SPARC: LEARNING SPECTRALLY STABLE ASSIGNMENTS FOR GENERALIZED CATEGORY DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Generalized Category Discovery (GCD) aims to recognize known categories and discover novel categories in unlabeled data. Despite recent progress driven by strong representation learning, GCD models may rely on appearance cues such as texture, illumination, and background. These cues can distort feature similarity and lead to unstable category assignments. Existing methods mainly enforce consistency between spatially augmented views, while the effect of frequency-domain appearance changes on category assignments remains underexplored. Moreover, prior Fourier-based methods typically use frequency operations for augmentation or feature enhancement, rather than explicitly improving assignment stability. To address this issue, we propose **SPARC**, a plug-and-play module for **SP**ectral **A**ssignment **R**egularization for **C**ategories. SPARC contains two components. Image-level Consistency Regularization (ICR) mixes low-frequency amplitudes from different images while retaining the source phase, and encourages the mixed image to preserve the assignment of the clean image. Token-level Consistency Regularization (TCR) identifies patch tokens that remain stable after spectral filtering, pools them into a Stable Semantic Token (SST), and aligns its assignment with that of the global descriptor. For parametric hosts, SST is additionally used as the readout during inference. Experiments on representative GCD baselines with DINO and DINOv2 demonstrate performance improvements across CIFAR-10/100, ImageNet-100, CUB, Stanford-Cars, and FGVC-Aircraft.

1 INTRODUCTION

Once framed as Novel Category Discovery (NCD) and later broadened to Generalized Category Discovery (GCD) (Vaze et al., 2022), category discovery has become an important open-world problem. GCD aims to recognize known categories while simultaneously discovering novel categories from unlabeled data. Recent advances in GCD have shown that strong visual representations (Caron et al., 2021; Oquab et al., 2024), combined with contrastive learning (Vaze et al., 2022; Rastegar et al., 2024) and prototype-based clustering (Wen et al., 2023; Liu & Han, 2025), can effectively transfer semantic knowledge from known classes to unseen ones. However, the quality of category discovery remains fundamentally dependent on the stability of the learned semantic structure. In practice, visual representations can encode appearance cues that are not always aligned with category semantics. For example, prior studies have documented texture bias in image classifiers and the influence of image backgrounds on object recognition (Geirhos et al., 2019; Xiao et al., 2021). When such cues dominate feature similarity, they may also affect category discovery: samples from different categories may be grouped together because they share similar appearance, whereas samples from the same category may be separated under changes in texture, illumination, or background.

To alleviate these issues, GCD methods commonly learn view-consistent representations using *spatial-domain* augmentations such as cropping, flipping, and color jittering. This applies to both *non-parametric* approaches, which infer novel categories from feature similarities (Vaze et al., 2022; Rastegar et al., 2024), and *parametric* approaches, which learn category assignments through trainable prototypes or classifiers (Wen et al., 2023; Wang et al., 2024). Although their assignment mechanisms differ, both largely establish consistency across spatially augmented views without explicitly accounting for changes in spectral statistics. As a result, their representations and category assignments may remain sensitive to non-semantic spectral cues (see Fig. 1(a)).

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.

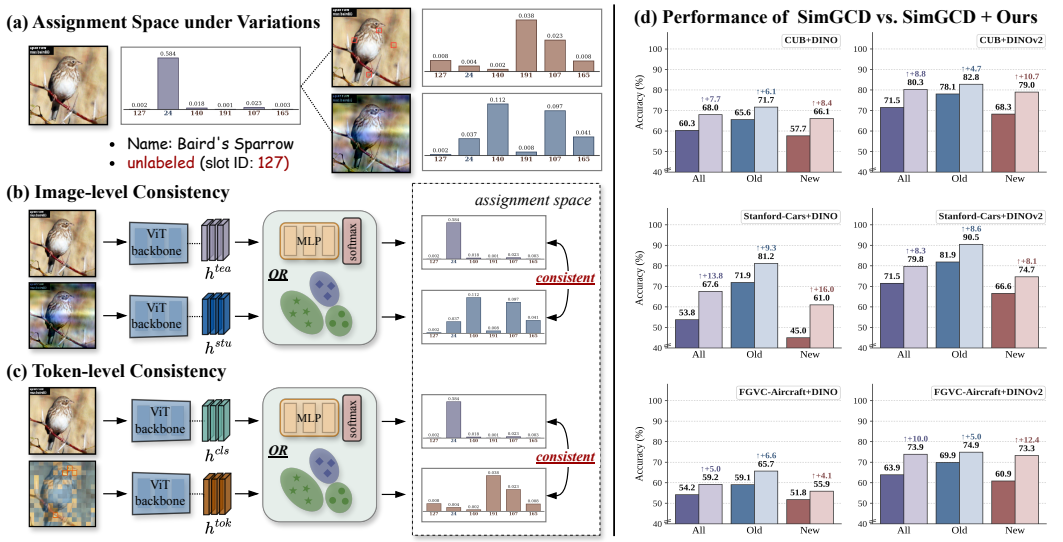

Figure 1: **(a)** Top-5 predicted slots of a sample image and the corresponding variations of a trained SimGCD (Wen et al., 2023) model. **(b)** Image-level consistency in the assignment space. **(c)** Token-level consistency in the assignment space. **(d)** Comparison of *All*-, *Old*-, and *New*-class *ACC* on CUB, Stanford-Cars, and FGVC-Aircraft with DINO and DINOv2 backbones. Results are reported for SimGCD versus SimGCD+**SPARC**.

Recently, several studies have explored *frequency-domain* learning for category discovery under domain-shifted (Feng & Ge, 2025) and continual (Feng et al., 2026) settings. Existing approaches have demonstrated that spectral information can improve robustness to domain shifts or enhance fine-grained discrimination by manipulating low- and high-frequency components. However, these prior works primarily treat Fourier operations (Oppenheim & Lim, 1981) as tools for feature enhancement, domain adaptation, or frequency-based augmentation. Their focus remains largely on improving representation quality in the feature space. In contrast, the central objective of GCD is not merely to learn discriminative representations, but to establish reliable semantic assignments for clustering and category discovery.

From this perspective, we argue that *semantic assignments themselves should remain stable under semantic-preserving spectral perturbations*. Specifically, frequency manipulations that alter appearance statistics while preserving semantic structure should not significantly change the model's category assignment distribution. This observation motivates a new form of regularization for GCD: spectral assignment consistency (see Fig. 1(b)). Instead of enforcing direct feature alignment between clean and transformed views, we encourage consistency in the prototype assignment space, which is more closely aligned with the clustering objective of GCD.

Moreover, existing GCD frameworks are predominantly built upon self-supervised Vision Transformer backbones pretrained with DINO (Caron et al., 2021) or DINOv2 (Oquab et al., 2024), whose patch-token representations naturally preserve rich spatial semantic structures. Importantly, not all patch tokens exhibit the same degree of robustness under spectral perturbations. Tokens corresponding to semantic object structures tend to remain stable, whereas tokens dominated by texture or background artifacts often fluctuate significantly. This suggests that spectral stability itself can serve as a cue for identifying reliable local semantics (see Fig. 1(c)).

Motivated by these observations, we propose **SPARC**, a plug-and-play module for **S**pectral **A**ssignment **R**egularization for **C**ategories. SPARC regularizes the host in assignment space through two complementary consistency objectives without modifying the host backbone; for *parametric* hosts, SST replaces the CLS readout during training and inference.

Image-level Consistency Regularization (ICR) mixes low-frequency amplitude while preserving the source phase, and requires the mixed view to keep the clean image's prototype assignment. Token-level Consistency Regularization (TCR) pools spectrally stable patch tokens into a local descriptor, the Stable Semantic Token (SST), and aligns it with the global descriptor in the same assignment space. On a *parametric* host, this SST replaces the CLS readout in training and evaluation.

Extensive experiments show that **SPARC** consistently improves category discovery across multiple GCD frameworks and datasets, yielding more reliable assignments under spectral appearance changes (see Fig. 1(d)).

Our contributions are summarized as follows: **(i)** We identify spectral assignment consistency as a previously overlooked property in GCD: semantic assignments should remain stable under semantic-preserving spectral perturbations, yet existing methods do not regularize this stability explicitly. **(ii)** We propose **SPARC**, a plug-and-play module that regularizes prototype assignments under spectral perturbations without altering the host backbone, while using SST as the inference readout for parametric hosts. **(iii)** Within **SPARC**, Image-level Consistency Regularization (ICR) enforces assignment invariance under low-frequency amplitude mixing, and Token-level Consistency Regularization (TCR) aligns the global assignment with a Stable Semantic Token (SST) pooled from spectrally stable patches. **(iv)** Extensive experiments across multiple datasets and representative *parametric* and *non-parametric* GCD baselines demonstrate consistent improvements in clustering quality and novel-category discovery.

2 Related Work

Generalized Category Discovery (GCD). Category discovery was first studied as Novel Category Discovery (NCD) (Han et al., 2019), which transfers knowledge from labeled seen classes to cluster unlabeled data assumed to contain only novel classes. GCD (Vaze et al., 2022) relaxes this by mixing known and novel categories in the unlabeled pool. Most methods start from a strong backbone, extract a global descriptor, and perform category assignment in the resulting embedding space. Within this assignment stage, existing methods fall into several groups. *Non-parametric* approaches, starting from the original GCD pipeline, were refined by SelEx (Rastegar et al., 2024) via hierarchical semi-supervised k-means. *Parametric* methods replace clustering with a jointly trained classifier: SimGCD (Wen et al., 2023) combines contrastive learning with entropy regularization, later improved by DebGCD (Liu & Han, 2025), SEAL (He et al., 2025), etc. Complementary to global objectives, several works redesign the views used in the GCD task. MOS (Peng et al., 2025) constructs an object-centric view via saliency-based foreground–background separation and jointly learns from the original image and the object-only crop to model object–scene associations. HiLo (Wang et al., 2025b) adapts PatchMix in the embedding space as a domain-aware augmentation tailored to GCD under distribution shift. Another line intervenes at the token or part level. SPTNet (Wang et al., 2024) introduces spatial prompt tuning so that the model attends to discriminative object parts. APL (Dai et al., 2025) employs shared learnable part queries with DINO priors and an all-min contrastive loss to discover consistent object parts. AF (Xu et al., 2025) prunes non-informative tokens via multi-scale importance measurement and adaptive pruning to combat distracted attention. TAR (Yang et al., 2026) suppresses high-norm attention artifacts and refines the token sequence to recover local cues. These approaches demonstrate the value of stable local or multi-granularity evidence, yet their operations remain predominantly spatial-domain or token-sequence manipulations within the Vision Transformer.

Fourier-Domain Learning. The Fourier transform separates appearance statistics from structural content. Classical signal analyses show that phase often carries substantial structural information for image reconstruction, whereas amplitude is more closely tied to style-like cues (Oppenheim & Lim, 1981). A related robustness literature further suggests that models often rely on frequency-specific shortcuts: e.g., texture- or background-biased decisions (Geirhos et al., 2019; Xiao et al., 2021), and sensitivity to particular frequency bands under corruption or adversarial perturbation (Yin et al., 2019). Building on amplitude–phase decoupling, FDA (Yang & Soatto, 2020) swaps low-frequency amplitude across domains for adaptation, noting that amplitude largely encodes style while phase preserves structure. Xu et al. (2021) extend this idea with amplitude mixing and consistency between original and Fourier-augmented predictions. Subsequent FDA-style operators diversify spectral appearance while aiming to preserve semantics: frequency-space domain randomization (Huang et al., 2021), amplitude–phase recombination (Chen et al., 2021) and so on. More recently, this idea has been introduced into open-world discovery settings. FREE (Feng & Ge, 2025) uses amplitude differences for domain separation and amplitude perturbations for robustness under domain-shifted GCD, while PRISM (Feng et al., 2026) emphasizes high-frequency cues to separate known from novel categories in continual discovery. Across these lines, Fourier processing mainly improves domain robustness, representation quality, or frequency-aware augmentation. In contrast, GCD hinges on the stability of category assignments: even with unchanged object layout, frequency-dependent

appearance shifts can still flip prototype membership. This motivates our view of Fourier transforms not merely as feature enhancers but as controlled semantic-preserving perturbations for assignment-space regularization.

3 PRELIMINARIES

Problem Setting. Let $\mathcal{D}_l = \{(x_i, y_i)\}$ be labeled samples from known categories $\mathcal{Y}_l$, and $\mathcal{D}_u = \{x_i\}$ unlabeled samples from $\mathcal{Y}_l \cup \mathcal{Y}_n$ with $\mathcal{Y}_n \cap \mathcal{Y}_l = \emptyset$. GCD aims to recognize unlabeled samples from $\mathcal{Y}_l$ and to cluster those from $\mathcal{Y}_n$. In a mini-batch $\mathcal{B} = \mathcal{B}_l \cup \mathcal{B}_u$, supervised terms use $\mathcal{B}_l$ and unsupervised clustering uses all of $\mathcal{B}$. We attach our module to an existing GCD method (the *host*), which provides a ViT f_θ, an objective $\mathcal{L}_{\text{host}}$, and an assignment bank $P = \{p_k\}_{k=1}^K \subset \mathbb{R}^d$. The backbone returns $f_\theta(x) = (h^{\text{cls}}(x), \mathbf{T}(x))$, with CLS descriptor $h^{\text{cls}}(x) \in \mathbb{R}^d$ and patch tokens $\mathbf{T}(x) = [\mathbf{t}_1; \ldots; \mathbf{t}_N] \in \mathbb{R}^{N \times d}$ ($N{=}G_h G_w$). P is a set of learnable prototypes or clustering centers for a GCD host.

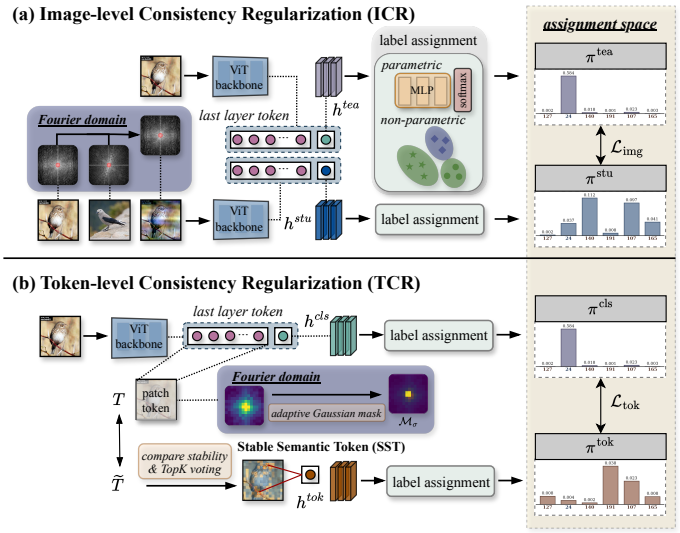

Figure 2: **SPARC** overview. **(a)** Image-level Consistency Regularization (ICR) and corresponding $\mathcal{L}_{\text{img}}$. **(b)** Token-level Consistency Regularization (TCR) with Stable Semantic Token (SST) and corresponding $\mathcal{L}_{\text{tok}}$.

Soft Assignment. Category discovery ultimately produces a soft assignment over K categories. For a descriptor $h \in \mathbb{R}^d$,

$$z_k(h; P) = \frac{h^\top p_k}{\|h\|_2 \|p_k\|_2}, \qquad \pi(h; P, \tau) = \text{softmax}\big(z(h; P)/\tau\big) \in \mathbb{R}^K. \tag{1}$$

Given a teacher–student pair $(h^{\text{tea}}, h^{\text{stu}})$, disagreement in assignment space is measured by

$$\mathcal{L}_{\text{KL}}(h^{\text{tea}}, h^{\text{stu}}; P) = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \text{KL}\Big(\text{sg}\left[\pi(h_i^{\text{tea}}; P, \tau_{\text{tea}})\right] \Big\| \pi(h_i^{\text{stu}}; P, \tau_{\text{stu}}) \Big), \tag{2}$$

where $\text{sg}[\cdot]$ freezes the teacher assignment. Both spectral losses below are instances of Eq. 2: they regularize assignments rather than features.

Spectral Operators. For a real tensor $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$, let $\widehat{\mathbf{X}} = \mathcal{F}(\mathbf{X})$ denote its per-channel DFT, with amplitude $\mathbf{A} = |\widehat{\mathbf{X}}|$ and phase $\mathbf{\Phi} = \angle \widehat{\mathbf{X}}$. On the Nyquist-normalized radius $r_{u,v} \in [0, 1]$ from DC, the Gaussian low-pass $\mathcal{M}_\sigma(u, v) = \exp(-r_{u,v}^2/(2\sigma^2))$ yields

$$\widetilde{\mathbf{X}} = \mathcal{F}^{-1}\big(\mathcal{M}_\sigma \odot \mathbf{A} \odot e^{\mathrm{i}\mathbf{\Phi}}\big). \tag{3}$$

Amplitude is attenuated while phase is left unchanged. ICR applies a real FFT to RGB images; TCR applies a shifted complex FFT to patch-token grids (Appendix A).

4 METHODOLOGY

SPARC regularizes the host assignment under spectral perturbations without changing the backbone or the host's spatial views (Fig. 2). It comprises three complementary ingredients. Image-level Consistency Regularization (ICR) mixes low-frequency amplitude and requires the mixed assignment to match that of the clean image (Sec. 4.1). Token-level Consistency Regularization (TCR) pools spectrally stable patches into a local descriptor and aligns it with the global descriptor in assignment space (Sec. 4.2). The Stable Semantic Token (SST) is a readout, not a third loss: on a parametric

host it replaces h^{cls} with this local descriptor in $\mathcal{L}_{\mathrm{host}}$ and at evaluation (Sec. 4.3). Both regularizers are instances of Eq. 2: they penalize disagreement in $\pi(\cdot; P)$ rather than in feature space, so the objective stays aligned with the clustering criterion used at evaluation.

4.1 Image-level Consistency Regularization via Amplitude Mixing

ICR keeps $\pi(\cdot; P)$ fixed when low-frequency appearance changes but spatial structure does not. Let $A(x) = |\mathcal{F}(x)|$ and $\Phi(x) = \angle\mathcal{F}(x)$. On this radial coordinate we define a mixing window of bandwidth $\rho \in (0, 1]$,

$$\mathcal{W}_{u,v}(\rho) = \mathbb{I}\big[r_{u,v} \le \rho\big], \tag{4}$$

so only coefficients inside a low-frequency disk are eligible for mixing. A donor x' is a cyclic shift of the minibatch (no self-pair) and $\lambda \sim \mathrm{Unif}[0, 1]$ is drawn per sample.

$$\widetilde{A} = A(x) + \lambda\, \mathcal{W}(\rho) \odot \big(A(x') - A(x)\big), \qquad \widetilde{x} = \mathcal{F}^{-1}\big(\widetilde{A}\, e^{\mathrm{i}\Phi(x)}\big). \tag{5}$$

Phase from x preserves object layout; the mixed amplitude may transfer part of the target image's low-frequency appearance statistics. We normalize the mixed image $\widetilde{x}$ using the same channel-wise mean and standard deviation as the host inputs. The mixed image is read out as h^{tok} when SST is enabled and as h^{cls} otherwise; write h^{stu} for this student descriptor and h^{tea} for the corresponding clean readout. Taking the clean view as teacher and the mixed view as student,

$$\mathcal{L}_{\mathrm{img}} = \mathcal{L}_{\mathrm{KL}}(h^{\mathrm{tea}}, h^{\mathrm{stu}}; P). \tag{6}$$

$\widetilde{x}$ enters only $\mathcal{L}_{\mathrm{img}}$; gradients do not alter the host's spatial augmentation pipeline, and the bank P is shared with $\mathcal{L}_{\mathrm{host}}$.

4.2 Token-level Consistency Regularization via Spectral Pooling

TCR aligns the global assignment with a local descriptor formed from patches that survive a spectral low-pass. Reshape the patch tokens $\mathbf{T} \in \mathbb{R}^{N\times d}$ to a spatial grid $\mathbf{Z} \in \mathbb{R}^{d\times G_h\times G_w}$ ($N{=}G_hG_w$) and form $\widehat{\mathbf{Z}} = \mathrm{FFTShift}(\mathcal{F}(\mathbf{Z})) = \mathbf{A} \odot e^{\mathrm{i}\boldsymbol{\Phi}}$. Per-sample bandwidth is the energy-weighted mean radius

$$\bar{r} = \frac{\sum_{u,v} r_{u,v}\, \|\mathbf{A}_{:,u,v}\|_2^2}{\sum_{u,v} \|\mathbf{A}_{:,u,v}\|_2^2 + \epsilon}, \qquad \sigma = \kappa\, \bar{r}, \tag{7}$$

with scale $\kappa{>}0$; filtered tokens $\widetilde{\mathbf{T}}$ then follow Eq. 3 with mask $\mathcal{M}_\sigma$. A location is treated as spectrally stable when low-pass filtering barely moves its activation. For sample b, patch index n, and channel c,

$$S_{b,n,c} = \frac{|T_{b,n,c}|}{\max\big(|\widetilde{T}_{b,n,c} - T_{b,n,c}|, \epsilon\big)}, \qquad V_{b,n} = \sum_{c=1}^{d} \mathbb{I}\big[n \in \mathrm{TopK}_c(S_{b,:,c}; K_v)\big], \tag{8}$$

where $\mathrm{TopK}_c(\cdot; K_v)$ returns the K_v indices of largest score along the patch axis of channel c. Each channel therefore casts one vote for each of its most stable locations; $V_{b,n}$ aggregates these votes. The Stable Semantic Token (SST) averages the K_s patches with the largest votes,

$$h_b^{\mathrm{tok}} = \tfrac{1}{K_s} \sum_{n\in\mathcal{S}_b} \mathbf{t}_{b,n}, \qquad \mathcal{S}_b = \mathrm{TopK}_n(V_{b,n}, K_s). \tag{9}$$

Unless stated otherwise we set $K_v{=}K_s$ (shared TopK). $\mathcal{L}_{\mathrm{tok}}$ is Eq. 2 between h^{cls} and h^{tok}; the teacher is the descriptor used by $\mathcal{L}_{\mathrm{host}}$ (Sec. 4.3). This design encourages agreement with the host's selected readout.

4.3 Host Instantiation and Stable Semantic Token

SST selects the descriptor for $\mathcal{L}_{\mathrm{host}}$ and evaluation; both consistency teachers follow that choice. This avoids a train–test mismatch: the representation optimized by the host is the one scored at test time.

Parametric host (e.g., SimGCD (Wen et al., 2023)). With SST enabled, supervised CE, clustering, contrastive terms, and evaluation all use h^{tok}:

$$\mathcal{L}_{\mathrm{img}} = \mathcal{L}_{\mathrm{KL}}(h^{\mathrm{tok}}, h^{\mathrm{stu}}; P), \qquad \mathcal{L}_{\mathrm{tok}} = \mathcal{L}_{\mathrm{KL}}(h^{\mathrm{tok}}, h^{\mathrm{cls}}; P), \tag{10}$$

where h^{stu} is the amplitude-mixed view under the same readout. CLS remains only as the student of $\mathcal{L}_{\mathrm{tok}}$, which encourages the global token to agree with spectrally stable local evidence.

Non-parametric host (e.g., SelEx (Rastegar et al., 2024)). P is estimated in the embedding (e.g., spherical k-means), so replacing the readout would change the geometry that clustering depends on. $\mathcal{L}_{\text{host}}$ and evaluation therefore stay on h^{cls} (SST off), and the spectral terms remain auxiliary:

$$\mathcal{L}_{\text{img}} = \mathcal{L}_{\text{KL}}(h^{\text{cls}}, h^{\text{stu}}; P), \qquad \mathcal{L}_{\text{tok}} = \mathcal{L}_{\text{KL}}(h^{\text{cls}}, h^{\text{tok}}; P). \tag{11}$$

h^{tok} is still computed for TCR but does not replace the non-parametric host representation.

4.4 Overall Objective

The full training objective combines the host loss with the two spectral regularizers:

$$\mathcal{L} = \mathcal{L}_{\text{host}} + \lambda_{\text{img}}\gamma_{\text{img}}(t)\,\mathcal{L}_{\text{img}} + \lambda_{\text{tok}}\gamma_{\text{tok}}(t)\,\mathcal{L}_{\text{tok}}. \tag{12}$$

Supervised host terms use $\mathcal{B}_l$; unsupervised host terms and both spectral losses use $\mathcal{B}$. Each branch $b \in \{\text{img}, \text{tok}\}$ is introduced by $\gamma_b(t) = \min(1, t/W_b)$ over the first W_b epochs, so early estimates of P are not locked by full-strength auxiliaries. Either weight may be set to zero for ablation. At inference, ICR and $\mathcal{L}_{\text{tok}}$ are disabled; evaluation uses SST on SimGCD and CLS on SelEx.

5 Experiments

5.1 Implementation Details

We evaluate **SPARC** on two representative hosts: the parametric method SimGCD (Wen et al., 2023) and the non-parametric method SelEx (Rastegar et al., 2024). Datasets cover both generic recognition—CIFAR-10/100 (Krizhevsky et al., 2009) and ImageNet-100 (Deng et al., 2009)—and the fine-grained SSB suite—CUB (Wah et al., 2011), Stanford-Cars (Krause et al., 2013), and FGVC-Aircraft (Maji et al., 2013)—using the standard GCD splits introduced by Vaze et al. (2022). Unless noted, hosts are initialized from DINO (Caron et al., 2021) (ViT-B/16) or DINOv2 (Oquab et al., 2024) (ViT-B/14); results with other backbones are deferred to Table 6. We report clustering accuracy (*ACC*) on unlabeled images for all categories (*All*), known categories (*Old*), and novel categories (*New*), after Hungarian matching to ground-truth labels. **SPARC** is attached as training-time ICR/TCR auxiliaries without changing the host optimizer and original spatial augmentations. For the *parametric* SimGCD, SST replaces CLS readout in $\mathcal{L}_{\text{host}}$ and at evaluation; on *non-parametric* SelEx, SST is used only within TCR. Further hyperparameters, dataset statistics, and extended ablations are provided in the Appendix.

5.2 Quantitative Results

Tables 1 and 2 report *All*/*Old*/*New ACC*. Within each backbone block, **bold** / underline denote the best / second-best entry in that column.

Performance on Generic Datasets. Table 1 reports *All*/*Old*/*New ACC* on CIFAR-10 (Krizhevsky et al., 2009), CIFAR-100 (Krizhevsky et al., 2009), and ImageNet-100 (Deng et al., 2009). With DINO, SimGCD+**SPARC** raises mean *All*/*New* from 86.7%/84.6% to 88.5%/86.7% ($+1.8$/$+2.1$ percentage points; relative $+2.1\%$/$+2.5\%$), driven mainly by CIFAR-100 ($+2.8$/$+3.6$) and ImageNet-100 ($+1.7$/$+2.3$); SelEx+**SPARC** improves mean *All*/*New* by $+1.7$/$+2.4$, with the largest lift on ImageNet-100 ($+4.0$/$+4.8$). With DINOv2, where strong hosts already leave limited headroom, SimGCD+**SPARC** still reaches 93.8%/92.6% mean *All*/*New* ($+1.4$/$+1.3$), again helped by ImageNet-100 ($+2.8$/$+3.7$); SelEx+**SPARC** adds $+0.8$/$+1.5$, including a $+3.8$ *New* gain on CIFAR-100. Overall, spectral assignment regularization remains useful on generic data even when absolute margins are smaller than on fine-grained benchmarks.

Performance on Fine-grained Datasets. Table 2 reports the same protocol on the SSB suite: CUB (Wah et al., 2011), Stanford-Cars (Krause et al., 2013), and FGVC-Aircraft (Maji et al., 2013). Here both absolute and relative gains are larger. With DINO, SimGCD+**SPARC** lifts mean *All*/*New* from 56.1%/51.5% to 64.7%/60.6% ($+8.6$/$+9.1$; relative $+15.3\%$/$+17.7\%$), peaking on Stanford-Cars ($+13.8$/$+16.0$; relative $+25.7\%$/$+35.6\%$); SelEx+**SPARC** improves mean *All*/*New* by $+7.1$/$+9.9$, with strong *New* gains on CUB ($+8.3$) and Stanford-Cars ($+14.7$). With DINOv2, SimGCD+**SPARC** reaches 78.0%/75.7% mean *All*/*New* ($+9.0$/$+10.4$; relative $+13.0\%$/$+15.9\%$), including double-digit *New* gains on CUB ($+10.7$) and FGVC-Aircraft ($+12.4$); SelEx+**SPARC** contributes a further $+2.3$/$+2.7$ on average and attains the best mean *All*/*Old*/*New* in the DINOv2 block. Thus **SPARC** delivers consistent *All*/*New* gains on both backbones, showing that spectral assignment regularization is especially effective under fine-grained appearance variation.

Table 1: Clustering accuracy (*ACC*, %) on generic datasets for *All*/*Old*/*New*. The rightmost block is the mean over CIFAR-10, CIFAR-100, and ImageNet-100.

	Method[Venue]	CIFAR-10			CIFAR-100			ImageNet-100			Average		
		All	Old	New	All	Old	New	All	Old	New	All	Old	New
DINO	GCD (Vaze et al., 2022)[CVPR 2022]	91.5	97.9	88.2	73.0	76.2	66.5	74.1	89.8	66.3	79.5	88.0	73.7
	PromptCAL (Zhang et al., 2023)[CVPR 2023]	**97.9**	96.6	98.5	81.2	84.2	75.3	83.1	92.7	78.3	87.4	91.2	84.0
	GPC (Zhao et al., 2023)[ICCV 2023]	90.6	97.6	87.0	75.4	84.6	60.1	75.3	93.4	66.7	80.4	91.9	71.3
	InfoSieve (Rastegar et al., 2023)[NeurIPS 2023]	94.8	97.7	93.4	78.3	82.2	70.5	80.5	93.8	73.8	84.5	91.2	79.2
	SPTNet (Wang et al., 2024)[ICLR 2024]	97.3	95.0	98.6	81.3	84.3	75.6	85.4	93.2	81.4	88.0	90.8	85.2
	DebGCD (Liu & Han, 2025)[ICLR 2025]	97.2	94.8	98.4	83.0	84.6	79.9	85.9	94.3	81.6	88.7	91.2	86.6
	AF (Xu et al., 2025)[ICCV 2025]	97.8	95.9	**98.8**	82.2	85.0	76.5	85.4	94.6	80.8	88.5	91.8	85.4
	SEAL (He et al., 2025)[NeurIPS 2025]	97.2	94.7	98.4	82.1	81.7	83.0	84.6	90.9	81.3	88.0	89.1	87.6
	†CPF-GCD (Dai et al., 2026)[ICML 2026]	96.8	97.7	96.3	82.0	84.4	77.3	85.5	94.7	80.8	88.1	92.3	84.8
	CURE (Bian et al., 2026)[ICML 2026]	97.5	97.9	97.3	**85.7**	**85.3**	**86.6**	87.0	95.3	**82.9**	**90.1**	**92.8**	**88.9**
	SimGCD (Wen et al., 2023)[ICCV 2023]	97.1	95.1	98.1	80.1	81.2	77.8	83.0	93.1	77.9	86.7	89.8	84.6
	+ **SPARC**	97.8	96.4	98.4	82.9	83.6	81.4	84.7	93.8	80.2	88.5	91.3	86.7
	Inc. Δ	**+0.7**	**+1.3**	**+0.3**	**+2.8**	**+2.4**	**+3.6**	**+1.7**	**+0.7**	**+2.3**	**+1.8**	**+1.5**	**+2.1**
	SelEx (Rastegar et al., 2024)[ECCV 2024]	95.9	**98.1**	94.8	82.3	**85.3**	76.3	83.1	93.6	77.8	87.1	92.3	83.0
	+ **SPARC**	97.3	97.2	97.3	82.1	84.9	76.4	**87.1**	**96.2**	82.6	88.8	**92.8**	85.4
	Inc. Δ	**+1.4**	**-0.9**	**+2.5**	**-0.2**	**-0.4**	**+0.1**	**+4.0**	**+2.6**	**+4.8**	**+1.7**	**+0.5**	**+2.4**
DINOv2	GCD (Vaze et al., 2022)[CVPR 2022]	97.8	**99.0**	97.1	79.6	84.5	69.9	78.5	89.5	73.0	85.3	91.0	80.0
	DebGCD (Liu & Han, 2025)[ICLR 2025]	98.9	97.5	99.6	**90.1**	**90.9**	88.6	**93.2**	**97.0**	**91.2**	**94.1**	95.1	**93.1**
	SEAL (He et al., 2025)[NeurIPS 2025]	98.9	98.1	99.3	89.8	90.4	**89.5**	91.3	93.3	90.3	93.3	93.9	93.0
	CURE (Bian et al., 2026)[ICML 2026]	98.8	98.8	98.8	89.1	90.1	87.1	91.2	96.7	88.4	93.0	95.2	91.4
	SimGCD (Wen et al., 2023)[ICCV 2023]	98.7	96.7	**99.7**	88.5	89.2	87.2	89.9	95.5	87.1	92.4	93.8	91.3
	+ **SPARC**	**99.1**	98.3	99.5	89.7	90.8	87.4	92.7	96.5	90.8	93.8	95.2	92.6
	Inc. Δ	**+0.4**	**+1.6**	**-0.2**	**+1.2**	**+1.6**	**+0.2**	**+2.8**	**+1.0**	**+3.7**	**+1.4**	**+1.4**	**+1.3**
	SelEx (Rastegar et al., 2024)[ECCV 2024]	98.5*	98.8*	98.5*	87.7*	90.8*	81.5*	90.9*	96.2*	88.3*	92.4	**95.3**	89.4
	+ **SPARC**	98.8	98.2	99.1	89.7	90.8	85.3	91.1	96.6	88.4	93.2	95.2	90.9
	Inc. Δ	**+0.3**	**-0.6**	**+0.6**	**+2.0**	**+0.0**	**+3.8**	**+0.2**	**+0.4**	**+0.1**	**+0.8**	**-0.1**	**+1.5**

†results based on the best baseline (Rastegar et al., 2024). *results from (Liu et al., 2025).

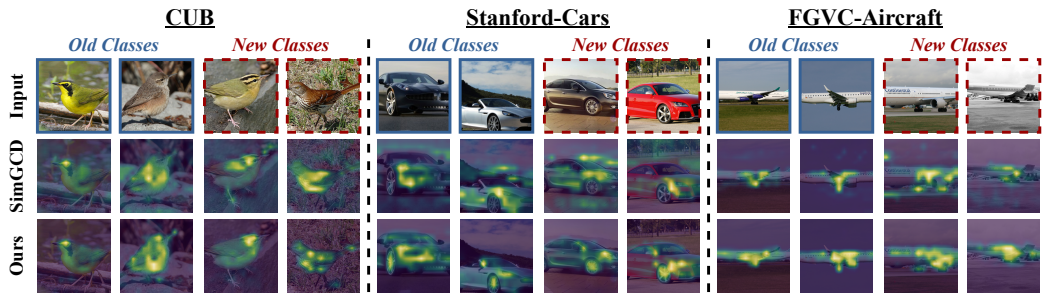

Figure 3: Last-layer CLS-to-patch attention maps for SimGCD (Wen et al., 2023) and SimGCD+**SPARC** on CUB (Wah et al., 2011), Stanford-Cars (Krause et al., 2013), and FGVC-Aircraft (Maji et al., 2013) (DINO backbone). Columns alternate *Old* and *New* samples.

5.3 Qualitative Comparison

Fine-grained GCD demands attention to category-discriminative object parts rather than background clutter (Vaze et al., 2022; Wang et al., 2024). Fig. 3 compares the last-layer CLS-to-patch attention maps of SimGCD and SimGCD+**SPARC** on CUB, Stanford-Cars, and FGVC-Aircraft. Under SimGCD, attention frequently spreads over background regions, specular reflections, or other non-discriminative areas, and this tendency is particularly pronounced on *New* samples. By contrast, **SPARC** concentrates on distinctive local structures—such as beak shape and plumage patterns on birds, headlights and grille geometry on cars, and cabin windows or wing contours on aircraft—while effectively suppressing irrelevant context. This object-centric focus provides a more stable spatial support for prototype assignment, and is consistent with the quantitative gains reported above.

5.4 Ablation Study

Table 2: Clustering accuracy (*ACC*, %) on fine-grained SSB datasets for *All*/*Old*/*New*. The rightmost block is the mean over CUB, Stanford-Cars, and FGVC-Aircraft.

	Method[Venue]	CUB			Stanford-Cars			FGVC-Aircraft			Average		
		All	Old	New	All	Old	New	All	Old	New	All	Old	New
DINO	GCD (Vaze et al., 2022)[CVPR 2022]	51.3	56.6	48.7	39.0	57.6	29.9	45.0	41.1	46.9	45.1	51.8	41.8
	PromptCAL (Zhang et al., 2023)[CVPR 2023]	62.9	64.4	62.1	50.2	70.1	40.6	52.2	52.2	52.3	55.1	62.2	51.7
	GPC (Zhao et al., 2023)[ICCV 2023]	52.0	55.5	47.5	38.2	58.9	27.4	43.3	40.7	44.8	44.5	51.7	39.9
	μGCD (Vaze et al., 2023)[NeurIPS 2023]	65.7	68.0	64.6	56.5	68.1	50.9	53.8	55.4	53.0	58.7	63.8	56.2
	InfoSieve (Rastegar et al., 2023)[NeurIPS 2023]	69.4	77.9	65.2	55.7	74.8	46.4	56.3	63.7	52.5	60.5	72.1	54.7
	FlipClass (Lin et al., 2024)[NeurIPS 2024]	71.3	71.3	71.3	63.1	81.7	53.8	59.3	66.9	55.4	64.6	73.3	60.2
	DebGCD (Liu & Han, 2025)[ICLR 2025]	66.3	71.8	63.5	65.3	81.6	57.4	61.7	63.9	**60.6**	64.4	72.4	60.5
	MOS (Peng et al., 2025)[CVPR 2025]	69.6	72.3	68.2	64.6	80.9	56.7	61.1	66.9	58.2	65.1	73.4	61.0
	APL (Dai et al., 2025)[CVPR 2025]	68.5	73.1	66.2	62.3	80.7	53.4	60.9	63.5	59.6	63.9	72.4	59.7
	AF (Xu et al., 2025)[ICCV 2025]	69.0	74.3	66.3	67.0	80.7	60.4	59.4	68.1	55.0	65.1	74.4	60.6
	†ConGCD (Tang et al., 2025)[ICCV 2025]	**81.7**	80.4	**82.4**	57.5	77.5	47.9	**62.5**	70.2	58.7	67.2	76.0	63.0
	SEAL (He et al., 2025)[NeurIPS 2025]	66.2	72.1	63.2	65.3	79.3	58.5	62.0	65.3	60.4	64.5	72.2	60.7
	TAR (Yang et al., 2026)[CVPR 2026]	70.6	74.8	68.5	65.7	**82.4**	57.5	62.0	69.1	58.7	66.1	75.4	61.6
	LEAH (Duan et al., 2026)[AAAI 2026]	69.3	76.3	65.9	65.5	80.4	58.2	58.3	64.7	55.1	64.4	73.8	59.7
	†CPF-GCD (Dai et al., 2026)[ICML 2026]	79.8	**80.5**	79.4	61.8	69.0	58.1	58.9	**79.0**	49.2	66.8	**76.2**	62.2
	RCA-AFGCD (Yu et al., 2026)[ICML 2026]	71.2	76.4	68.6	65.0	79.2	58.2	60.1	67.1	56.6	65.4	74.2	61.1
	CURE (Bian et al., 2026)[ICML 2026]	75.2	75.5	75.1	**68.7**	80.8	63.0	-	-	-	-	-	-
	SimGCD (Wen et al., 2023)[ICCV 2023]	60.3	65.6	57.7	53.8	71.9	45.0	54.2	59.1	51.8	56.1	65.5	51.5
	+ **SPARC**	68.0	71.7	66.1	67.6	81.2	61.0	58.5	66.2	54.7	64.7	73.0	60.6
	Inc. Δ	**+7.7**	**+6.1**	**+8.4**	**+13.8**	**+9.3**	**+16.0**	**+4.3**	**+7.1**	**+2.9**	**+8.6**	**+7.5**	**+9.1**
	SelEx (Rastegar et al., 2024)[ECCV 2024]	73.6	75.3	72.8	58.5	75.6	50.3	57.1	64.7	53.3	63.1	71.9	58.8
	+ **SPARC**	79.6	76.5	81.1	68.6	76.1	**65.0**	62.4	67.1	60.1	**70.2**	73.2	**68.7**
	Inc. Δ	**+6.0**	**+1.2**	**+8.3**	**+10.1**	**+0.5**	**+14.7**	**+5.3**	**+2.4**	**+6.8**	**+7.1**	**+1.3**	**+9.9**
DINOv2	GCD (Vaze et al., 2022)[CVPR 2022]	71.9	71.2	72.3	65.7	67.8	64.7	55.4	47.9	59.2	64.3	62.3	65.4
	DebGCD (Liu & Han, 2025)[ICLR 2025]	77.5	80.8	75.8	75.4	87.7	69.5	71.9	76.0	69.8	74.9	81.5	71.7
	APL (Dai et al., 2025)[CVPR 2025]	75.1	79.1	73.2	73.4	87.6	66.7	68.8	74.1	66.6	72.4	80.3	68.8
	†ConGCD (Tang et al., 2025)[ICCV 2025]	86.3	87.4	85.8	79.8	93.1	73.3	**81.7**	**83.3**	**81.0**	82.6	87.9	80.0
	SEAL (He et al., 2025)[NeurIPS 2025]	76.7	78.3	75.9	77.7	88.7	72.4	74.6	73.2	75.3	76.3	80.1	74.5
	CURE (Bian et al., 2026)[ICML 2026]	81.3	80.4	81.7	79.9	88.1	76.0	-	-	-	-	-	-
	SimGCD (Wen et al., 2023)[ICCV 2023]	71.5	78.1	68.3	71.5	81.9	66.6	63.9	69.9	60.9	69.0	76.6	65.3
	+ **SPARC**	80.3	82.8	79.0	79.8	90.5	74.7	73.9	74.9	73.3	78.0	82.7	75.7
	Inc. Δ	**+8.8**	**+4.7**	**+10.7**	**+8.3**	**+8.6**	**+8.1**	**+10.0**	**+5.0**	**+12.4**	**+9.0**	**+6.1**	**+10.4**
	SelEx (Rastegar et al., 2024)[ECCV 2024]	87.4	85.1	88.5	82.2	**93.7**	76.7	79.8	82.3	78.6	83.1	87.0	81.3
	+ **SPARC**	**90.5**	**88.6**	**91.5**	**84.0**	93.5	**79.7**	81.6	83.1	80.8	**85.4**	**88.4**	**84.0**
	Inc. Δ	**+3.1**	**+3.5**	**+3.0**	**+1.8**	**-0.2**	**+3.0**	**+1.8**	**+0.8**	**+2.2**	**+2.3**	**+1.4**	**+2.7**

†results based on the best baseline (Rastegar et al., 2024).

Table 3: Component ablation of ICR, SST, and TCR on CUB, Stanford-Cars, and FGVC-Aircraft (SimGCD host, DINO backbone). Metrics are *All*/*Old*/*New ACC* (%).

Components			CUB			Stanford-Cars			FGVC-Aircraft			Average		
ICR	SST	TCR	All	Old	New	All	Old	New	All	Old	New	All	Old	New
✗	✗	✗	60.3	65.6	57.7	53.8	71.9	45.0	54.2	59.1	51.8	56.1	65.5	51.5
✓	✗	✗	62.0	66.7	61.4	55.8	72.4	47.8	56.5	63.4	53.1	58.1	67.5	54.1
✗	✓	✗	63.3	70.5	59.7	63.0	80.6	54.5	57.1	64.8	53.3	61.1	72.0	55.8
✗	✗	✓	63.5	67.7	61.4	54.0	71.9	45.4	55.5	61.8	52.3	57.7	67.1	53.0
✓	✓	✗	63.0	**71.9**	58.6	63.1	79.4	55.3	56.4	61.9	53.7	60.8	71.1	55.9
✗	✓	✓	63.5	70.7	59.8	63.9	79.8	56.2	**59.2**	65.7	**55.9**	62.2	72.1	57.3
✓	✓	✓	**68.0**	71.7	**66.1**	**67.6**	**81.2**	**61.0**	58.5	**66.2**	54.7	**64.7**	**73.0**	**60.6**

Effectiveness of the Components. Tab. 3 ablates the three components of **SPARC**—Image-level Consistency Regularization (ICR), Stable Semantic Token (SST), and Token-level Consistency Regularization (TCR)—on a SimGCD host with a DINO backbone. Each component alone improves over the baseline. SST yields the largest single-factor gain, most notably raising *All ACC* on Stanford-Cars from 53.8% to 63.0% (+9.2; 17.1% relative). Combining SST with TCR brings further gains, and the full model attains the best results on CUB (68.0%) and Stanford-Cars (67.6%), while remaining competitive on FGVC-Aircraft (58.5%). These results indicate that the three components are complementary and usually most effective when used jointly.

Effectiveness of Image-level Consistency Regularization (ICR). Fig. 4 visualizes the ICR mixing operator. The first two columns show source and target images from CUB; the next four columns show mixing with $\rho = 0.03, 0.05, 0.1, 0.2$ (left to right). As ρ grows, low-frequency appearance—color, illumination, and large-scale background—is progressively taken from the target, while phase from the source keeps object layout readable.

Table 4: Ablation of amplitude-mixing radius ρ, bandwidth scale κ, and voting TopK on CUB (SimGCD host, DINO backbone). Baseline *All*/*Old*/*New*: 60.3 / 65.6 / 57.7. (a) ρ (with κ=0.5, TopK=4 fixed); (b) κ (with ρ=0.05, TopK=4 fixed); (c) TopK (with ρ=0.05, κ=0.5 fixed).

(a) amplitude-mixing radius ρ

ρ	All	Old	New
0.03	67.5	**74.5**	64.0
0.05	**68.0**	71.7	**66.1**
0.1	67.4	72.9	64.7
0.2	64.4	70.9	61.2

(b) bandwidth scale κ

κ	All	Old	New
0.25	67.1	**72.9**	64.2
0.5	**68.0**	71.7	**66.1**
0.75	66.5	72.1	63.7
1.0	62.7	69.5	59.4

(c) voting TopK

TopK	All	Old	New
2	65.5	69.8	63.4
4	**68.0**	**71.7**	**66.1**
8	65.4	69.4	63.5
16	62.5	68.9	59.4

Additionally, Table 4 (a) ablates the only free geometric parameter of ICR—the normalized mixing radius ρ. A small ρ already improves over the baseline: even a narrow low-frequency window exposes and suppresses the appearance shortcuts the host overuses. Pushing ρ too far begins to overwrite genuine semantic content instead of isolating spurious low-frequency cues, and accuracy falls below the best setting.

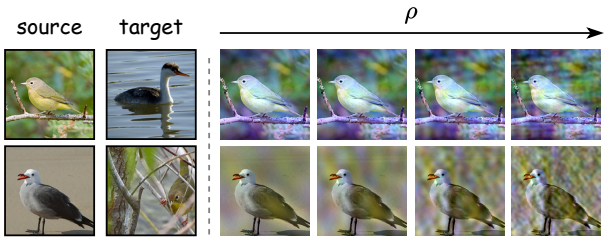

Figure 4: ICR amplitude mixing on CUB. From left to right: source and target images, followed by mixed views as ρ increases.

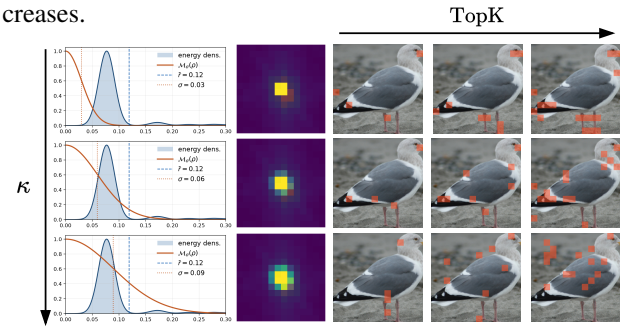

Figure 5: TCR pipeline of **SPARC** on CUB (Wah et al., 2011). Rows vary $\kappa \in \{0.25, 0.5, 0.75\}$; rightmost columns show patches retained by TopK $\in \{4, 8, 16\}$.

Effectiveness of Token-level Consistency Regularization (TCR). Fig. 5 illustrates one forward pass of TCR. From left to right: (i) the energy-weighted radial density and the adaptive Gaussian mask $\mathcal{M}_\sigma$ in Eq. 7, with σ=$\kappa\,\bar{r}$; (ii) the token-grid spectrum with $\mathcal{M}_\sigma$ overlaid, showing how a larger κ widens the low-pass support in Eq. 3; and (iii)–(v) patches retained by channel-wise stability voting and TopK pooling across different κ values (Eqs. 8–9). Together with Tab. 4 (b) and (c), a moderate κ concentrates the selected patches on distinctive object parts—such as the beak, feather tips, tail, and legs—and keeps them object-centric: κ=0.5 and TopK=4 attain the best *All*/*New* (68.0%/66.1%). When κ and TopK become too large, votes spill onto background and texture regions. We therefore adopt κ=0.5 and TopK=4 in the main experiments; the full $\kappa\times$TopK grid is reported in Table 9. In summary, these suggest that voting and TopK selection tend to favor discriminative object regions over surrounding texture or background shortcuts, which is the intended role of the SST.

6 Conclusion

We present **SPARC**, a plug-and-play module for **SP**ectral **A**ssignment **R**egularization for **C**ategories. Motivated by the premise that category assignments should remain stable under controlled spectral perturbations, SPARC regularizes prototype assignments rather than directly aligning features. Image-level Consistency Regularization (ICR) mixes low-frequency amplitudes while retaining the source phase and encourages the mixed view to preserve the clean assignment. Token-level Consistency Regularization (TCR) pools spectrally stable patches into a Stable Semantic Token (SST) and aligns its assignment with that of the global descriptor. For parametric hosts, SST additionally replaces the CLS readout during training and inference. Across generic and fine-grained GCD benchmarks with representative hosts and backbones, SPARC achieves positive average improvements, with particularly strong gains on fine-grained datasets. The ablations further suggest that ICR, SST, and TCR provide complementary benefits, although their effects vary across datasets and hyperparameter settings. Overall, **SPARC** provides a frequency-aware approach to improving assignment stability under controlled appearance changes in Generalized Category Discovery.

AI USE STATEMENT

We used AI tools only to assist with and polish the writing of this paper, including grammar checking, wording refinement, and readability improvement. AI tools were not used to generate research ideas, design experiments, analyze results, or draw conclusions.

REFERENCES

Yuwei Bian, Shidong Wang, and Haofeng Zhang. CURE: Consistency-under-Unified semantic regularization for generalized category discovery. In *ICML*, 2026.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

Guangyao Chen, Peixi Peng, Li Ma, Jia Li, Lin Du, and Yonghong Tian. Amplitude-phase recombination: Rethinking robustness of convolutional neural networks in frequency domain. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Boyang Dai, Chaoqi Chen, and Yizhou Yu. Identifying latent concepts and structures for generalized category discovery. In *ICML*, 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 *CVPR*, 2025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

Yu Duan, Junzhi He, Zhanxuan Hu, Mengda Ji, Rong Wang, and Quanxue Gao. Learning intrinsic hierarchy for generalized category discovery. In *AAAI*, 2026.

Wei Feng and Zongyuan Ge. Generalized category discovery under domain shift: A frequency domain perspective. In *NeurIPS*, 2025.

Wei Feng, Sijin Zhou, Yiwen Jiang, and Zongyuan Ge. PRISM: Progressive robust learning for open-world continual category discovery. In *ICLR*, 2026.

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. In *ICLR*, 2019.

Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *ICCV*, 2019.

Zhenqi He, Yuanpei Liu, and Kai Han. SEAL: Semantic-aware hierarchical learning for generalized category discovery. In *NeurIPS*, 2025.

Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. FSDR: Frequency space domain randomization for domain generalization. In *CVPR*, 2021.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *ICCV Workshops*, 2013.

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

Haonan Lin, Wenbin An, Jiahao Wang, Yan Chen, Feng Tian, Mengmeng Wang, Qianying Wang, Guang Dai, and Jingdong Wang. Flipped classroom: Aligning teacher attention with student in generalized category discovery. In *NeurIPS*, 2024.

Yuanpei Liu and Kai Han. DebGCD: Debiased learning with distribution guidance for generalized category discovery. In *ICLR*, 2025.

Yuanpei Liu, Zhenqi He, and Kai Han. Hyperbolic category discovery. In *CVPR*, 2025.

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

Alan V. Oppenheim and Jae S. Lim. The importance of phase in signals. *Proceedings of the IEEE*, 1981.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024.

Rabah Ouldnoughi, Chia-Wen Kuo, and Zsolt Kira. CLIP-GCD: Simple Language-Guided Generalized Category Discovery. *arXiv preprint arXiv:2305.10420*, 2023.

Zhengyuan Peng, Jinpeng Ma, Zhimin Sun, Ran Yi, Haichuan Song, Xin Tan, and Lizhuang Ma. MOS: Modeling object-scene associations in generalized category discovery. In *CVPR*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Sarah Rastegar, Hazel Doughty, and Cees Snoek. Learn to categorize or categorize to learn? Self-Coding for generalized category discovery. In *NeurIPS*, 2023.

Sarah Rastegar, Mohammadreza Salehi, Yuki M Asano, Hazel Doughty, and Cees G M Snoek. SelEx: Self-expertise in fine-grained generalized category discovery. In *ECCV*, 2024.

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.

Luyao Tang, Kunze Huang, Chaoqi Chen, Yuxuan Yuan, Chenxin Li, Xiaotong Tu, Xinghao Ding, and Yue Huang. Dissecting generalized category discovery: Multiplex consensus under self-deconstruction. In *ICCV*, 2025.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *CVPR*, 2022.

Sagar Vaze, Andrea Vedaldi, and Andrew Zisserman. No representation rules them all in category discovery. In *NeurIPS*, 2023.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, California Institute of Technology, 2011.

Enguang Wang, Zhimao Peng, Zhengyuan Xie, Fei Yang, Xialei Liu, and Ming-Ming Cheng. GET: Unlocking the Multimodal Potential of CLIP for generalized category discovery. In *CVPR*, 2025a.

Hongjun Wang, Sagar Vaze, and Kai Han. SPTNet: An efficient alternative framework for generalized category discovery with spatial prompt tuning. In *ICLR*, 2024.

Hongjun Wang, Sagar Vaze, and Kai Han. HiLo: A learning framework for generalized category discovery robust to domain shifts. In *ICLR*, 2025b.

Xin Wen, Bingchen Zhao, and Xiaojuan Qi. Parametric classification for generalized category discovery: A baseline study. In *ICCV*, 2023.

Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *ICLR*, 2021.

Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A Fourier-based framework for domain generalization. In *CVPR*, 2021.

Qiyu Xu, Zhanxuan Hu, Yu Duan, Ercheng Pei, and Yonghang Tai. A hidden stumbling block in generalized category discovery: Distracted attention. In *ICCV*, 2025.

Muli Yang, Jie Yin, Yanan Gu, Cheng Deng, Hanwang Zhang, and Hongyuan Zhu. Consistent prompt tuning for generalized category discovery. *IJCV*, 2025.

Xingyu Yang, Yu Zhang, Siya Mi, and Xiu-Shen Wei. TAR: Token-aware refinement for fine-grained generalized category discovery. In *CVPR*, 2026.

Yanchao Yang and Stefano Soatto. FDA: Fourier domain adaptation for semantic segmentation. In *CVPR*, 2020.

Dong Yin, Raphael Gontijo Lopes, Jonathon Shlens, Ekin D. Cubuk, and Justin Gilmer. A Fourier Perspective on model robustness in computer vision. In *NeurIPS*, 2019.

Jiawei Yu, Zijian Gao, Tianjiao Wan, Xuan Liu, Cheng Yang, and Kele Xu. Reliable confidence alignment for generalized category discovery. In *ICML*, 2026.

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 *CVPR*, 2023.

Bingchen Zhao, Xin Wen, and Kai Han. Learning semi-supervised Gaussian mixture models for generalized category discovery. In *ICCV*, 2023.

Haiyang Zheng, Nan Pu, Wenjing Li, Nicu Sebe, and Zhun Zhong. Textual knowledge matters: Cross-modality co-teaching for generalized visual class discovery. In *ECCV*, 2024.

A Implementation Details

A.1 Dataset Statistics

We follow the standard GCD split of Vaze et al. (2022). Half of the classes are treated as known (*Old*), except on CIFAR-100, where 80% of the classes are known. For each known class, half of the images form the labeled set $\mathcal{D}_l$ and the remainder are placed in the unlabeled set $\mathcal{D}_u$, which also contains all images from novel classes $\mathcal{Y}_n$. Dataset sizes are summarized in Table 5, where $M = |\mathcal{Y}_l|$ and $K = |\mathcal{Y}_l \cup \mathcal{Y}_n|$. As in prior GCD work (Vaze et al., 2022; Wen et al., 2023; Rastegar et al., 2024), the total number of categories K is assumed known at training and evaluation; when K is unknown, it can be estimated by existing procedures (Han et al., 2019; Vaze et al., 2022), which we do not study here.

Table 5: Overview of evaluation datasets: labeled/unlabeled sizes ($|\mathcal{D}_l|$, $|\mathcal{D}_u|$) and class counts ($M = |\mathcal{Y}_l|$, $K = |\mathcal{Y}_l \cup \mathcal{Y}_n|$). "FG" denotes fine-grained.

Dataset	FG	$\lvert\mathcal{D}_l\rvert$	M	$\lvert\mathcal{D}_u\rvert$	K
CIFAR-10 (Krizhevsky et al., 2009)	✗	12.5K	5	37.5K	10
CIFAR-100 (Krizhevsky et al., 2009)	✗	20.0K	80	30.0K	100
ImageNet-100 (Deng et al., 2009)	✗	31.9K	50	95.3K	100
CUB (Wah et al., 2011)	✓	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

A.2 Evaluation Metrics

Following the standard transductive GCD protocol (Vaze et al., 2022), *ACC* is evaluated on the unlabeled pool $\mathcal{D}_u$, which participates in unsupervised training but whose labels are used only for post-hoc evaluation. With ground-truth labels y_i and predictions $\hat{y}_i$,

$$ACC = \frac{1}{|\mathcal{D}_u|} \sum_{i=1}^{|\mathcal{D}_u|} \mathbb{1}\big(y_i = h(\hat{y}_i)\big),$$

where h is the optimal one-to-one mapping obtained by the Hungarian algorithm on *all* unlabeled samples (following the $\mathrm{v2}$ protocol of Vaze et al. (2022) and Wen et al. (2023)). *Old* and *New ACC* reuse this single All-set mapping and then average only over the corresponding subsets; we do not recompute a separate Hungarian map on *Old* or *New* alone. This is the protocol used by SimGCD and SelEx, and is the number reported throughout.

A.3 Training Protocol

We attach **SPARC** to the official SimGCD (Wen et al., 2023) and SelEx (Rastegar et al., 2024) recipes and keep each host's optimizer, schedule, and spatial views unchanged. Backbones are DINO ViT-B/16 (Caron et al., 2021) and DINOv2 ViT-B/14 (Oquab et al., 2024); CLIP and DINOv3 comparisons are reported in Table 6. Images are resized to 224×224. Each host uses two independently augmented spatial views per image. Blocks 0–10 of the ViT are frozen; the last block, projection head, and (for SimGCD) classifier/prototypes are trained.

Both hosts use SGD with momentum 0.9, learning rate 0.1, weight decay 5×10^{-5}, batch size 128, and 200 epochs, with a cosine schedule decaying to 10^{-3} of the initial learning rate. On SimGCD, the clustering teacher temperature is linearly warmed from 0.07 to 0.04 over the first 30 epochs and the student/clustering temperature is 0.1; prototypes are learnable vectors in P and are updated by SGD through $\mathcal{L}_{\mathrm{host}}$. On SelEx, supervised/unsupervised contrastive losses use the official temperature 0.07; the assignment bank P is refreshed from embeddings by spherical k-means (hierarchical self-expertise), not by gradient updates to a parametric classifier.

SPARC adds one spectral mixed view per image for ICR: a donor is obtained by a cyclic shift of the minibatch with offset in $\{1, \ldots, |\mathcal{B}|-1\}$ (no self-pairing; $|\mathcal{B}|=1$ skips mixing), and $\lambda \sim \mathrm{Unif}[0, 1]$ is drawn independently per sample. This mixed view is used only in $\mathcal{L}_{\mathrm{img}}$ and does not replace the host's two spatial views. ICR is applied after denormalizing RGB to $[0, 1]$ with a real FFT (no shift; conjugate symmetry is implicit) and the real inverse, without clipping; TCR uses a complex FFT with $\mathrm{FFTShift}$/$\mathrm{IFFTShift}$ and the real part. Numerical floors use $\epsilon=10^{-6}$; TopK ties keep the first index. Unless a table notes otherwise, we use the CUB-validated defaults from Tables 4 and 9: ICR radius $\rho=0.05$, TCR scale $\kappa=0.5$, pooling/voting $K_s=K_v=4$, and auxiliary weights

λ_{img}=0.25, λ_{tok}=0.05. Both spectral weights are linearly warmed up over the first W=30 epochs via $\gamma_b(t) = \min(1, t/W)$ in Eq. 12; this schedule is independent of SimGCD's teacher-temperature warmup. The assignment KL uses student temperature 0.1 and a stop-gradient teacher assignment. On SimGCD, the Stable Semantic Token replaces the CLS readout inside $\mathcal{L}_{\text{host}}$ and at evaluation; on SelEx, SST is computed only for TCR and does not replace the host embedding used by SS-k-means.

B Additional Quantitative Results

B.1 Experiments with different backbones

We further evaluate **SPARC** using two additional backbones: CLIP ViT-B/16 (Radford et al., 2021) and DINOv3 ViT-B/16 (Siméoni et al., 2025) backbones.

Table 6: Results on fine-grained datasets (*All*/*Old*/*New ACC*, %) with CLIP and DINOv3 backbones. The rightmost block reports the mean across datasets.

	Method[Venue]	CUB			Stanford-Cars			FGVC-Aircraft			Average		
		All	Old	New	All	Old	New	All	Old	New	All	Old	New
CLIP	GCD (Vaze et al., 2022)[CVPR 2022]	57.6	65.2	53.8	65.1	75.9	59.8	45.3	44.4	45.8	56.0	61.8	53.1
	CLIP-GCD (Ouldnoughi et al., 2023)[arXiv 2023]	62.8	77.1	55.7	70.6	88.2	62.2	50.0	56.6	46.5	61.1	74.0	54.8
	TextGCD (Zheng et al., 2024)[ECCV 2024]	76.6	80.6	74.7	86.9	87.4	86.7	-	-	-	-	-	-
	CPT (Yang et al., 2025)[IJCV 2025]	70.1	73.5	68.4	74.2	84.3	69.3	-	-	-	-	-	-
	GET (Wang et al., 2025a)[CVPR 2025]	77.0	78.1	76.4	78.5	86.8	74.5	58.9	59.6	58.5	71.5	74.8	69.8
	SimGCD (Wen et al., 2023)[ICCV 2023]	71.7	76.5	69.4	70.0	83.4	63.5	54.3	58.4	52.2	65.3	72.8	61.7
	+ SPARC	73.9	77.9	71.9	73.9	85.1	68.5	61.7	65.0	60.1	69.8	76.0	66.8
	Inc. Δ	**+2.2**	**+1.4**	**+2.5**	**+3.9**	**+1.7**	**+5.0**	**+7.4**	**+6.6**	**+7.9**	**+4.5**	**+3.2**	**+5.1**
DINOv3	SimGCD (Wen et al., 2023)[ICCV 2023]	75.9	83.8	72.0	73.9	79.4	71.3	71.4	84.4	65.0	73.7	82.5	69.4
	+ SPARC	82.1	85.7	80.3	78.0	83.7	75.2	75.8	83.0	72.2	78.6	84.1	75.9
	Inc. Δ	**+6.2**	**+1.9**	**+8.3**	**+4.1**	**+4.3**	**+3.9**	**+4.4**	**-1.4**	**+7.2**	**+4.9**	**+1.6**	**+6.5**
	SelEx (Rastegar et al., 2024)[ECCV 2024]	83.5	78.1	86.2	78.8	90.3	73.3	73.9	77.3	72.4	78.7	81.9	77.3
	+ SPARC	85.6	85.6	85.7	84.0	93.4	79.5	77.9	80.1	76.6	82.5	86.4	80.6
	Inc. Δ	**+2.1**	**+7.5**	**-0.5**	**+5.2**	**+3.1**	**+6.2**	**+4.0**	**+2.8**	**+4.2**	**+3.8**	**+4.5**	**+3.3**

B.2 Effects of ICR loss weight

This hyperparameter is the coefficient of ICR. The loss asks the soft assignment to stay fixed when low-frequency amplitude is mixed and the source phase is kept. By fixing the remaining hyperparameters (e.g., $\lambda_{\text{tok}} = 0.05$), we keep the configuration aligned with Tab. 2. As shown in Tab. 7, raising λ_{img} from 0.15 to 0.25 improves *All ACC* from 65.3% to 68.0% with DINO and from 78.6% to 80.3% with DINOv2; 0.35 begins to degrade both. A moderate weight encourages invariance under appearance change, while an excessively large weight may enforce invariance even when the perturbation changes class-relevant information.

B.3 Effects of TCR loss weight

This hyperparameter is the coefficient of TCR, which aligns the SST with the CLS token in assignment space. By fixing the remaining hyperparameters (e.g., $\lambda_{\text{img}} = 0.25$), we keep the configuration aligned with Tab. 2. As shown in Tab. 8, raising λ_{tok} from 0.02 to 0.05 improves *All ACC* from 66.0% to 68.0% with DINO and from 79.0% to 80.3% with DINOv2. Larger weights hurt: with DINO, *All* and *New ACC* drops by 1.9 and 1.8 percentage points, respectively. A moderate token constraint keeps the assignment tied to spectrally stable local evidence; an excessive weight

Table 7: Ablation of ICR weight λ_{img} on CUB (SimGCD host; λ_{tok}=0.05 fixed). Metrics are *All*/*Old*/*New ACC* (%).

	DINO			DINOv2		
λ_{img}	All	Old	New	All	Old	New
baseline	60.3	65.6	57.7	71.5	78.1	68.3
SST only	63.3	70.5	59.7	72.1	79.6	68.4
0.15	65.3	70.2	62.8	78.6	82.0	76.9
0.25	**68.0**	**71.7**	**66.1**	**80.3**	**82.8**	**79.0**
0.35	64.8	71.7	61.3	78.9	81.8	77.5

Table 8: Ablation of TCR weight λ_{tok} on CUB (SimGCD host; λ_{img}=0.25 fixed). Metrics are *All*/*Old*/*New ACC* (%).

	DINO			DINOv2		
λ_{tok}	All	Old	New	All	Old	New
baseline	60.3	65.6	57.7	71.5	78.1	68.3
SST only	63.3	70.5	59.7	72.1	79.6	68.4
0.02	66.0	**72.9**	62.5	79.0	80.7	78.2
0.05	**68.0**	71.7	**66.1**	**80.3**	**82.8**	**79.0**
0.10	66.1	69.9	64.3	77.2	82.1	74.7

Table 9: Joint ablation of bandwidth scale κ and voting TopK on CUB (SimGCD host, DINO backbone; ρ=0.05, λ_{img}=0.25, λ_{tok}=0.05). Each cell reports *All*/*Old*/*New ACC* (%). Best *ACC* is bold.

$\kappa \setminus$ TopK	2	4	8	16
0.25	65.7 / 71.2 / 62.9	67.1 / 72.9 / 64.2	67.1 / 71.2 / 65.1	64.7 / 72.7 / 60.7
0.50	65.5 / 69.8 / 63.4	**68.0** / 71.7 / **66.1**	65.4 / 69.4 / 63.5	62.5 / 68.9 / 59.4
0.75	67.2 / **74.1** / 63.7	66.5 / 72.1 / 63.7	65.4 / 68.3 / 63.9	59.8 / 66.1 / 56.6
1.00	65.8 / 68.0 / 64.7	62.7 / 69.5 / 59.4	62.5 / 66.8 / 60.4	58.1 / 66.0 / 54.1

over-commits the model to that pooled descriptor. We therefore keep λ_{tok}=0.05 for both backbones, with λ_{img} fixed at 0.25.

B.4 Joint sweep of κ and TopK

Tab. 4 reports one-dimensional sweeps of κ and TopK. Here we fix ρ=0.05, λ_{img}=0.25, and λ_{tok}=0.05, and sweep the full grid $\kappa \in \{0.25, 0.5, 0.75, 1.0\} \times \text{TopK} \in \{2, 4, 8, 16\}$ on CUB with a SimGCD host and DINO backbone (Tab. 9). The best cell remains $(\kappa, \text{TopK}) = (0.5, 4)$ at $68.0/71.7/66.1$ (*All*/*Old*/*New*), matching the main setting. Performance is stable in a neighbourhood of this point (e.g., κ=0.25 with $\text{TopK} \in \{4, 8\}$), but degrades when both the low-pass support and the pooled set grow large together—consistent with votes spilling onto background texture (Fig. 5).

C Additional Qualitative Results

C.1 Assignment Space Analysis

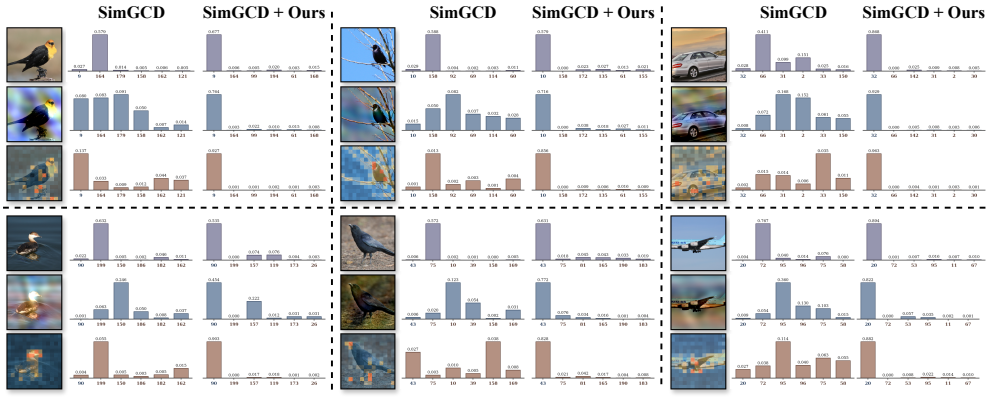

Figure 6: Soft assignment distributions under controlled spectral perturbations on CUB (Wah et al., 2011), Stanford-Cars (Krause et al., 2013), and FGVC-Aircraft (Maji et al., 2013). Blue/red slot IDs denote *known*/*novel* categories.

Fig. 6 probes assignment stability under semantic-preserving spectral perturbations. For each sample we apply the same low-frequency amplitude mixing used by ICR and plot the resulting soft-assignment distribution over prototypes. Under SimGCD, the mass often drifts away from the correct slot or spreads across unrelated categories, even though object layout is unchanged. With **SPARC**, the peak remains on the correct prototype and competing slots stay suppressed across CUB, Stanford-Cars, and FGVC-Aircraft. This shows that the baseline assignment is unstable under spec-

Table 10: Assignment 2×2 (%) on the same CUB $\mathcal{D}_u$ (n=4496), matching the three rows of Fig. 6. Columns are **SPARC**, rows are SimGCD. (a,c) correct (✓) / wrong (✗). (b) stable (✓) / flip (✗) under ICR mix.

		(a) Clean Image		(b) Image Variation		(c) Token Variation	
		✓	✗	✓	✗	✓	✗
SimGCD	✓	50.2	9.7	85.9	0.9	8.6	1.4
	✗	17.1	23.0	12.9	0.4	58.6	31.4

tral variations, whereas **SPARC** maintains the matched prediction in these examples—matching the premise that GCD should regularize assignments, not only features.

The figure has three rows: the clean host, the same host after ICR mixing, and a token pooled from TCR-selected patches. Tab. 10 puts numbers on those three rows, using the same 4496 unlabeled CUB images for every block (DINO, SimGCD vs. **SPARC**). Rows are SimGCD and columns are **SPARC**. A single `v2` Hungarian map (Appendix A) is fit on each model's clean-host outputs and reused in (b) and (c), so the three blocks differ only in the readout, not in the image set.

(a) compares prediction correctness on the clean images (CLS vs. SST). The 17.1 vs. 9.7 off-diagonal is the +7.4 All-*ACC* gap (59.9→67.3): **SPARC** corrects more SimGCD errors than it introduces. (These clean-image accuracies are obtained from the diagnostic analysis run and therefore differ slightly from the main-run results.) (b) compares prediction stability after ICR mixing (ρ=0.05, five shared donors). Here ✓/✗ mean the aligned prediction is unchanged / flipped on a majority of mixes. SimGCD flips while **SPARC** stays on 12.9% of the pool; the reverse is 0.9%. (c) asks whether both trained heads can read the same SST token (TCR vote, κ=0.5, TopK=4). **SPARC** does (67.2 All-*ACC*); SimGCD, trained on CLS, reaches only 10.0%.. These percentages are the population support for Fig. 6; (c) is a probe of the selected token and does not replace the main-table host *ACC*.

D Case Analysis

Tab. 11 lists the three *Old* and three *New* classes with the largest per-class *ACC* gains on CUB (SimGCD host, DINO backbone). *Old* classes improve (e.g., Black Tern and Red-faced Cormorant, both +53.3%), but the largest lifts appear on *New* classes. Two of the three have baseline accuracy 0%, with most images absorbed by a single unrelated prototype (e.g., Lazuli Bunting→Forster's Tern). Adding **SPARC** reverses this collapse: these classes rise to 90.0%–96.7%.

Table 11: Per-class *ACC* on CUB: classes with the largest gains under **SPARC**. "Wrong Category" is SimGCD's most frequent incorrect label for that class (share of images in parentheses); Δ is the absolute change in per-class *ACC*.

			ACC (%)			
Slot ID	Class Name	Type	Baseline	+ **SPARC**	Δ	Wrong Category
142	Black Tern	Old	33.3	86.7	+53.3	Evening Grosbeak (13%)
24	Red-faced Cormorant	Old	20.0	73.3	+53.3	Gadwall (47%)
6	Least Auklet	Old	50.0	87.5	+37.5	European Goldfinch (38%)
15	Lazuli Bunting	New	0.0	96.7	+96.7	Forster's Tern (57%)
76	Dark-eyed Junco	New	0.0	90.0	+90.0	Cedar Waxwing (77%)
166	Golden-winged Warbler	New	6.7	93.3	+86.7	Geococcyx (73%)