WHEN [CLS] FALLS SHORT: OBJECT-AWARE FEATURE DISPLACEMENT FOR ROBUST GENERALIZED CATEGORY DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Generalized Category Discovery (GCD) aims to recognize known categories and discover novel ones from unlabeled data by leveraging knowledge from a partially labeled set. Most existing methods primarily rely on the global [CLS] token of Vision Transformers as the image representation. Although holistic and highly semantic, the view-invariant [CLS] feature may retain background and appearance cues that remain stable across augmentations, reducing the separability of novel categories. In this paper, we propose **O**bject-**A**ware Feature **Disp**lacement (OADisp), a plug-and-play paradigm that improves the global representation without replacing [CLS] or retraining the backbone. **OADisp** exploits the rich local semantics encoded in patch features to form complementary guidance for adapting the original [CLS] toward an image-conditioned target, thereby preserving holistic global semantics while incorporating localized discriminative cues. Specifically, we instantiate this guidance with object-aware semantic descriptors constructed through mask-guided aggregation of fragmented patch features, and further regularize their extraction with frequency-domain perturbation consistency to improve guidance reliability. *Without any domain-shift-specific design*, we extensively evaluate **OADisp** on a wide range of 9 datasets under both standard and the more challenging Domain-Generalized settings, together with comprehensive ablation studies and quantitative analyses. **OADisp** consistently improves overall performance, demonstrating the effectiveness and robustness of the resulting representations for open-world category discovery.

1 INTRODUCTION

Modern visual recognition systems (He et al., 2016; Ren et al., 2015; He et al., 2017; Liu et al., 2018) have achieved remarkable success under closed-world settings with exhaustive annotations (Deng et al., 2009; Lin et al., 2014), but they often struggle in dynamic real-world environments where unannotated data and novel categories continuously emerge. While conventional Semi-Supervised Learning (SSL) (Tarvainen & Valpola, 2017; Oliver et al., 2018; Berthelot et al., 2019; Sohn et al., 2020) utilizes unlabeled data within a fixed label space, and Novel Class Discovery (NCD) (Han et al., 2019; 2020; Zhao & Han, 2021) identifies new classes under the strict assumption that unlabeled data contains only novel categories, both fall short of realistic open-world scenarios. To bridge this gap, Generalized Category Discovery (GCD) (Vaze et al., 2022) has been introduced. GCD aims to categorize unlabeled images drawn from both known and novel classes by transferring supervision from a partially labeled set of known categories. By jointly handling the recognition of seen classes and the discovery of unseen ones, GCD provides a more practical and realistic formulation for open-world visual perception. The central challenge of GCD is fundamentally representational: the model must learn category features that are discriminative enough to preserve known-class boundaries, yet sufficiently transferable to support the discovery of previously unseen categories.

Recent progress (Vaze et al., 2022; Rastegar et al., 2024; Choi et al., 2024; Wen et al., 2023; Cao et al., 2025; Liu & Han, 2025; Liu et al., 2025; He et al., 2025) in GCD relies almost exclusively on holistic image embeddings, typically utilizing the global [CLS] token from a DINO-pretrained Vision Transformer (ViT) (Caron et al., 2021; Oquab et al., 2024). While appealing due to their strong nearest-neighbor behavior and unified representation of holistic context, DINO's view-invariance objective

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

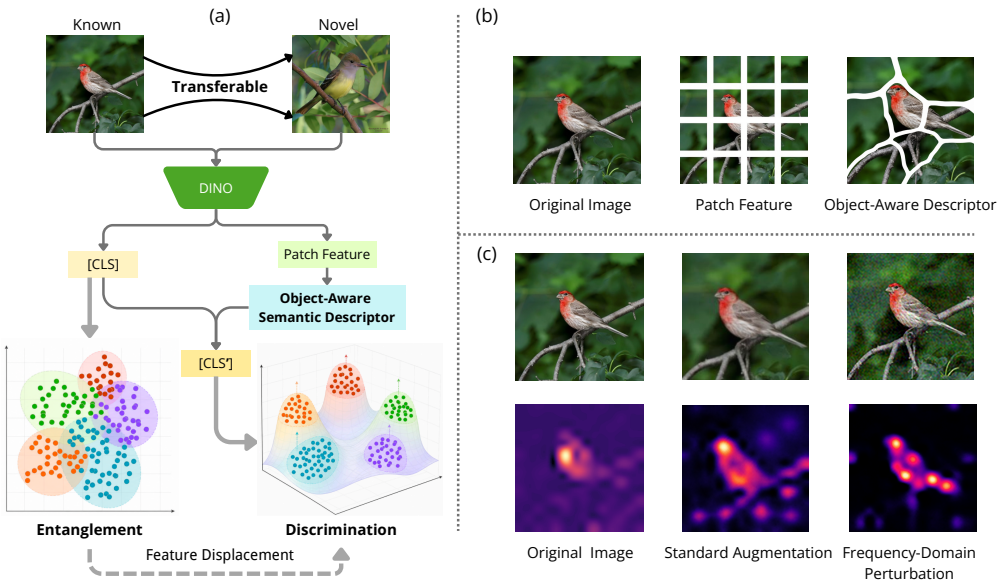

Figure 1: Motivation of **OADisp**. (a) The global [CLS] may entangle category semantics with background and appearance biases, causing overlapping representations, which we guide through object-aware feature displacement to improve discrimination. (b) Fragmented raw patch features motivate object-aware descriptor construction. (c) Frequency-domain perturbation provides targeted regularization for object-aware descriptors beyond standard pixel-space augmentations.

encourages consistency across augmented views, but factors that remain stable under standard pixel-space augmentations may also be retained in the [CLS] representation, including spurious background context and appearance biases (Beyer et al., 2022). Geometrically, such entanglement can cause visually similar but semantically distinct novel categories to occupy overlapping regions in the representation spaces, as shown in Fig. 1(a). A recent attempt addresses this by replacing [CLS] with patch-level features to recover local semantics (Dai et al., 2025). Nevertheless, this strategy can be suboptimal, as it overlooks the holistic semantics and stable global structure encoded in [CLS].

In this work, we propose **O**bject-**A**ware Feature **Disp**lacement (OADisp), a novel plug-and-play paradigm that steers the global [CLS] feature with complementary object-aware guidance derived from patch features, effectively leveraging both holistic image-level semantics and local discriminative cues. Specifically, **OADisp** constructs an image-conditioned target state from auxiliary semantic descriptors and displaces the original [CLS] feature toward this target. For guiding the displacement, we instantiate object-aware semantic descriptors by aggregating patch features through mask-guided reweighting, rather than directly using raw patch tokens, which are spatially fragmented and may correspond to object parts, background regions, or mixed boundary areas, as shown in Fig. 1(b). However, object-level aggregation alone may still retain appearance and background biases, owing to its unsupervised nature (Dai et al., 2026). To better regularize the object-aware descriptors, for which standard pixel-space augmentations are less effective, we introduce frequency-domain perturbation consistency by perturbing low-frequency amplitude while preserving phase, as illustrated in Fig. 1(c). *Without any domain-shift specific design*, extensive experimental results on both standard GCD benchmarks and Domain Generalization settings (Rathore et al., 2025a) demonstrate consistent improvements across multiple baselines. Our contributions are summarized as follows:

(1) We propose **O**bject-**A**ware Feature **Disp**lacement (OADisp), a plug-and-play paradigm that refines the global [CLS] feature through displacement guided by object-aware semantic descriptors, preserving holistic semantics while improving category separability.

(2) We instantiate the descriptors via mask-guided patch aggregation and further regularize them with frequency-domain perturbation consistency to provide more reliable object-level guidance for feature displacement.

(3) *Without any domain-shift-specific design*, we extensively evaluate our approach on *9 datasets* under both standard and more challenging Domain Generalization settings, across *4 distinct*

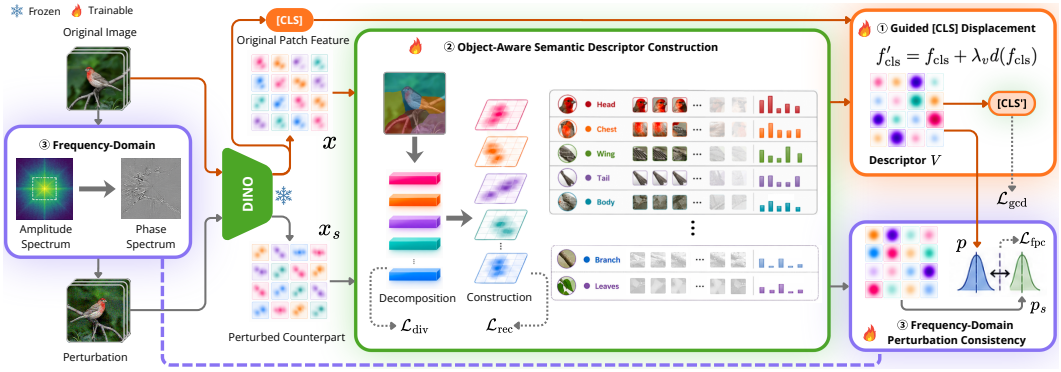

Figure 2: Overview of the proposed **OADisp**. The DINO backbone is frozen throughout training, while the learnable **OADisp** modules and baseline are jointly optimized.

baselines. **OADisp** consistently improves performance, strongly validating the effectiveness and robustness of the resulting representations.

2 Preliminaries

Problem Setup. Generalized category discovery (GCD) aims to recognize labeled known classes while discovering unlabeled novel classes in a unified label space. Formally, let $\mathcal{C}^k$ and $\mathcal{C}^n$ denote the sets of known and novel categories, respectively, with $\mathcal{C}^k \cap \mathcal{C}^n = \varnothing$. We are given a labeled set $\mathcal{D}_l = \{(x_i, y_i)\}_{i=1}^{N_l}$ with $y_i \in \mathcal{C}^k$, and an unlabeled set $\mathcal{D}_u = \{x_j\}_{j=1}^{N_u}$ drawn from $\mathcal{C}^k \cup \mathcal{C}^n$. The goal is to learn a model $h : \mathcal{X} \to \mathcal{C}^k \cup \mathcal{C}^n$ from $\mathcal{D}_l$ and $\mathcal{D}_u$. Let $K = |\mathcal{C}^k|$, $N = |\mathcal{C}^n|$, and $C = K + N$. Following prior work (Wen et al., 2023; He et al., 2025), we assume C is known during training. When it is unavailable, existing methods (Vaze et al., 2022) can be used for estimation.

DINO-pretrained Vision Transformer. Given an input image $x \in \mathbb{R}^{H\times W\times 3}$, we adopt a DINO-pretrained Vision Transformer (ViT) (Caron et al., 2021; Dosovitskiy et al., 2020) $\phi(\cdot)$ as the visual backbone. The image is first divided into $P = HW/L^2$ non-overlapping patches, where L denotes the patch size. Each patch is linearly projected into a d-dimensional token, and a learnable [CLS] token is prepended to the patch sequence together with positional embeddings. The resulting tokens are then processed by a stack of Transformer blocks, producing the final output $Z = [f_{\text{cls}}; F]$, where $f_{\text{cls}} \in \mathbb{R}^d$ denotes the global [CLS] feature and $F = \{f_p\}_{p=1}^{P} \in \mathbb{R}^{P\times d}$ denotes the patch features. Existing GCD methods typically perform category prediction based on f_{cls}, whereas our method further exploits the patch features F to construct more informative representations.

3 Methodology

Our method aims to enhance the discriminability of the [CLS] token for GCD without retraining the backbone. As shown in Figure 2, we first formulate a feature displacement driven by abstract local semantic guidance to adapt the global representation (Sec. 3.1). We then instantiate this guidance with object-aware semantic descriptors constructed through mask-guided patch aggregation (Sec. 3.2) and further regularize their extraction with frequency-domain perturbation consistency to obtain more reliable object-level guidance (Sec. 3.3). Finally, the resulting adapted [CLS] can be seamlessly integrated into existing GCD pipelines in a plug-and-play manner (Sec. 3.4). Throughout training, we keep the DINO backbone frozen and jointly optimize the baseline and all learnable **OADisp** modules.

3.1 Guided Global Feature Displacement

In GCD, the global [CLS] token $f_{\text{cls}} \in \mathbb{R}^d$ extracted from DINO provides a compact and holistic summary of the input image. DINO learns such a representation by enforcing consistency across different augmented views of the same image. However, this view-invariance objective does not guarantee that the invariant factors are purely category semantics. Consequently, background or appearance cues that are insufficiently altered by the augmentations can also be absorbed into the

global representation. For instance, models may exploit category-correlated backgrounds and texture cues as predictive shortcuts (Geirhos et al., 2018; Xiao et al., 2020). Therefore, f_{cls} may entangle core category semantics with spurious context and appearance biases. In GCD, such entanglement is particularly problematic, as semantically distinct novel categories may occupy overlapping regions in the representation space, resulting in poor separability.

Sample-Adaptive Feature Displacement. Instead of discarding the [CLS] representation or retraining the DINO backbone, we seek to unlock its discriminative potential through a simple yet effective feature adaptation. Although the [CLS] feature provides strong global semantics, it may remain entangled with background and appearance cues, limiting the separability of semantically distinct novel categories. To improve its separability while preserving the original global representation, we introduce a *sample-adaptive displacement direction*, which is determined individually for each image by its image-conditioned target, to locally shift [CLS] away from its entangled state. This displacement is performed directly in the original feature space, providing an efficient correction without changing the feature dimensionality or retraining the backbone. Formally, given an image-conditioned target state $c(f_{\text{cls}})$, we adapt f_{cls} by a direct linear update:

$$f'_{\text{cls}} = f_{\text{cls}} + \lambda_v d(f_{\text{cls}}) = f_{\text{cls}} + \lambda_v \frac{c(f_{\text{cls}}) - f_{\text{cls}}}{\|c(f_{\text{cls}}) - f_{\text{cls}}\|_2 + \epsilon}, \tag{1}$$

where $d(f_{\text{cls}})$ denotes the displacement direction, λ_v controls the displacement strength.

Image-conditioned Target State. We now describe how the target state $c(f_{\text{cls}})$ is constructed. Rather than relying solely on the global [CLS] token, we introduce a set of auxiliary semantic descriptors, denoted as $V = \{v_k\}_{k=1}^{N_s}$. These descriptors are constructed from patch-level tokens and are intended to provide complementary local cues for the current image. Since the global [CLS] token and patch tokens were jointly learned during DINO pretraining (Caron et al., 2021), they naturally reside in a shared representational space. This property allows us to directly measure the semantic compatibility between the global feature and descriptor. Given V, we ℓ_2-normalize the global feature as $\hat{f}_{\text{cls}}$ and the descriptors as $\hat{v}_k$ and compute their compatibility scores by cosine similarity, and convert them into normalized assignment weights:

$$\alpha_k = \text{softmax}\left(\hat{f}_{\text{cls}}^{\top} \cdot \hat{v}_k\right). \tag{2}$$

The image-conditioned target state is then obtained as the weighted combination of v_k:

$$c(f_{\text{cls}}) = \sum_{k=1}^{N_s} \alpha_k v_k. \tag{3}$$

Conceptually, $c(f_{\text{cls}})$ represents a target state formed by weighting the semantic descriptors according to their compatibility with the current global representation. In this way, the proposed displacement complements the original DINO representation: *the [CLS] token preserves holistic global semantics, while the auxiliary descriptors contribute complementary local cues, jointly steering the resulting representation f'_{cls} toward being more discriminative and robust for open-world category discovery.*

3.2 Object-Aware Semantic Descriptor Construction

To instantiate the semantic descriptor set V introduced in Section 3.1, a straightforward choice is to directly use the native DINO patch tokens $F = \{f_p\}_{p=1}^{P} \in \mathbb{R}^{P \times d}$. However, individual patch tokens correspond to fixed spatial regions and may capture object parts, background regions, or mixed content around boundaries (Fig. 1(b)), making them spatially fragmented and semantically ambiguous. We therefore construct V as object-aware semantic descriptors by aggregating semantically related patches into coherent units, providing more reliable local semantic guidance for the image-conditioned target state. Visualization results are in Appendix F.

Object-Centric Decomposition. To achieve this, we perform an object-centric decomposition to generate spatial masks that dynamically group patch tokens into semantically coherent components. We instantiate this using a slot-based attention mechanism (Locatello et al., 2020) (details in Appendix C), which aggregates patch information into N_s latent slots, $S = \{s_k\}_{k=1}^{N_s}$. We further decode each slot with a spatial-broadcast decoder (Seitzer et al., 2023), producing a per-slot reconstruction $\hat{F}_k \in \mathbb{R}^{P \times d}$ and mask logits $l_k \in \mathbb{R}^P$. The masks are obtained by applying a softmax across slots:

$$m_k = \text{softmax}_k(l_k), \quad M = \{m_k\}_{k=1}^{N_s} \in \mathbb{R}^{N_s \times P}. \tag{4}$$

Here, each slot s_k can be viewed as *a latent object-level hypothesis*, while each mask m_k provides *a spatial assignment* of patch regions to the corresponding latent component.

Since the decomposition is learned without direct object-level supervision, we regularize it with reconstruction and diversity objectives. Specifically, the per-slot reconstructions are combined according to their masks to recover the original patch feature map as $\hat{F} = \sum_{k=1}^{N_s} \hat{F}_k \odot m_k$. We define

$$\mathcal{L}_{\text{rec}} = \|\hat{F} - F\|_F^2, \qquad \mathcal{L}_{\text{div}} = \left\| \hat{S}\hat{S}^\top - I \right\|_F^2, \tag{5}$$

where $\hat{s}_k = s_k / \|s_k\|_2$ and $\hat{S} = [\hat{s}_1; \ldots; \hat{s}_{N_s}] \in \mathbb{R}^{N_s \times d_s}$. The reconstruction loss $\mathcal{L}_{\text{rec}}$ preserves decomposition fidelity by grounding slots and masks in the patch feature map, while the diversity loss $\mathcal{L}_{\text{div}}$ reduces inter-slot redundancy and encourages complementary latent components.

Mask-Guided Construction. With the masks M obtained above, we construct the object-aware semantic descriptor v_k by reweighting the patch features according to each mask. For each m_k, we normalize it over the spatial dimension to obtain $\hat{m}_{k,p}$ and compute the mask-guided aggregation:

$$v_k = \sum_{p=1}^{P} \hat{m}_{k,p} f_p. \tag{6}$$

Here, $v_k \in \mathbb{R}^d$ denotes the semantic descriptor associated with the k-th latent component. Collecting all descriptors yields $V = \{v_k\}_{k=1}^{N_s} \in \mathbb{R}^{N_s \times d}$.

3.3 Frequency-Domain Perturbation Consistency

To obtain more reliable object-aware semantic descriptors, we further regularize their extraction with a perturbation better suited to the aggregated object-level representation. Standard self-supervised augmentations in DINO, such as cropping, color jittering, and blurring, operate through generic pixel-space transformations and are not specifically designed to regularize the aggregated object-aware descriptors (Fig.1(c)). We therefore introduce a frequency-domain perturbation x_s that modifies low-frequency amplitude while preserving phase. Low-frequency amplitude is closely associated with global appearance and contextual statistics, whereas phase retains much of the spatial structural information (Oppenheim & Lim, 2005; Yang & Soatto, 2020). Enforcing consistency between their category predictions thus promotes more reliable object-level semantic guidance while reducing sensitivity to appearance and contextual variations.

Frequency-Domain Perturbation. Formally, let the Fourier transform of x be $\mathcal{F}(x) = A(x)e^{j\phi(x)}$, where $A(x)$ and $\phi(x)$ denote the amplitude and phase spectra, respectively. We define a centered binary mask M_r to select a low-frequency region, and compute its batch-wise mean amplitude $\bar{A}_{\text{low}}$. We perturb only the selected low-frequency amplitude while keeping the phase unchanged:

$$A_s = A(x) - \alpha M_r \odot \big(A(x) - \bar{A}_{\text{low}}\big) + M_r \odot \xi, \tag{7}$$

where α controls the perturbation strength, $\odot$ denotes element-wise multiplication, and ξ is Gaussian noise applied within the selected region. The frequency-perturbed counterpart is then obtained by $x_s = \mathcal{F}^{-1}\big(A_s e^{j\phi(x)}\big)$. By selectively varying low-frequency appearance and contextual statistics while preserving phase, x_s provides a targeted counterpart of x for the subsequent descriptor consistency regularization.

To regularize the object-aware descriptors under the proposed perturbation, we obtain a compact descriptor representation by averaging them: $\bar{v} = \frac{1}{N_s}\sum_{k=1}^{N_s} v_k$. Processing x_s through the same extraction pipeline yields the corresponding descriptor set V_s and its averaged representation $\bar{v}_s$. We then project $\bar{v}$ and $\bar{v}_s$ into the category space to obtain distributions $p = \text{softmax}(g_c(\bar{v}))$ and $p_s = \text{softmax}(g_c(\bar{v}_s))$. Here, g_c is a learnable linear projection. Finally, we impose a symmetric Kullback-Leibler (KL) divergence:

$$\mathcal{L}_{\text{fpc}} = \frac{1}{2}\Big(D_{\text{KL}}(p \,\|\, p_s) + D_{\text{KL}}(p_s \,\|\, p)\Big). \tag{8}$$

By encouraging consistent predictions under low-frequency perturbations, $\mathcal{L}_{\text{fpc}}$ reduces the sensitivity of descriptor extraction to appearance and contextual variations, yielding more reliable object-aware semantic descriptors for guiding feature displacement.

Table 1: Evaluation on the four fine-grained datasets under standard GCD.

Methods	CUB			Stanford Cars			FGVC-Aircraft			Herbarium-19		
	All	Known	Novel	All	Known	Novel	All	Known	Novel	All	Known	Novel
k-means (Arthur et al., 2007)	34.3	38.9	32.1	12.8	10.6	13.8	16.0	14.4	16.8	13.0	12.2	13.4
RankStats+ (Han et al., 2021)	33.3	51.6	24.2	28.3	61.8	12.1	26.9	36.4	22.2	27.9	55.8	12.8
UNO+ (Fini et al., 2021)	35.1	49.0	28.1	35.5	70.5	18.6	40.3	56.4	32.2	28.3	53.7	14.7
ORCA (Cao et al., 2021)	35.3	45.6	30.2	23.5	50.1	10.7	22.0	31.8	17.1	20.9	30.9	15.5
GCD (Vaze et al., 2022)	51.3	56.6	48.7	39.0	57.6	29.9	45.0	41.1	46.9	35.4	51.0	27.0
PromptCAL (Zhang et al., 2023)	62.9	64.4	62.1	50.2	70.1	40.6	52.2	52.2	52.3	37.0	52.0	28.9
GPC (Zhao et al., 2023)	52.0	55.5	47.5	38.2	58.9	27.4	43.3	40.7	44.8	-	-	-
CiPR (Hao et al., 2024)	57.1	58.7	55.6	47.0	61.5	40.1	-	-	-	-	-	-
SPTNet (Wang et al., 2024)	65.8	68.8	65.1	59.0	79.2	49.3	59.3	61.8	58.1	43.4	58.7	35.2
SelEx (Rastegar et al., 2024)	73.6	75.3	72.8	57.1	64.7	53.3	58.5	75.6	50.3	-	-	-
LegoGCD (Cao et al., 2024)	63.8	71.9	59.8	57.3	75.7	48.4	55.0	61.5	51.7	45.1	57.4	38.5
AllGCD (Cao et al., 2025)	68.4	75.1	65.1	60.5	77.0	52.5	57.4	62.4	54.9	45.2	58.3	38.0
SimGCD (Wen et al., 2023)	60.6	64.4	58.6	52.8	68.6	45.1	54.5	61.5	51.1	43.4	57.1	36.1
+ **OADisp**	65.5	71.0	62.7	58.0	76.4	49.8	57.9	63.9	54.9	45.3	58.6	38.1
Δ	+4.9	+6.6	+4.1	+5.2	+7.8	+4.7	+3.4	+2.4	+3.8	+1.9	+1.5	+2.0
DebGCD (Liu & Han, 2025)	65.3	72.0	62.0	64.5	81.5	56.3	59.2	63.5	57.0	44.1	58.1	36.5
+ **OADisp**	68.4	75.9	64.8	68.8	86.8	60.0	61.7	68.0	58.5	45.5	58.9	38.3
Δ	+3.1	+3.9	+2.8	+4.3	+5.3	+3.7	+2.5	+4.5	+1.5	+1.4	+0.8	+1.8
SEAL (He et al., 2025)	65.8	76.7	60.4	63.2	78.5	55.8	60.4	66.2	57.5	45.1	42.7	47.8
+ **OADisp**	67.9	77.2	63.3	67.2	84.1	59.1	62.3	67.4	59.8	47.7	45.0	50.6
Δ	+2.1	+0.5	+2.9	+4.0	+5.6	+3.3	+1.9	+1.2	+2.3	+2.6	+2.3	+2.8

3.4 Integration into GCD Pipelines

Our method can be seamlessly integrated into existing GCD pipelines in a plug-and-play manner. Given a baseline GCD objective $\mathcal{L}_{\text{gcd}}$ defined on the global representation, we adapt the original $f\text{cls}$ through the proposed object-aware feature displacement to obtain f'_{cls} for category prediction. The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\text{gcd}} + \lambda_{\text{fpc}}\mathcal{L}_{\text{fpc}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{div}}\mathcal{L}_{\text{div}}, \tag{9}$$

where λ_{fpc}, λ_{rec}, and λ_{div} balance the frequency-domain perturbation consistency, reconstruction, and diversity regularization terms, respectively.

4 Experiment

4.1 Experiment Setup

Datasets. We evaluate our method on both standard and Domain-Generalized GCD benchmarks. Following prior standard GCD works, we first conduct experiments on four fine-grained datasets (CUB (Welinder et al., 2010), Stanford Cars (Krause et al., 2013), FGVC-Aircraft (Maji et al., 2013), and Herbarium19 (Tan et al., 2019)) and two generic datasets (CIFAR-100 (Krizhevsky et al., 2009), and ImageNet-100 (Deng et al., 2009)). We adopt the commonly used class-split protocol (Vaze et al., 2022; Wen et al., 2023). To further validate the generalization ability of our method under domain shifts, we also evaluate it on three standard DG-GCD benchmarks (PACS (Li et al., 2017), Office-Home (Venkateswara et al., 2017), and DomainNet (Peng et al., 2019)), following the established protocol (Rathore et al., 2025a;b) for constructing class splits and source-target domain pairs. *Notably, the proposed **OADisp** does not involve any explicit design for domain shift.* Detailed dataset statistics and split information are provided in the Appendix B.

Evaluation Metric and Implementation Details. Following standard practice (Vaze et al., 2022), we evaluate the performance of our method using clustering accuracy (ACC) on the unlabeled set $\mathcal{D}_u$, which is defined as $\text{ACC} = \frac{1}{|\mathcal{D}_u|}\sum_{i=1}^{|\mathcal{D}_u|}\mathbb{1}\left(y_i = h(\hat{y}_i)\right)$, where y_i and $\hat{y}_i$ are the ground-truth and predicted cluster labels, and h denotes the optimal one-to-one mapping between them obtained by the Hungarian algorithm (Kuhn, 1955). We report ACC on *All*, *Known*, and *Novel* classes. For implementation details, we integrate our method as a plug-and-play module into several GCD baselines, and all settings follow the corresponding original papers. More details are in the Appendix B.

4.2 Main Results

Results on Fine-Grained Datasets. As shown in Table 1, we integrate our proposed **OADisp** as a plug-and-play module into several state-of-the-art GCD baselines, including SimGCD (Wen et al., 2023), DebGCD (Liu & Han, 2025), and SEAL (He et al., 2025). The results demonstrate that our method consistently achieves significant performance gains across all fine-grained datasets. Specifically, when applied to the SimGCD baseline, **OADisp** yields an improvement of +4.9% in *All* on CUB, and notably, maintains a solid +1.9% improvement even on the challenging Herbarium-19 dataset. Importantly, rather than sacrificing known-class accuracy to facilitate novel-class discovery, our approach achieves synergistic improvements on both fronts, as evidenced by simultaneous increases of +2.4% in *Known* and +3.8% in *Novel* on the FGVC-Aircraft dataset. Furthermore, **OADisp** demonstrates strong compatibility when integrated with more advanced frameworks. For instance, it provides a gain of +4.3% and +4.0% in *All* on Stanford Cars when combined with DebGCD (Liu & Han, 2025) and SEAL (He et al., 2025), respectively. Appendix D further compares **OADisp** with the related plug-and-play APL (Dai et al., 2025).

Results on Generic Datasets. As summarized in Table 2, the results demonstrate that **OADisp** maintains its effectiveness even in broader classification scenarios. On the CIFAR-100 benchmark, our module significantly enhances DebGCD (Liu & Han, 2025), achieving the substantial +3.2% and +6.3% improvement in *All* and *Novel*. Similarly, on the larger-scale ImageNet-100 dataset, **OADisp** provides consistent benefits across different baselines. Even when integrated with high-performance frameworks like SEAL (He et al., 2025), although known-class accuracy experiences a marginal dip, our method still secures a net improvement of +1.2% in overall performance. These gains demonstrate that guiding feature displacement with object-aware semantic descriptors improves fine-grained discrimination while enhancing the robustness of global representations across a wide spectrum of visual categories, supporting the practical effectiveness of **OADisp** for open-world category discovery.

Table 2: Evaluation on the two generic datasets under standard GCD.

Method	CIFAR-100			ImageNet-100		
	All	Known	Novel	All	Known	Novel
k-means (Arthur et al., 2007)	52.0	52.2	50.8	72.7	75.5	71.3
RankStats+ (Han et al., 2021)	58.2	77.6	19.3	37.1	61.6	24.8
UNO+ (Fini et al., 2021)	69.5	80.6	47.2	70.3	95.0	57.9
ORCA (Cao et al., 2021)	73.5	92.6	63.9	81.8	86.2	79.6
GCD (Vaze et al., 2022)	73.0	76.2	66.5	74.1	89.8	66.3
PromptCAL (Zhang et al., 2023)	81.2	84.2	75.3	83.1	92.7	78.3
GPC (Zhao et al., 2023)	75.4	84.6	60.1	75.3	93.4	66.7
CiPR (Hao et al., 2024)	81.5	82.4	79.7	80.5	84.9	78.3
SPTNet (Wang et al., 2024)	81.3	84.3	75.6	85.4	93.2	81.4
SelEx (Rastegar et al., 2024)	82.3	85.3	76.3	83.1	93.6	77.8
LegoGCD (Cao et al., 2024)	81.8	81.4	82.5	86.3	94.5	82.1
AllGCD (Cao et al., 2025)	82.3	82.3	82.2	86.5	94.7	82.3
SimGCD (Wen et al., 2023)	80.9	80.4	81.7	83.6	91.7	79.6
+ **OADisp**	81.7	81.4	82.4	86.1	93.5	82.4
Δ	+0.8	+1.0	+0.7	+2.5	+1.8	+2.8
DebGCD (Liu & Han, 2025)	81.9	83.6	78.5	85.2	95.7	80.0
+ **OADisp**	85.1	85.2	84.8	86.6	96.8	81.4
Δ	+3.2	+1.4	+6.3	+1.4	+0.9	+1.4
SEAL (He et al., 2025)	79.6	80.8	77.1	82.2	94.8	76.1
+ **OADisp**	81.9	82.6	80.5	83.4	94.2	78.1
Δ	+2.3	+1.8	+3.4	+1.2	-0.6	+2.0

Results on DG Datasets. We further evaluate the robustness of **OADisp** under the Domain Generalization (DG) setting, as summarized in Table 3. When integrated into the state-of-the-art HIDISC (Rathore et al., 2025b) baseline, **OADisp** consistently improves performance across all three benchmarks under substantial distribution shifts. Notably, on PACS, **OADisp** improves *Novel* accuracy by +4.8%, demonstrating its effectiveness for discovering unseen categories in target domains. Although the gains on Office-Home and DomainNet are more moderate, they are achieved *without introducing any specific design tailored to domain shifts.* Together with the consistent improvements under standard GCD, these results highlight the strong *transferability* of object-aware feature displacement across different settings and support the *robustness* of the learned representation.

4.3 Ablation Study

Component Analysis. To systematically assess the contribution of each component, we conduct an ablation study on CUB, as shown in Table 4.

- For variant **#2**, we remove the object-aware descriptor construction and *directly use raw patch features as the descriptors* for displacement. Despite retaining both feature displacement and frequency perturbation, this variant achieves only 62.5% in *All* and 59.2% in *Novel*, whereas the object-aware construction in **#4** improves them to 65.5% and 62.7%. This direct comparison demonstrates that structured, mask-guided aggregation provides substantially more effective local guidance than using fragmented raw patch features.

Table 3: Evaluation on the three datasets under the Domain Generalization setting.

Methods	PACS			Office-Home			DomainNet		
	All	Known	Novel	All	Known	Novel	All	Known	Novel
ViT (Dosovitskiy et al., 2020)	42.0	50.9	33.2	26.2	29.1	21.6	25.4	26.5	22.4
GCD (Vaze et al., 2022)	52.3	62.2	38.4	52.7	54.2	50.3	27.4	27.9	26.1
DG^{2}CD-Net (Rathore et al., 2025a)	73.3	75.3	72.6	53.9	53.4	54.3	29.0	30.4	25.5
HIDISC (Rathore et al., 2025b)	75.9	77.8	73.2	58.1	60.8	53.4	30.4	31.1	28.1
+ **OADisp**	77.0	76.4	78.0	59.1	61.9	54.1	31.4	31.8	29.9
Δ	+1.1	-1.4	+4.8	+1.0	+1.1	+0.7	+1.0	+0.7	+1.8

Table 4: Ablation of proposed components.

	Guided [CLS] Displacement	Object-Aware Descriptor	Frequency Perturbation	All	Known	Novel
#1	-	-	-	60.6	64.4	58.6
#2	✓	-	✓	62.5	69.1	59.2
#3	✓	✓	-	63.4	67.6	61.3
#4	✓	✓	✓	65.5	71.0	62.7

Table 5: Ablation of slot numbers.

N_s	All	Known	Novel
3	64.8	71.1	61.6
4	65.1	69.8	62.7
5	65.5	71.0	62.7
6	64.4	69.5	61.9
7	63.9	68.4	61.6

- Removing frequency-domain perturbation **#3** also leads to a performance drop, indicating that it provides complementary regularization for descriptor construction. Overall, the full model **#4** achieves the best performance, validating the complementary design of the proposed components.
- Beyond component removal, Appendix E.3 further analyzes *alternative constructions of the displacement target* by varying the aggregation strategy (uniform mean pooling vs. [CLS]-conditioned weighting) and the descriptor set (raw patch features vs. object-aware descriptors), demonstrating the necessity of our design.

Hyperparameter Sensitivity. We further analyze the impact of the number of slots N_s in Table 5 on the CUB dataset. Specifically, when $N_s < 5$, the model's performance is slightly lower, as few slots may provide insufficient capacity to fully capture and decompose all components within a scene. On the other hand, when $N_s > 5$, accuracy begins to decline, which can be attributed to over-segmentation where a single semantic entity is fragmented into too many latent slots, thereby reducing the stability of the aggregation process. Overall, the results are relatively stable, indicating that our method is not sensitive to the number of slots. More ablation analyses for other hyperparameters, including λ_v,λ_{rec} and λ_{div}, are in Appendix E.

4.4 Analysis

Statistical Analysis. We analyze the effectiveness of **OADisp** from the perspective of representation separation. We employ the Silhouette Coefficients (Rousseeuw, 1987) to quantify the clustering quality of the learned representations. This metric measures the relative cohesion within a cluster and separation from other clusters, with higher values indicating better-separated representations. As shown in Fig. 3, the adapted representation f'_{cls} consistently achieves a higher Silhouette Score than the original f_{cls}. Specifically, compared to the baseline, **OADisp** yields a relative improvement of nearly 8% on the fine-grained CUB and over 27% on ImageNet-100. These results indicate that object-aware feature displacement produces a more separable global representation, providing further support for its effectiveness in category discovery.

Visualization. To intuitively assess the impact of **OADisp**, we compare the t-SNE distributions of the original f_{cls} and the adapted f'_{cls} embeddings in Figure 4. While the original f_{cls} forms recognizable groups, they still exhibit relatively blurred boundaries and local entanglement between several categories (Fig. 4a). In contrast, the adapted f'_{cls} further elevates the clustering quality, resulting in tighter intra-class cohesion and wider inter-class separation (Fig. 4b). This visualization suggests that guiding the global representation with object-aware semantic descriptors improves category separation, yielding a more discriminative representation space for category discovery.

Computational Overhead. Integrating **OADisp** into SimGCD increases the parameter count from 92.17M to 106.27M (+15.30%) and the batch inference time from 32.41ms to 44.02ms (1.358×). While this introduces additional computational cost, **OADisp** is evaluated *without setting-specific*

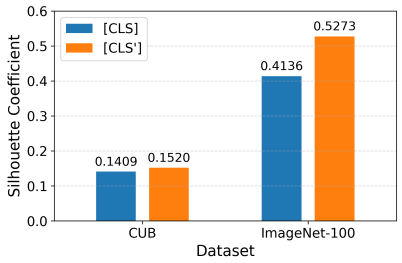

Figure 3: Silhouette Coefficients

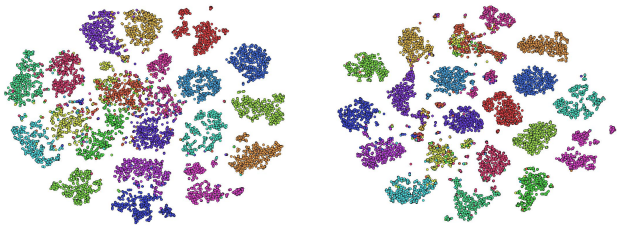

Figure 4: t-SNE Visualization on f_{cls} and f'_{cls}.

redesign across both standard and the more challenging DG setting, covering 9 datasets and 4 different baselines, and consistently delivers positive performance gains. We therefore view the additional overhead as a practical trade-off for the broad and robust improvements brought by the proposed plug-and-play representation adaptation.

5 Related Works

Generalized Category Discovery (GCD) aims to categorize unlabeled samples from both known and novel classes by leveraging labeled data from only known classes. Existing methods can be broadly grouped into three paradigms. Non-parametric methods (Vaze et al., 2022; Rastegar et al., 2024; Choi et al., 2024) usually learn clustering-friendly representations and perform label assignment through feature clustering. Parametric methods (Wen et al., 2023; Zhao et al., 2023) replace post-hoc clustering with learnable classifiers, facilitating better utilization of unlabeled data through pseudo-label learning (Cao et al., 2024; 2025; Liu & Han, 2025) or semantic hierarchical construction (Hao et al., 2024; Liu & Han, 2025; He et al., 2025). Prompt-learning methods (Zhang et al., 2023; Wang et al., 2024) provide an alternative perspective by learning prompts to adapt data representations. Notably, unlike most existing approaches that primarily rely on the [CLS] token, APL (Dai et al., 2025) advocates for the use of rich patch features. In contrast to both, we explore the complementary use of global and local features for robust category discovery. Moreover, given that category discovery often encounters domain shifts (Ben-David et al., 2006; Blanchard et al., 2011) in real-world scenarios, DG-GCD (Rathore et al., 2025a;b) has been introduced to extend this task by discovering categories in unseen target domains. In this work, we design a plug-and-play paradigm for standard GCD and further validate its effectiveness under the more challenging Domain Generalization setting.

Object-Centric Learning (OCL) aims to decompose visual scenes into a set of entity-level representations, providing structured abstractions beyond global image features. Early methods such as IODINE (Greff et al., 2019) learn object representations through iterative variational inference, while MONet (Burgess et al., 2019) decomposes scenes into object-like components using attention masks and variational autoencoding. Slot Attention (Locatello et al., 2020) later introduces a simple and general slot-based mechanism, where a set of learnable slots iteratively compete to bind with different objects or regions in the input. Beyond synthetic scene decomposition, slot-based OCL (Seitzer et al., 2023) has shown potential in more realistic vision tasks, including domain adaptation (Dai et al., 2026), generative modeling (Jiang et al., 2023), and robot interactions (Rezazadeh et al., 2024). Inspired by these works, we use object-level hypotheses not as final predictions, but as intermediate structural cues for aggregating patch features into object-aware semantic descriptors.

6 Conclusion

In this work, we presented **OADisp**, a plug-and-play approach for GCD that adapts the global [CLS] representation through feature displacement. By constructing object-aware semantic descriptors via mask-guided patch aggregation and further regularizing them with frequency-domain perturbation consistency, **OADisp** provides reliable local semantic guidance while preserving holistic global information. Extensive experiments across standard and the more challenging DG setting, together with comprehensive analyses and visualizations, demonstrate the effectiveness and robustness of our approach. We hope that object-aware feature displacement provides a simple yet effective direction for integrating complementary global and local semantics in open-world category discovery.

AI USE STATEMENT

In this work, we used generative AI tools for grammar correction and language polishing. We have not used generative AI tools for generating research ideas, developing methods, or conducting experiments, and other required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work. All AI-assisted edits were manually reviewed by the authors to ensure that they did not alter the intended meaning or technical content. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have made efforts to ensure the reproducibility of our work by providing detailed descriptions of the proposed method, experimental settings, implementation details, and evaluation protocols in the main paper and Appendix. Upon acceptance, we will publicly release the source code.

REFERENCES

David Arthur, Sergei Vassilvitskii, et al. k-means++: The advantages of careful seeding. In *Soda*, volume 7, pp. 1027–1035, 2007.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.

Lucas Beyer, Xiaohua Zhai, and Alexander Kolesnikov. Better plain vit baselines for imagenet-1k. *arXiv preprint arXiv:2205.01580*, 2022.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. *Advances in neural information processing systems*, 24, 2011.

Christopher P Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. Monet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*, 2019.

Kaidi Cao, Maria Brbic, and Jure Leskovec. Open-world semi-supervised learning. *arXiv preprint arXiv:2102.03526*, 2021.

Xinzi Cao, Xiawu Zheng, Guanhong Wang, Weijiang Yu, Yunhang Shen, Ke Li, Yutong Lu, and Yonghong Tian. Solving the catastrophic forgetting problem in generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16880–16889, 2024.

Xinzi Cao, Ke Chen, Feidiao Yang, Xiawu Zheng, Yonghong Tian, and Yutong Lu. Allgcd: Leveraging all unlabeled data for generalized category discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3293–3303, 2025.

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*, pp. 9650–9660, 2021.

Sua Choi, Dahyun Kang, and Minsu Cho. Contrastive mean-shift learning for generalized category discovery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 23094–23104, 2024.

Boyang Dai, Zeng Fan, Zihao Qi, Meng Lou, and Yizhou Yu. Cgsa: Class-guided slot-aware adaptation for source-free object detection. In *The Fourteenth International Conference on Learning Representations*, 2026.

Qiyuan Dai, Hanzhuo Huang, Yu Wu, and Sibei Yang. Adaptive part learning for fine-grained generalized category discovery: A plug-and-play enhancement. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 25444–25453, 2025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

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.

Enrico Fini, Enver Sangineto, Stéphane Lathuilière, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9284–9292, 2021.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. In *International conference on machine learning*, pp. 2424–2433. PMLR, 2019.

Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8401–8409, 2019.

Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. *arXiv preprint arXiv:2002.05714*, 2020.

Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Autonovel: Automatically discovering and learning novel visual categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6767–6781, 2021.

Shaozhe Hao, Kai Han, and Kwan-Yee K Wong. Cipr: An efficient framework with cross-instance positive relations for generalized category discovery. *Transactions on Machine Learning Research*, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2961–2969, 2017.

Zhenqi He, Yuanpei Liu, and Kai Han. Seal: Semantic-aware hierarchical learning for generalized category discovery. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Jindong Jiang, Fei Deng, Gautam Singh, and Sungjin Ahn. Object-centric slot diffusion. *Advances in Neural Information Processing Systems*, 36:8563–8601, 2023.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 554–561, 2013.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5542–5550, 2017.

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*, pp. 740–755. Springer, 2014.

Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8759–8768, 2018.

Yuanpei Liu and Kai Han. Debgcd: Debiased learning with distribution guidance for generalized category discovery. In *The Thirteenth International Conference on Learning Representations*, 2025.

Yuanpei Liu, Zhenqi He, and Kai Han. Hyperbolic category discovery. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9891–9900, 2025.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in neural information processing systems*, 33:11525–11538, 2020.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. *Advances in neural information processing systems*, 31, 2018.

Alan V Oppenheim and Jae S Lim. The importance of phase in signals. *Proceedings of the IEEE*, 69 (5):529–541, 2005.

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. *Transactions on Machine Learning Research Journal*, 2024.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1406–1415, 2019.

Sarah Rastegar, Mohammadreza Salehi, Yuki M Asano, Hazel Doughty, and Cees GM Snoek. Selex: Self-expertise in fine-grained generalized category discovery. In *European Conference on Computer Vision*, pp. 440–458. Springer, 2024.

Vaibhav Rathore, Saikat Dutta, Sarthak Mehrotra, Zsolt Kira, Biplab Banerjee, et al. When domain generalization meets generalized category discovery: An adaptive task-arithmetic driven approach. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4905–4915, 2025a.

Vaibhav Rathore, Divyam Gupta, and Biplab Banerjee. Hidisc: A hyperbolic framework for domain generalization with generalized category discovery. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

Alireza Rezazadeh, Athreyi Badithela, Karthik Desingh, and Changhyun Choi. Slotgnn: Unsupervised discovery of multi-object representations and visual dynamics. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 17508–17514. IEEE, 2024.

Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, Thomas Brox, et al. Bridging the gap to real-world object-centric learning. In *The Eleventh International Conference on Learning Representations*, 2023.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

Kiat Chuan Tan, Yulong Liu, Barbara Ambrose, Melissa Tulig, and Serge Belongie. The herbarium challenge 2019 dataset. *arXiv preprint arXiv:1906.05372*, 2019.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7492–7501, 2022.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5018–5027, 2017.

Hongjun Wang, Sagar Vaze, and Kai Han. Sptnet: An efficient alternative framework for generalized category discovery with spatial prompt tuning. *arXiv preprint arXiv:2403.13684*, 2024.

Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010.

Xin Wen, Bingchen Zhao, and Xiaojuan Qi. Parametric classification for generalized category discovery: A baseline study. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16590–16600, 2023.

Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020.

Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4085–4095, 2020.

Sheng Zhang, Salman Khan, Zhiqiang Shen, Muzammal Naseer, Guangyi Chen, and Fahad Shahbaz Khan. Promptcal: Contrastive affinity learning via auxiliary prompts for generalized novel category discovery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3479–3488, 2023.

Bingchen Zhao and Kai Han. Novel visual category discovery with dual ranking statistics and mutual knowledge distillation. *Advances in Neural Information Processing Systems*, 34:22982–22994, 2021.

Bingchen Zhao, Xin Wen, and Kai Han. Learning semi-supervised gaussian mixture models for generalized category discovery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16623–16633, 2023.

A THEORETICAL ANALYSIS

We provide a theoretical analysis of **OADisp** from three complementary perspectives. First, we analyze the reliability of object-aware semantic descriptors, showing how mask-guided aggregation can reduce the contamination of fragmented raw patch features under appropriate decomposition quality. Second, we characterize frequency-domain perturbation consistency as a local regularizer that penalizes the sensitivity of descriptor predictions along the applied perturbation directions. Finally, we study the effect of guided feature displacement on the global representation, showing that its bounded update can preserve sufficiently confident predictions and, under suitable alignment conditions, improve class margins and pairwise feature separation. Together, these analyses provide theoretical support for using reliable local semantic guidance to adapt the global [CLS] representation while controlling deviation from the original global representation.

A.1 SETTING AND NOTATION

We follow the notation and construction introduced in Section 3. Given an input image x, let $x_s = T_\eta(x)$ denote a frequency-perturbed counterpart associated with a sampled perturbation parameter η, as described in Section 3.3. The DINO-pretrained ViT produces the global [CLS] feature and patch features, denoted by $f_{\mathrm{cls}}(x)$ and $F(x) = \{f_p(x)\}_{p=1}^{P}$, respectively.

The object-aware semantic descriptor set $V(x) = \{v_k(x)\}_{k=1}^{N_s}$ is constructed by the mask-guided patch aggregation in Eq. 6. For frequency-domain perturbation consistency, we use the averaged descriptor representation $\bar{v}(x) = \frac{1}{N_s}\sum_{k=1}^{N_s} v_k(x)$, and define its category prediction as $p(x) = \mathrm{softmax}(g_c(\bar{v}(x)))$. Applying the same descriptor extraction pipeline to x_s gives $\bar{v}(x_s)$ and $p(x_s)$. For brevity, we write $p = p(x)$ and $p_s = p(x_s)$.

The image-conditioned target state $c(f_{\mathrm{cls}}(x))$ is defined by the compatibility-weighted combination of semantic descriptors in Eqs. 2–3. For notational convenience, define $f(x) = f_{\mathrm{cls}}(x), c(x) = c(f_{\mathrm{cls}}(x)), d(x) = \frac{c(x)-f(x)}{\|c(x)-f(x)\|_2+\epsilon}, f'(x) = f(x) + \lambda_v d(x)$. Notably, when analyzing a single fixed input, as in Theorem 1, we suppress the argument x and simply write f, c, d, and f'. For pairwise analysis, we retain the explicit dependence on the input.

A.2 ASSUMPTIONS

Assumption 1. Patch contamination and mask-guided aggregation. For a latent semantic component k, let $\mu_k \in \mathbb{R}^d$ denote its ideal semantic prototype in the patch-feature space, and let $\mathcal{R}_k \subseteq \{1, \dots, P\}$ denote the set of patches associated with this component. To isolate the effect of aggregation, we condition on the learned normalized mask $\hat{m}_k(x)$ and treat its weights as fixed in the following local analysis. For $p \in \mathcal{R}_k$, assume

$$f_p(x) = \mu_k + \xi_p, \qquad \mathbb{E}[\xi_p \mid \hat{m}_k] = 0, \qquad \mathbb{E}[\|\xi_p\|_2^2 \mid \hat{m}_k] \le \sigma^2, \tag{A.1}$$

and, for $p \neq q$,

$$\mathbb{E}[\xi_p^\top \xi_q \mid \hat{m}_k] = 0. \tag{A.2}$$

For patches outside the component, assume

$$\mathbb{E}\big[\|f_p(x) - \mu_k\|_2^2 \,\big|\, p \notin \mathcal{R}_k, \hat{m}_k\big] \ge c_{\mathrm{out}}^2, \qquad \|f_p(x) - \mu_k\|_2 \le C, \tag{A.3}$$

where $0 < c_{\mathrm{out}} \le C$. For $\hat{m}_k(x)$, define the mask leakage and effective in-component aggregation size as

$$\tau_k = \sum_{p \notin \mathcal{R}_k} \hat{m}_{k,p}(x), \qquad n_{\mathrm{eff},k} = \frac{1}{\sum_{p \in \mathcal{R}_k} \hat{m}_{k,p}(x)^2}. \tag{A.4}$$

Assumption 2. Local smoothness under frequency-domain perturbations. Let $T_\eta(x)$ denote a frequency-perturbed counterpart of x, where η parameterizes the low-frequency amplitude perturbation and $T_0(x) = x$. Define the descriptor logits under this perturbation as

$$z_\theta(\eta) = g_c(\bar{v}(T_\eta(x))). \tag{A.5}$$

We assume that $z_\theta(\eta)$ is twice continuously differentiable in a neighborhood of $\eta = 0$. For the perturbation distribution induced by our frequency-domain transformation, define its second-moment matrix as

$$\Sigma_{\mathrm{fd}} = \mathbb{E}_\eta\left[\eta\eta^\top\right]. \tag{A.6}$$

We consider the local regime in which $\|\eta\|_2$ is sufficiently small for a second-order expansion around $\eta = 0$ to be valid.

Assumption 3. Classifier regularity. For the following margin analysis, we consider a linear class-scoring head with class-r logit

$$h_r(f) = w_r^\top f + b_r, \tag{A.7}$$

with bounded classifier weights:

$$L_{\mathrm{row}} = \max_r \|w_r\|_2 < \infty. \tag{A.8}$$

A.3 Lemma 1. Mask-Guided Aggregation Reduces Patch-Level Contamination

Claim. Condition on a fixed learned mask $\hat{m}_k(x)$, and let a raw patch token be sampled uniformly from all P patches. For notational simplicity, we omit this conditioning in the expectations below. Its expected squared deviation from the component prototype μ_k satisfies

$$\mathbb{E}\|f_p(x) - \mu_k\|_2^2 \geq (1 - \pi_k)c_{\mathrm{out}}^2, \qquad \pi_k = \frac{|\mathcal{R}_k|}{P}. \tag{A.9}$$

In contrast, the mask-guided object-aware semantic descriptor $v_k(x) = \sum_{p=1}^P \hat{m}_{k,p}(x) f_p(x)$ satisfies

$$\mathbb{E}\|v_k(x) - \mu_k\|_2^2 \leq \frac{2\sigma^2}{n_{\mathrm{eff},k}} + 2C^2\tau_k^2. \tag{A.10}$$

Therefore, the object-aware semantic descriptor has a strictly smaller expected deviation from μ_k than a uniformly sampled raw patch whenever

$$\frac{2\sigma^2}{n_{\mathrm{eff},k}} + 2C^2\tau_k^2 < (1 - \pi_k)c_{\mathrm{out}}^2. \tag{A.11}$$

Proof. For a uniformly sampled raw patch, the law of total expectation gives

$$\mathbb{E}\|f_p(x) - \mu_k\|_2^2 = \pi_k\mathbb{E}\left[\|f_p(x) - \mu_k\|_2^2 \mid p \in \mathcal{R}_k\right] + (1 - \pi_k)\mathbb{E}\left[\|f_p(x) - \mu_k\|_2^2 \mid p \notin \mathcal{R}_k\right].$$

The first term is nonnegative, while Assumption 1 gives

$$\mathbb{E}\left[\|f_p(x) - \mu_k\|_2^2 \mid p \notin \mathcal{R}_k\right] \geq c_{\mathrm{out}}^2.$$

Hence,

$$\mathbb{E}\|f_p(x) - \mu_k\|_2^2 \geq (1 - \pi_k)c_{\mathrm{out}}^2,$$

which proves Eq. A.9.

We next analyze the mask-guided descriptor. Since the mask weights are normalized over spatial positions, $\sum_{p=1}^P \hat{m}_{k,p}(x) = 1$. Therefore, using the normalized mask weights, we decompose

$$v_k(x) - \mu_k = \sum_{p\in\mathcal{R}_k} \hat{m}_{k,p}(x)\big(f_p(x) - \mu_k\big) + \sum_{p\notin\mathcal{R}_k} \hat{m}_{k,p}(x)\big(f_p(x) - \mu_k\big).$$

Denote the two terms by A_k and B_k, respectively. Using $\|a + b\|_2^2 \leq 2\|a\|_2^2 + 2\|b\|_2^2$, we obtain

$$\mathbb{E}\|v_k(x) - \mu_k\|_2^2 \leq 2\mathbb{E}\|A_k\|_2^2 + 2\mathbb{E}\|B_k\|_2^2.$$

For the in-component term, Assumption 1 gives $f_p(x) - \mu_k = \xi_p$. Since the residuals are conditionally zero-mean and uncorrelated,

$$\mathbb{E}\|A_k\|_2^2 = \mathbb{E}\left\|\sum_{p\in\mathcal{R}_k} \hat{m}_{k,p}(x)\xi_p\right\|_2^2 \leq \sigma^2 \sum_{p\in\mathcal{R}_k} \hat{m}_{k,p}(x)^2 = \frac{\sigma^2}{n_{\mathrm{eff},k}}.$$

For the leaked out-of-component term, Assumption 1 implies

$$\|B_k\|_2 \le \sum_{p \notin \mathcal{R}_k} \hat{m}_{k,p}(x)\|f_p(x) - \mu_k\|_2 \le C \sum_{p \notin \mathcal{R}_k} \hat{m}_{k,p}(x) = C\tau_k.$$

Therefore,

$$\mathbb{E}\|B_k\|_2^2 \le C^2\tau_k^2.$$

Combining the two terms proves Eq. A.10.

Finally, if Eq. A.11 holds, then the upper bound in Eq. A.10 is strictly smaller than the lower bound in Eq. A.9. Hence,

$$\mathbb{E}\|v_k(x) - \mu_k\|_2^2 < \mathbb{E}\|f_p(x) - \mu_k\|_2^2.$$

The condition is favored by small mask leakage τ_k, a large effective aggregation size $n_{\mathrm{eff},k}$, and sufficiently distinct out-of-component patches characterized by c_{out}. This completes the proof.

A.4 Lemma 2. Frequency-Domain Perturbation Consistency Penalizes Local Sensitivity

Claim. Let $z_\theta(\eta) = g_c(\bar{v}(T_\eta(x)))\,, p_\eta = \mathrm{softmax}(z_\theta(\eta))$, and denote the unperturbed prediction by $p = p_0$. Let $J_\eta = \left.\frac{\partial z_\theta(\eta)}{\partial \eta}\right|_{\eta=0}$ be the Jacobian of the category logits with respect to the frequency-domain perturbation. For sufficiently small perturbations with second-moment matrix $\Sigma_{\mathrm{fd}} = \mathbb{E}[\eta\eta^\top]$, the symmetric consistency loss satisfies

$$\mathbb{E}_\eta\left[\mathcal{L}_{\mathrm{fpc}}(\eta)\right] = \frac{1}{2}\,\mathrm{tr}\left(J_\eta^\top G(p) J_\eta \Sigma_{\mathrm{fd}}\right) + o(\mathrm{tr}(\Sigma_{\mathrm{fd}}))\,, \tag{A.12}$$

where $G(p) = \mathrm{diag}(p) - pp^\top$. Moreover, the leading term can be written as

$$\mathrm{tr}\left(J_\eta^\top G(p) J_\eta \Sigma_{\mathrm{fd}}\right) = \left\| G(p)^{1/2} J_\eta \Sigma_{\mathrm{fd}}^{1/2} \right\|_F^2 \ge 0. \tag{A.13}$$

Therefore, minimizing $\mathcal{L}_{\mathrm{fpc}}$ locally penalizes prediction sensitivity along the frequency-perturbation directions induced by the proposed transformation.

Proof. Let $z = z_\theta(0), z_\eta = z_\theta(\eta), \Delta z = z_\eta - z$. By Assumption 2 and a Taylor expansion around $\eta = 0$,

$$\Delta z = J_\eta \eta + o(\|\eta\|_2).$$

For nearby softmax distributions, the second-order expansion of the KL divergence in logit space gives

$$D_{\mathrm{KL}}(p\|p_\eta) = \frac{1}{2}\Delta z^\top G(p)\Delta z + o(\|\Delta z\|_2^2),$$

where $p_\eta = \mathrm{softmax}(z_\eta)$. The reverse KL divergence has the same second-order term:

$$D_{\mathrm{KL}}(p_\eta\|p) = \frac{1}{2}\Delta z^\top G(p)\Delta z + o(\|\Delta z\|_2^2).$$

Hence,

$$\mathcal{L}_{\mathrm{fpc}}(\eta) = \frac{1}{2}\left(D_{\mathrm{KL}}(p\|p_\eta) + D_{\mathrm{KL}}(p_\eta\|p)\right) = \frac{1}{2}\Delta z^\top G(p)\Delta z + o(\|\Delta z\|_2^2).$$

Substituting $\Delta z = J_\eta\eta + o(\|\eta\|_2)$ yields

$$\mathcal{L}_{\mathrm{fpc}}(\eta) = \frac{1}{2}\eta^\top J_\eta^\top G(p) J_\eta \eta + o(\|\eta\|_2^2).$$

Taking expectation and using $\mathbb{E}[\eta^\top A\eta] = \mathrm{tr}\left(A\,\mathbb{E}[\eta\eta^\top]\right)$ gives

$$\mathbb{E}_\eta[\mathcal{L}_{\mathrm{fpc}}(\eta)] = \frac{1}{2}\,\mathrm{tr}\left(J_\eta^\top G(p) J_\eta \Sigma_{\mathrm{fd}}\right) + o(\mathrm{tr}(\Sigma_{\mathrm{fd}}))\,,$$

in the small-perturbation regime, proving Eq. A.12.

Since $G(p)$ and Σ_{fd} are positive semidefinite,

$$\mathrm{tr}\left(J_\eta^\top G(p) J_\eta \Sigma_{\mathrm{fd}}\right) = \mathrm{tr}\left(\Sigma_{\mathrm{fd}}^{1/2} J_\eta^\top G(p) J_\eta \Sigma_{\mathrm{fd}}^{1/2}\right) = \left\| G(p)^{1/2} J_\eta \Sigma_{\mathrm{fd}}^{1/2} \right\|_F^2,$$

which proves Eq. A.13. Thus, frequency-domain perturbation consistency locally suppresses variations in the category prediction along the perturbation directions emphasized by Σ_{fd}. This completes the proof.

A.5 Theorem 1. Bounded Guided Displacement Preserves Confident Predictions and Conditionally Improves Margins

Claim. For $\lambda_v \geq 0$, the guided feature displacement is bounded:

$$\|f' - f\|_2 \leq \lambda_v. \tag{A.14}$$

Under Assumption 3, this further implies that, for any class r,

$$|h_r(f') - h_r(f)| \leq L_{\text{row}} \lambda_v. \tag{A.15}$$

If the original prediction $y^\star = \arg\max_r h_r(f)$ has margin

$$\gamma = h_{y^\star}(f) - \max_{r \neq y^\star} h_r(f) > 2L_{\text{row}} \lambda_v, \tag{A.16}$$

then the predicted class remains unchanged after displacement.

Moreover, for a target class y and a competing class r, define their pairwise margin as $\gamma_{y,r}(f) = h_y(f) - h_r(f)$. If $\lambda_v > 0$ and the displacement direction satisfies $(w_y - w_r)^\top d > 0$, then the displacement strictly increases this margin:

$$\gamma_{y,r}(f') - \gamma_{y,r}(f) > 0. \tag{A.17}$$

Proof. From Section A.1,

$$f' - f = \lambda_v d, \qquad d = \frac{c - f}{\|c - f\|_2 + \epsilon}.$$

Since

$$\|d\|_2 = \frac{\|c - f\|_2}{\|c - f\|_2 + \epsilon} \leq 1,$$

we obtain

$$\|f' - f\|_2 = \lambda_v \|d\|_2 \leq \lambda_v,$$

which proves Eq. A.14.

For any class r, by Assumption 3 and the Cauchy–Schwarz inequality,

$$|h_r(f') - h_r(f)| = |w_r^\top (f' - f)| \leq \|w_r\|_2 \|f' - f\|_2 \leq L_{\text{row}} \lambda_v.$$

This proves Eq. A.15.

Let $y^\star$ denote the original predicted class. For any competing class $r \neq y^\star$,

$$h_{y^\star}(f') - h_r(f') = h_{y^\star}(f) - h_r(f) + \big(h_{y^\star}(f') - h_{y^\star}(f)\big) - \big(h_r(f') - h_r(f)\big).$$

Using Eq. A.15,

$$h_{y^\star}(f') - h_r(f') \geq h_{y^\star}(f) - h_r(f) - 2L_{\text{row}} \lambda_v.$$

Therefore, if Eq. A.16 holds, every pairwise margin between $y^\star$ and a competing class remains positive, and the predicted class is unchanged.

Finally, for a target class y and a competing class r,

$$\gamma_{y,r}(f') - \gamma_{y,r}(f) = (w_y - w_r)^\top (f' - f) = \lambda_v (w_y - w_r)^\top d.$$

Hence, if $\lambda_v > 0$ and $(w_y - w_r)^\top d > 0$, then

$$\gamma_{y,r}(f') - \gamma_{y,r}(f) > 0,$$

which proves Eq. A.17. This completes the proof.

A.6 Theorem 2. Effect of Guided Displacement on Pairwise Feature Separation

Claim. For two samples x_i and x_j, define $\delta_{ij} = f(x_i) - f(x_j), \Delta d_{ij} = d(x_i) - d(x_j)$. For $\lambda_v \geq 0$, the pairwise squared distance after displacement satisfies

$$\|f'(x_i) - f'(x_j)\|_2^2 = \|\delta_{ij}\|_2^2 + 2\lambda_v \langle \delta_{ij}, \Delta d_{ij} \rangle + \lambda_v^2 \|\Delta d_{ij}\|_2^2. \tag{A.18}$$

For a cross-class pair, if $\langle \delta_{ij}, \Delta d_{ij} \rangle \geq \rho_{\text{out}} > 0$, then

$$\|f'(x_i) - f'(x_j)\|_2^2 \geq \|f(x_i) - f(x_j)\|_2^2 + 2\lambda_v \rho_{\text{out}}. \tag{A.19}$$

In particular, if $\lambda_v > 0$, the pairwise squared distance strictly increases.

For a same-class pair, if the displacement directions are locally coherent such that $\|\Delta d_{ij}\|_2 \leq \rho_{\text{in}}$, then

$$\|f'(x_i) - f'(x_j)\|_2 \leq \|f(x_i) - f(x_j)\|_2 + \lambda_v \rho_{\text{in}}. \tag{A.20}$$

Proof. From the displacement rule,

$$f'(x) = f(x) + \lambda_v d(x).$$

Therefore,

$$f'(x_i) - f'(x_j) = f(x_i) - f(x_j) + \lambda_v \big(d(x_i) - d(x_j)\big) = \delta_{ij} + \lambda_v \Delta d_{ij}.$$

Expanding the squared norm gives

$$\|f'(x_i) - f'(x_j)\|_2^2 = \|\delta_{ij} + \lambda_v \Delta d_{ij}\|_2^2 = \|\delta_{ij}\|_2^2 + 2\lambda_v \langle \delta_{ij}, \Delta d_{ij} \rangle + \lambda_v^2 \|\Delta d_{ij}\|_2^2,$$

which proves Eq. A.18.

For a cross-class pair satisfying $\langle \delta_{ij}, \Delta d_{ij} \rangle \geq \rho_{\text{out}} > 0$, the final term in Eq. A.18 is nonnegative. Hence,

$$\|f'(x_i) - f'(x_j)\|_2^2 \geq \|\delta_{ij}\|_2^2 + 2\lambda_v \rho_{\text{out}},$$

which proves Eq. A.19. When $\lambda_v > 0$, the added term is strictly positive, so the squared distance between the two samples increases.

For a same-class pair, the triangle inequality gives

$$\|f'(x_i) - f'(x_j)\|_2 = \|\delta_{ij} + \lambda_v \Delta d_{ij}\|_2 \leq \|\delta_{ij}\|_2 + \lambda_v \|\Delta d_{ij}\|_2.$$

If $\|\Delta d_{ij}\|_2 \leq \rho_{\text{in}}$, then

$$\|f'(x_i) - f'(x_j)\|_2 \leq \|f(x_i) - f(x_j)\|_2 + \lambda_v \rho_{\text{in}},$$

which proves Eq. A.20.

Thus, under these sufficient alignment conditions, guided displacement can increase cross-class separation while controlling distortion among same-class samples. This completes the proof.

B Additional Details

B.1 Additional Dataset Details

We provide additional details of the datasets used in our experiments. The standard GCD benchmarks consist of four fine-grained image classification datasets and two generic image classification datasets. The DG-GCD benchmarks are further used to evaluate the cross-domain generalization ability of our method. The statistics of these datasets are summarized in Tables B.1 and B.2.

Fine-grained Datasets. We use four fine-grained datasets, including CUB (Welinder et al., 2010), Stanford Cars (Krause et al., 2013), FGVC-Aircraft (Maji et al., 2013), and Herbarium19 (Tan et al., 2019). CUB (Welinder et al., 2010) is a fine-grained bird classification dataset containing 200 bird species, and is widely used for evaluating models on subtle visual differences among closely related categories. Stanford Cars (Krause et al., 2013) contains 196 car categories, covering different makes, models, and years, and focuses on fine-grained vehicle recognition. FGVC-Aircraft (Maji et al., 2013) is designed for fine-grained aircraft classification, containing 100 aircraft model variants with images labeled according to their corresponding aircraft types. Herbarium19 (Tan et al., 2019) is a large-scale fine-grained dataset for plant species recognition, consisting of herbarium specimen images and involving a substantially larger number of classes than the other fine-grained benchmarks.

Generic Datasets. We use two generic image classification datasets, including CIFAR-100 (Krizhevsky et al., 2009) and ImageNet-100 (Deng et al., 2009). CIFAR-100 (Krizhevsky et al., 2009) is a widely used natural image dataset consisting of 100 classes, with 600 images per class. All images are resized to 32 × 32 pixels. ImageNet-100 (Deng et al., 2009) is a commonly adopted subset of ImageNet-1K, containing 100 randomly selected classes from the original ImageNet dataset.

Domain-Generalized GCD Datasets. To further evaluate the robustness of our method under domain shifts, we use three standard DG-GCD benchmarks, including PACS (Li et al., 2017), Office-Home (Venkateswara et al., 2017), and DomainNet (Peng et al., 2019). PACS (Li et al., 2017) contains four domains, namely `Photo`, `Art Painting`, `Cartoon`, and `Sketch`, covering seven object categories with significant visual style variations. Office-Home (Venkateswara et al., 2017) is a larger domain generalization benchmark with four domains, including `Art`, `Clipart`, `Product`, and `Real-World`, and contains 65 object categories commonly observed in office and home environments. DomainNet (Peng et al., 2019) is a large-scale multi-domain dataset containing six domains, including `Clipart`, `Infograph`, `Painting`, `Quickdraw`, `Real`, and `Sketch`, with 345 object categories.

Table B.1: Statistics of datasets used in the standard GCD setting. The column "FG" indicates whether the dataset is fine-grained. $|\mathcal{Y}_l|$, $|\mathcal{Y}_u|$ represents the labelled and unlabelled sets. $|\mathcal{D}_l|$, $|\mathcal{D}_u|$ represents the counts of images.

Dataset	FG	$\lvert\mathcal{D}_l\rvert$	$\lvert\mathcal{Y}_l\rvert$	$\lvert\mathcal{D}_u\rvert$	$\lvert\mathcal{Y}_u\rvert$
CUB (Welinder et al., 2010)	✓	1.5K	100	4.5K	200
Stanford Cars (Krause et al., 2013)	✓	2.0K	98	6.1K	196
FGVC-Aircraft (Maji et al., 2013)	✓	1.7K	50	5.0K	100
Herbarium19 (Tan et al., 2019)	✓	8.9K	341	25.4K	683
CIFAR-100 (Krizhevsky et al., 2009)	✗	20.0K	80	30.0K	100
ImageNet-100 (Deng et al., 2009)	✗	31.9K	50	95.3K	100

Table B.2: Statistics of datasets used in the DG-GCD setting. K/N denotes the known-to-novel class ratio. The source-target partitioning uses domain rotation (each domain takes turns acting as the source).

Dataset	Domains	Images	Classes	K/N Ratio
PACS (Li et al., 2017)	4	9,991	7	4:3
Office-Home (Venkateswara et al., 2017)	4	15,588	65	40:25
DomainNet (Peng et al., 2019)	6	586,575	345	250:95

B.2 Additional Implementation Details

For the standard GCD setting, we follow the commonly used class-split protocol (Vaze et al., 2022). Specifically, for all fine-grained datasets and ImageNet-100, 50% of the classes are treated as known classes and the remaining classes as novel classes, while for CIFAR-100, 80 classes are used as known classes. Within the known classes, 50% of the samples are labeled to form the labeled set, and the remaining known-class samples are combined with all novel-class samples to construct the unlabeled set.

For the DG-GCD setting, we follow the established known/novel splits (Rathore et al., 2025a), where the known-to-novel class ratios are 4/3 for PACS, 40/25 for Office-Home, and 250/95 for DomainNet. We adopt a leave-one-domain-as-source protocol (Rathore et al., 2025a;b), where each domain is selected as the source domain once and the remaining domains are used as target domains. Specifically for DomainNet, closely following the DG setting (Rathore et al., 2025a), we evaluate a subset of source-target configurations with `Sketch`, `Painting`, and `Clipart` as source domains, and test each source on the remaining selected domains. The final performance is averaged over all source-target combinations. For each source-target configuration, only labeled samples from the

known classes in the source domain are used for training, while the target domain is completely inaccessible during training and is used only for evaluation. At test time, all target-domain samples are treated as unlabeled, and the model is required to cluster both known and novel categories. In practice, all our experiments are run on a single H100.

B.3 Hyperparameter Settings

We summarize the default hyperparameter configurations of the **OADisp** module, which are consistently applied across all standard GCD (Vaze et al., 2022) and DG-GCD (Rathore et al., 2025a) benchmarks to ensure a fair and stable evaluation of the learned representations. Specifically, we set the number of latent slots to $N_s = 5$ and the adaptation strength to $\lambda_v = 0.05$ in Eq.1. For the joint optimization objective defined in Eq.9, we set the frequency-domain perturbation weight to $\lambda_{\text{fpc}} = 5$, the reconstruction loss weight to $\lambda_{\text{rec}} = 1$, and the diversity regularization weight to $\lambda_{\text{div}} = 0.1$.

For frequency-domain perturbation, we set $\alpha = 0.5$ and independently sample the low-frequency mask ratio $r \sim \mathcal{U}(0.03, 0.06)$ for each image. The mask covers a centered $2b \times 2b$ region in the FFT-shifted spectrum, where $b = \lfloor r \min(H, W) \rfloor$. Gaussian noise is generated as $\xi = \sigma_{\text{fd}} \mu_A Z$, with $\sigma_{\text{fd}} = 0.1$, $Z \sim \mathcal{N}(0, 1)$, and μ_A denoting the mini-batch mean amplitude. In practice, we clamp the perturbed amplitude to non-negative values before inverse Fourier transformation.

C Implementation of Slot Attention

Given the patch features $F = \{f_p\}_{p=1}^{P} \in \mathbb{R}^{P \times d}$, Slot Attention module (Locatello et al., 2020) iteratively aggregates information from the patch set into N_s latent slots. At iteration t, the k-th slot is updated as:

$$\alpha_{k,p}^{(t)} = \text{softmax}_k \left(\frac{(s_k^{(t)} W_q)(f_p W_k)^\top}{\sqrt{d}} \right), \qquad a_{k,p}^{(t)} = \frac{\alpha_{k,p}^{(t)}}{\sum_{q=1}^{P} \alpha_{k,q}^{(t)}},$$
$$s_k^{(t+1)} = \text{GRU} \left(\sum_{p=1}^{P} a_{k,p}^{(t)} f_p W_v,\ s_k^{(t)} \right). \tag{C.1}$$

where $s_k^{(t)} \in \mathbb{R}^{d_s}$ denotes the state of the k-th slot at iteration t, and W_q, W_k, and W_v are learnable projections. After T iterations, we obtain the slot representations $S = \{s_k\}_{k=1}^{N_s}$. In practice, we set $N_s = 5$ and $T = 3$ by default (Seitzer et al., 2023).

D Additional Experimental Result

Table D.1: Comparison with APL on four fine-grained datasets under standard GCD .

Methods	CUB			Stanford Cars			FGVC-Aircraft			Herbarium-19		
	All	Known	Novel	All	Known	Novel	All	Known	Novel	All	Known	Novel
SimGCD (Wen et al., 2023)	60.6	64.4	58.6	52.8	68.6	45.1	54.5	61.5	51.1	43.4	57.1	36.1
Δ (+APL) (Dai et al., 2025)	+3.2	+2.0	+4.1	+4.5	+6.6	+3.6	+1.7	-1.5	+3.3	+1.3	+1.1	+1.2
Δ (+**OADisp**) (Ours)	+4.9	+6.6	+4.1	+5.2	+7.8	+4.7	+3.4	+2.4	+3.8	+1.9	+1.5	+2.0

Table D.2: Comparison with APL on three datasets under DG-GCD.

Methods	PACS			Office-Home			DomainNet		
	All	Known	Novel	All	Known	Novel	All	Known	Novel
HIDISC (Rathore et al., 2025b)	75.9	77.8	73.2	58.1	60.8	53.4	30.4	31.1	28.1
Δ (+APL) (Dai et al., 2025)	-3.3	-6.1	+1.0	-2.1	-3.0	-0.5	-0.1	-0.1	+0.2
Δ (+**OADisp**) (Ours)	+1.1	-1.4	+4.8	+1.0	+1.1	+0.7	+1.0	+0.7	+1.8

Tables D.1 and D.2 compare our proposed **OADisp** with APL (Dai et al., 2025), a related plug-and-play method that directly exploits patch features. Under standard GCD, our approach yields larger *All* improvements on all four fine-grained datasets. When directly transferred to DG-GCD without

Table E.1: Ablation study of λ_v.

λ_v	All	Known	Novel
0	60.6	64.4	58.6
0.01	63.9	72.8	59.4
0.05	65.5	71.0	62.7
0.1	64.4	72.3	60.5
1.0	56.7	64.2	53.0

Table E.2: Ablation study of λ_{rec} and λ_{div}.

λ_{rec}	λ_{div}	All	Known	Novel
1.0	0.1	65.5	71.0	62.7
1.0	0.05	64.5	68.4	62.5
1.0	0.2	64.7	71.7	61.3
0.5	0.1	64.3	69.5	61.4
2.0	0.1	64.0	69.1	61.5

additional domain-specific modification, APL decreases *All* accuracy on PACS, Office-Home, and DomainNet by 3.3, 2.1, and 0.1 points, respectively, whereas **OADisp** improves them by 1.1, 1.0, and 1.0 points and consistently benefits Novel classes. These results suggest that complementing the holistic [CLS] representation with structured patch-derived guidance is more effective and robust than replacing it with patch features, particularly under domain shifts.

E Ablation Study

E.1 Displacement Strength Sensitivity

We evaluate the sensitivity of **OADisp** to the displacement strength parameter λ_v, which controls how far the global [CLS] feature is shifted toward the image-conditioned guidance target. As shown in Table E.1, the performance on the CUB dataset follows a bell-shaped trend as λ_v increases. Setting $\lambda_v = 0$ corresponds to the model is baseline. We observe that introducing a small amount of guidance consistently improves the clustering accuracy, with the peak performance of 65.5% *All* and 62.7% *Novel* achieved at $\lambda_v = 0.05$. However, increasing λ_v beyond this range gradually degrades performance, with a substantial drop at $\lambda_v = 1.0$. This degradation occurs because an excessively large λ_v forces the global feature to deviate too far from its original holistic summary, which may distort useful global information encoded in the original [CLS] representation. This suggests that moderate displacement provides a favorable balance between preserving the original holistic [CLS] representation and incorporating object-aware local semantics, whereas overly aggressive displacement may distort useful global information. Consequently, we fix $\lambda_v = 0.05$ for all experiments.

E.2 Loss Weight Sensitivity.

We analyze the sensitivity to the reconstruction and diversity loss weights, λ_{rec} and λ_{div}, on the CUB dataset. Following a one-at-a-time protocol, we vary one weight while keeping the other fixed at its default value. The results are summarized in Table E.2. The performance remains reasonably stable under moderate changes to both regularization weights. Varying λ_{div} from 0.05 to 0.2 changes the *All* accuracy by only 0.8–1.0 points relative to the default setting. The reconstruction weight has a slightly larger influence, which is expected since $\mathcal{L}_{\text{rec}}$ directly grounds the learned slots and masks in the original patch feature map. Nevertheless, changing λ_{rec} by a factor of two results in only a 1.2–1.5 point decrease in *All* accuracy. Importantly, we use the same default weights, $\lambda_{\text{rec}} = 1.0$ and $\lambda_{\text{div}} = 0.1$, across all nine datasets without dataset-specific tuning. Together, these results indicate that the object-aware descriptor construction does not require extensive hyperparameter tuning across different datasets.

E.3 Ablation of Descriptor Construction and Target Aggregation

We further analyze the two design choices used to form the target for guided [CLS] displacement: descriptor construction and target aggregation in Table E.3. All variants retain the same rule in Eq. 1 and Section 3.3, while varying the descriptor set and the strategy used to construct the image-conditioned target. For *Raw Patch Features*, we directly treat the native patch tokens as the descriptor set. For *Object-Aware Descriptors*, we use the mask-guided descriptors constructed in Eq. 6. Regarding target aggregation, *Uniform Mean* assigns equal weights to all descriptors (mean pooling), whereas *[CLS]-Conditioned* uses the compatibility-based weighting in Eqs. 2–3.

Table E.3: Ablation study of Descriptor Construction and Target Aggregation.

	Descriptor Construction	Target Aggregation	All	Known	Novel
#1	Raw Patch Features	Uniform Mean	61.4	68.2	58.0
#2	Raw Patch Features	[CLS]-Conditioned	62.5	69.1	59.2
#3	Object-Aware Descriptors	Uniform Mean	63.0	68.8	60.1
#4	Object-Aware Descriptors	[CLS]-Conditioned	**65.5**	**71.0**	**62.7**

The results show that both object-aware descriptor construction and [CLS]-conditioned aggregation contribute to the performance. Comparing #1 with #3, replacing raw patch features with object-aware descriptors under uniform aggregation improves *All* from 61.4% to 63.0% (+1.6%) and *Novel* from 58.0% to 60.1% (+2.1%). Under [CLS]-conditioned aggregation, the same replacement from #2 to #4 yields larger gains of +3.0% in *All* and +3.5% in *Novel*, further showing that object-aware descriptors provide more effective local semantic guidance than fragmented raw patches. Meanwhile, [CLS]-conditioned aggregation consistently outperforms uniform averaging: #2 improves over #1 by +1.1% in *All*, while #4 improves over #3 by +2.5%. Notably, the larger gain in the object-aware setting (+2.5% vs. +1.1%) suggests that compatibility-based target aggregation can better exploit semantically coherent descriptors, highlighting the complementary interaction between descriptor construction and displacement guidance.

F Visualization Results

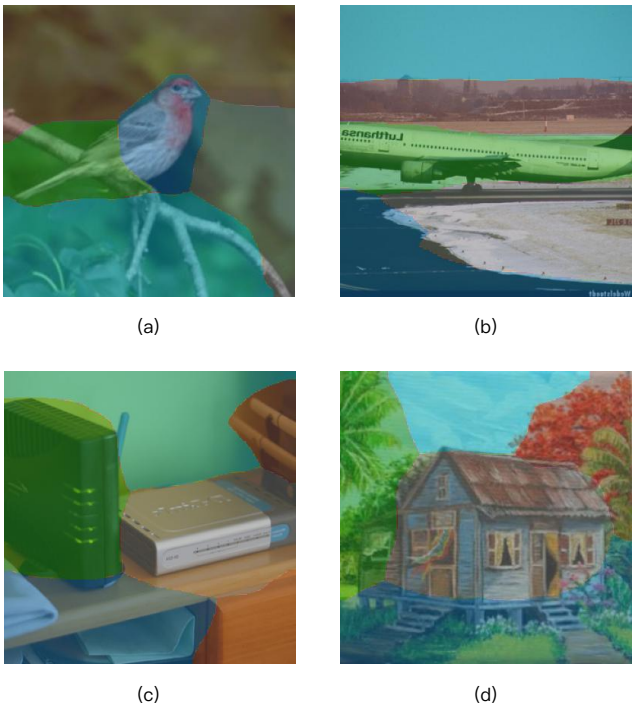

Figure F.1: Visualization of object-centric decomposition.

To qualitatively validate the object-centric decomposition used in **OADisp**, we visualize the learned spatial masks across four diverse datasets: (a) CUB (Welinder et al., 2010), (b) FGVC-Aircraft (Maji et al., 2013), (c) ImageNet-100 (Deng et al., 2009), and (d) PACS (Li et al., 2017). As shown in Figure F.1, the latent slots decompose complex scenes into semantically coherent components without explicit object-level supervision. Across fine-grained, generic, and non-photorealistic domains with distribution shifts, the learned masks consistently capture category-relevant foreground regions while separating substantial background content. These results support the use of the learned masks for

aggregating semantically related patch features, enabling the construction of effective object-aware semantic descriptors.