The $K = 6$ pseudo-temporal views supply six independent query grids. They are neither six positives in one denominator nor negatives for one another. InfoNCE compares the matched teacher location with the competing spatial locations, encouraging spatial discriminability when the teacher features distinguish those locations.

Loss Formulation. We next analyze the loss formulation used in CISC under the same setup as Table 5. The comparison includes five variants: (i) removing CISC entirely, (ii) replacing the

point-to-point InfoNCE loss with an L_1 loss, (iii) replacing it with an L_2 loss, (iv) replacing it with a cosine-similarity loss, and (v) using the default InfoNCE formulation. This study is meant to test whether the gain comes from the proposed contrastive correspondence objective rather than from simpler pairwise feature-matching losses.

Table 7: Loss-form ablation for Context-Invariant Spatiotemporal Consistency, measured on COCO-Panoptic region classification and H-CLIP semantic segmentation.

	Region Classification (Acc@1)			Semantic Segmentation (mIoU)	
Variant	Boxes	Thing	Stuff	ADE150	Context59
w/o CISC	74.5	72.4	50.9	35.0	60.4
L_1 loss	75.7	75.1	54.3	36.3	61.3
L_2 loss	76.1	76.0	55.4	36.7	61.7
Cosine loss	76.6	77.1	56.7	37.3	62.1
InfoNCE loss	**77.1**	**78.3**	**57.9**	**37.9**	**62.6**

Temperature Sensitivity. We additionally vary the InfoNCE temperature while retaining the same training and evaluation protocol. Table 8 shows that the tested temperatures produce ranges of 1.1, 0.2, and 0.7 Acc@1 points for Boxes, Thing, and Stuff, respectively. Performance is relatively stable over this range, with the default $\tau_c = 0.2$ giving the highest Stuff accuracy.

Table 8: **CISC temperature sensitivity.** We report COCO-Panoptic region classification Acc@1 under the same protocol.

Temperature τ_c	Boxes	Thing	Stuff
0.1	76.1	**78.5**	57.2
0.2 (default)	77.1	78.3	**57.9**
0.4	**77.2**	**78.5**	57.7

D.3 Ablation on SGSR Design

We analyze the SGSR formulation while preserving the DINOv2-based structural guidance. The comparison includes three settings: (i) removing SGSR entirely, (ii) removing the response-normalization step before DINO-guided aggregation, and (iii) using the default SGSR formulation with response normalization. This comparison measures the downstream effect of including response normalization in the evaluated SGSR formulation.

Table 9: Ablation on SGSR design, comparing removal of SGSR and the effect of response normalization before DINO-guided aggregation, measured on COCO-Panoptic region classification and H-CLIP semantic segmentation.

	Region Classification (Acc@1)			Semantic Segmentation (mIoU)	
Variant	Boxes	Thing	Stuff	ADE150	Context59
w/o SGSR	76.0	77.4	57.8	36.3	61.4
w/o response normalization	76.6	77.8	57.8	37.1	62.0
with response normalization	**77.1**	**78.3**	**57.9**	**37.9**	**62.6**

D.4 Sensitivity to Loss Weights

We further study the sensitivity to λ_{CISC} and λ_{SGSR} using COCO-Panoptic region classification. We sweep λ_{CISC} in $\{0.1, 0.25, 0.5\}$ while fixing $\lambda_{\text{SGSR}} = 1.0$, and sweep λ_{SGSR} in $\{0.5, 1.0, 2.0\}$ while fixing $\lambda_{\text{CISC}} = 0.25$. The default setting used throughout the main paper is $\lambda_{\text{CISC}} = 0.25$ and $\lambda_{\text{SGSR}} = 1.0$.

Table 10: Sensitivity to the loss weights λ_{CISC} and λ_{SGSR}. We report COCO-Panoptic region classification using Boxes, Thing, and Stuff Acc@1.

(a) λ_{CISC} sensitivity

Value	Boxes	Thing	Stuff
0.1	76.2	76.9	55.9
0.25	**77.1**	**78.3**	**57.9**
0.5	76.8	77.9	57.2

(b) λ_{SGSR} sensitivity

Value	Boxes	Thing	Stuff
0.5	76.6	77.7	56.9
1.0	**77.1**	**78.3**	**57.9**
2.0	76.8	77.8	57.1

Joint Coefficient Analysis. To examine the interaction between the objectives, we also evaluate the four combinations of the low and high coefficients in a 2×2 factorial comparison. Table 11 complements the individual sweeps in Table 10. Changing λ_{CISC} has a small effect at either tested value of λ_{SGSR}. Increasing λ_{SGSR} from 0.5 to 2.0 lowers Stuff accuracy at both CISC weights, indicating that stronger structural regularization is not uniformly beneficial. Across the four joint settings, Boxes and Thing accuracies span 0.3 and 0.5 points, while Stuff spans 2.8 points. The default setting provides the highest Boxes and Thing accuracy among the listed configurations and remains close to the highest Stuff accuracy.

Table 11: **Joint sensitivity to the CISC and SGSR weights.** Each entry reports Boxes / Thing / Stuff Acc@1. The default $(0.25, 1.0)$ gives $77.1/78.3/57.9$.

	$\lambda_{\text{SGSR}} = 0.5$	$\lambda_{\text{SGSR}} = 2.0$
$\lambda_{\text{CISC}} = 0.1$	76.7 / 77.5 / 57.4	77.0 / 77.4 / 55.8
$\lambda_{\text{CISC}} = 0.5$	76.7 / 77.9 / 58.1	76.7 / 77.5 / 55.3

D.5 Interpreting the Sampling Design Choices

Several sampling-related design choices in SSCL are best understood as controlling the hardness of the pseudo-temporal supervision rather than as introducing new semantic information. These hyperparameters primarily affect the CISC branch: the random grid size $n \in \{2, \ldots, 6\}$ determines the granularity of candidate regions, the number of anchors per image controls how many local correspondences are enforced in each update, and the number of student views K controls how broadly each anchor is exposed to context variation. Together, they determine whether CISC sees mostly easy crops with near-identical context or genuinely challenging crops that require local semantic stability under substantial changes in framing and surrounding content.

The intended trade-off is straightforward. If the crop distribution is too conservative, the model receives little pressure to suppress context-induced drift because the teacher and student views remain nearly identical. If it is too aggressive, the projected anchor can become too small relative to the crop window, or too weakly constrained spatially, which makes dense correspondence noisier and can reduce optimization stability. The default choices in the paper therefore aim for a middle ground: the crops are large enough to vary context meaningfully, but still constrained enough that the anchor identity remains recoverable after projection to the common patch grid.

SGSR should be interpreted differently. It operates on full-image dense response maps induced by the current image-level labels, so it is not governed by the teacher-student crop geometry in the same direct way as CISC. Its main trade-off is instead between semantic faithfulness and structural smoothing: stronger structure-guided propagation can make responses more region-complete and less noisy, but it can also over-smooth boundaries or spread errors when the VFM affinity graph connects the wrong locations. The connection to sampling is therefore indirect—both branches still optimize the same student backbone on the same training images—but the crop hyperparameters are primarily control knobs for CISC rather than SGSR.

D.6 Cross-Task Improvement Patterns

Table 12 summarizes gains over DeCLIP at each backbone scale across several downstream tasks. The gains are positive for all listed settings.

Table 12: **Representative gains over DeCLIP on selected benchmarks.** We report the improvement of our adapted backbone over the corresponding DeCLIP-based transfer baseline at the same backbone scale.

Benchmark	Metric	ViT-B/16	ViT-L/14
OVIS	AP	+2.0	+1.7
LV-VIS test	AP_N	+1.0	+2.2
COCO-Panoptic stuff	Acc@1	+5.2	+3.7
OV-COCO novel	AP_{50}	+1.6	+0.9
OV-LVIS rare	AP_r	+1.7	+1.4
CAT-Seg / ADE150	mIoU	+1.3	+1.2
CAT-Seg / Context59	mIoU	+1.5	+0.6

Table 12 makes the pattern concrete. On COCO-Panoptic stuff masks, the gain over DeCLIP reaches +5.2 Acc@1 for ViT-B/16 and +3.7 for ViT-L/14, which is substantially larger than the corresponding box-level gain. This pattern is consistent with improved dense representation quality on diffuse or weakly bounded regions. Since the comparison evaluates the complete SSCL model, it does not isolate the contribution of either objective. A similar effect appears on OVIS, where clutter, occlusion, and deformation place strong demands on context-robust local features.

The gains vary with backbone scale and task. ViT-B/16 shows larger improvements on OVIS and COCO-Panoptic stuff classification, whereas ViT-L/14 shows a larger gain on LV-VIS test novel-category AP_N. Both scales benefit across all listed benchmarks.

D.7 Generalization Across Frozen Visual Foundation Models

We examine whether SGSR depends on the particular VFM used to construct its spatial affinities. We replace DINOv2 [19] with DINOv3 [50] while keeping the other training settings unchanged, and compare both variants with SSCL without SGSR. The VFM remains frozen and supplies category-agnostic structural guidance, while semantic supervision comes from the VLM's text-patch responses. All variants use the same evaluation protocol without VFM-specific tuning.

Table 13 reports region classification and zero-shot semantic segmentation results. Both VFMs improve all three metrics over the variant without SGSR. DINOv2 performs slightly better on thing classification and VOC20, while DINOv3 performs better on COCO-Stuff. These results indicate that SGSR can benefit from either structural prior, with task-dependent differences rather than a uniform advantage for one VFM.

Table 13: **Generalization across frozen VFMs.** We vary the auxiliary VFM while keeping the other training settings fixed. Thing classification is measured by Acc@1, and zero-shot segmentation is measured by mIoU.

VFM	Thing Acc@1	VOC20 mIoU	COCO-Stuff mIoU
w/o SGSR	77.4	84.5	24.6
DINOv2	**78.3**	**85.6**	25.1
DINOv3	78.0	85.2	**25.7**

D.8 Generalization to SigLIP 2

The main experiments use EVA-CLIP at two model scales. To evaluate transfer to a different VLM family, we additionally apply SSCL to SigLIP 2-B/16 [51] and measure region classification performance. This experiment changes the target VLM being adapted, unlike the auxiliary VFM replacement in Section D.7.

As shown in Table 14, SSCL improves Boxes, Thing, and Stuff Acc@1 by 1.6, 1.8, and 1.9 points, respectively. These consistent improvements complement the EVA-CLIP results and support the applicability of the adaptation framework to both evaluated VLM families.

Table 14: **Region classification with SigLIP 2-B/16.** We compare the SigLIP 2 baseline with its SSCL-adapted counterpart and report Acc@1.

Model	Boxes	Thing	Stuff
SigLIP 2-B/16	71.6	69.9	49.8
SigLIP 2-B/16 + SSCL	**73.2**	**71.7**	**51.7**

D.9 Transfer to Referring Image Segmentation

Referring image segmentation identifies a target from a natural-language expression that can contain directional or relational cues. It therefore provides a useful task-level check on a representation trained for context consistency. RefCOCO was collected using ReferItGame [52] and permits expressions that describe a target by its position [53]. We evaluate SSCL on the validation split.

We compare the pretrained and SSCL-adapted ViT-B/16 encoders using the same fixed, training-free CaR inference pipeline [54]. Table 15 shows improvements in both mIoU and oIoU. The result supports transfer to referring segmentation under this pipeline. It does not isolate the contribution of CISC or establish performance on a subset of expressions requiring relational reasoning.

Table 15: **Referring image segmentation on RefCOCO validation.** Both encoders use the same fixed CaR inference pipeline without additional downstream training.

Encoder	mIoU	oIoU
Pretrained ViT-B/16	29.9	26.8
ViT-B/16 + SSCL	**33.3**	**29.9**

D.10 Transfer to Open-Vocabulary Part Segmentation

Part segmentation tests whether the adapted representation transfers to concepts at a finer granularity than whole objects. Choi et al. [55] evaluate CAT-Seg [4] as a fine-tuned baseline for part segmentation. We use the same downstream framework to compare the two initializations. Starting from the two CAT-Seg checkpoints trained under our matched open-vocabulary semantic segmentation setup, with pretrained CLIP and SSCL encoders respectively, we fine-tune both models on Pascal-Part-116 [56] using identical settings. We report mIoU on the seen and unseen splits and their harmonic mean, h-IoU.

Table 16 shows that SSCL improves both seen and unseen part segmentation, increasing h-IoU from 26.5 to 30.2. This comparison indicates that SSCL provides a stronger initialization for downstream part segmentation under the evaluated transfer setup.

Table 16: **Open-vocabulary part segmentation on Pascal-Part-116.** Both CAT-Seg models use the same downstream fine-tuning settings. Seen and unseen scores are mIoU, and h-IoU is their harmonic mean.

Model	Seen	Unseen	h-IoU
Pretrained + CAT-Seg	27.9	25.2	26.5
SSCL + CAT-Seg	**31.1**	**29.3**	**30.2**

D.11 Transfer to In-Context Semantic Segmentation

In-context semantic segmentation specifies target concepts through annotated visual support examples rather than category names. INSID3 [57] uses prototypes from annotated support examples to select and aggregate candidate regions in a query image without additional training. We adapt its encoder interface to EVA02-CLIP-B/16 and compare the pretrained and SSCL-adapted checkpoints under otherwise identical settings on COCO-20i and LVIS-92i. We also report results for DINOv2-B/14 and DINOv3-B/16 as reference encoders in this evaluation.

Table 17 shows that SSCL improves the pretrained EVA02-CLIP-B/16 checkpoint by 30.50 and 18.41 mIoU points on COCO-20i and LVIS-92i, respectively. The DINO rows provide reference scores rather than a matched ablation of the SSCL objective across encoder families. These results demonstrate improved transfer to visually specified concepts and complement the text-conditioned evaluations in the main paper.

Table 17: **In-context semantic segmentation with INSID3.** We report mIoU using training-free in-context segmentation. The two EVA02-CLIP checkpoints share the same encoder interface and downstream settings.

Encoder	COCO-20i	LVIS-92i
EVA02-CLIP-B/16	20.19	13.19
EVA02-CLIP-B/16 + SSCL	50.69	**31.60**
DINOv2-B/14	41.91	29.42
DINOv3-B/16	**52.08**	29.90

D.12 Comparison with DenseRC

Liu et al. [58] report region classification without RPN proposals in their Table 1 and CAT-Seg [4] segmentation in their Table 4. We compare SSCL with the corresponding DenseRC results at the ViT-B/16 and ViT-L/14 scales. Table 18 reports Boxes and Stuff Acc@1 together with ADE150 and Context59 mIoU at both backbone scales.

With ViT-B/16, SSCL improves Boxes and Stuff classification and Context59 segmentation, while matching DenseRC on ADE150. With ViT-L/14, SSCL improves Stuff classification and both segmentation metrics, whereas DenseRC achieves higher Boxes accuracy. The comparison therefore shows improvements on most evaluated settings while preserving the task-specific differences between the two methods.

Table 18: **Comparison with DenseRC at two backbone scales.** Region classification uses Acc@1, and CAT-Seg semantic segmentation uses mIoU. Best results are bolded separately for each backbone scale.

Method	Backbone	Region Classification		Semantic Segmentation	
		Boxes	Stuff	ADE150	Context59
DenseRC	ViT-B/16	76.7	55.8	**37.6**	61.3
SSCL	ViT-B/16	**77.1**	**57.9**	**37.6**	**62.1**
DenseRC	ViT-L/14	**83.2**	58.2	41.5	64.0
SSCL	ViT-L/14	81.2	**58.3**	**41.9**	**64.5**

E Related Work

E.1 Open-Vocabulary Dense Perception

Open-vocabulary dense perception extends vision-language representations to localization, segmentation, and video understanding. Object detection methods include ViLD [6], OV-DETR [43], BARON-KD [44], CORA [45], PromptOVD [46], RO-ViT [47], and CFM-ViT [48]. Semantic segmentation methods such as LSeg [2], OVSeg [3], CAT-Seg [4], H-CLIP [5], SAN [36], ODISE [37], MAFT [38], FC-CLIP [39], and FrozenSeg [40] combine vision-language semantics with dense decoders. CLIP-VIS [9], OV2Seg [10], and OVFormer [11] extend open-vocabulary instance segmentation to video. SSCL studies a shared backbone adaptation strategy across these task families.

E.2 Transferring Vision-Language Models to Dense Prediction

Image-text pretraining in CLIP-based models [1, 30, 59] leaves dense alignment under-constrained. DenseCLIP [12], OVSeg [3], CAT-Seg [4], and H-CLIP [5] adapt representations with task-specific

decoders. Other methods pre-adapt the encoder before downstream transfer. RegionCLIP [13] strengthens region-text correspondence through region-centric pseudo supervision. CLIPSelf [8] distills crop representations into pooled dense features, and DeCLIP [15] adds VFM structural guidance. SSCL belongs to this second family and addresses context-induced feature drift and noisy concept-grounded responses after region-level self-distillation.

Recent approaches further explore the construction and transfer of dense representations. DenseRC [58] reconstructs dense CLIP features from multiple layers. We compare it with SSCL on region classification and CAT-Seg segmentation in Section D.12. SPAR [60] targets efficient any-resolution segmentation through single-pass inference, OVRCOAT [61] addresses objectness bias and region-to-text misalignment in panoptic segmentation, and DSLO [62] derives segmentation maps without iterative logits optimization. These methods address different inference mechanisms or task settings.

E.3 Vision Foundation Models as Structural Priors

VFMs [16–19] provide strong spatial organization even without category-specific supervision from the target task. Recent dense prediction methods increasingly use these models as sources of structural priors, affinity graphs, masks, or teacher features. In SSCL, the VFM remains frozen and supplies spatial affinities for refining CLIP-induced text responses. This separates the source of semantic supervision from the structural prior.

CLIP-DINOiser and DeCLIP [14, 15] transfer visual structure to dense CLIP representations, while CorrCLIP [63] reconstructs patch correlations for segmentation. SSCL uses VFM affinities to refine text-conditioned responses during adaptation.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We accurately state the proposed method, main contributions, and evaluated experimental scope in the abstract and introduction, without extending the claims beyond the reported results.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitation that anchor-centric pseudo-temporal views only approximate contextual changes in real videos, and identify richer context modeling as future work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: Our contributions are primarily methodological and empirical, focusing on dense CLIP adaptation and evaluation across open-vocabulary dense perception tasks. We do not present formal theoretical results, so assumptions and proofs are not applicable.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide the proposed objectives, training procedure, benchmark protocols, model configurations, optimization settings, and key hyper parameters needed to reproduce the main experimental results supporting our claims.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We will make the code and models publicly available, but the current submission does not yet provide open access to the code with full reproduction instructions.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify the training and evaluation details, including model configurations, region sampling and pseudo-temporal view construction, optimizer, learning rate, batch size, training epochs, loss weights, and task-specific evaluation protocols.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We do not report error bars or statistical significance tests. The main results are based on standard benchmark evaluation protocols, and the reported improvements are compared under matched experimental settings.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the GPU types, GPU memory and per-epoch training time for the dense adaptation stage, and separately specify that downstream training and evaluation are conducted on RTX A6000 GPUs due to architecture- and dependency-specific compatibility requirements.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and conduct our study using standard public datasets and benchmark protocols, without human-subject experiments or private/sensitive data. We also preserve submission anonymity as required.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss that SSCL may help reduce dense annotation requirements for dense perception, and we do not anticipate obvious negative societal impacts from this general representation adaptation method

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We do not release scraped datasets or high-risk models such as pretrained language models or image generators. The code and models to be released are intended for dense visual representation learning on standard public benchmarks, so additional safeguards for high-risk release are not applicable.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We credit the original creators of the datasets, pretrained models, and downstream frameworks used in our experiments, and we state that these assets are used following their official licenses and terms of use.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: We do not release new datasets or other new assets with the current submission. Code and model checkpoints will be documented when they are made publicly available.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: We do not conduct crowdsourcing experiments or research with human subjects; all experiments are based on standard public benchmarks and evaluation protocols.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: We do not conduct crowdsourcing experiments or research with human subjects; therefore, participant risk disclosure and IRB approval are not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: We do not use LLMs as important, original, or non-standard components of the core methodology. LLMs were used only for grammar polishing and formatting assistance, without affecting the method design, experiments, results, or scientific claims.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.