DRP: LONG-TAILED DEEP CLUSTERING VIA DENOISING-SUPPORTED REPULSIVE DIRICHLET PROCESS

Anonymous authors
Paper under double-blind review

ABSTRACT

In long-tailed deep clustering, dense head regions tend to occupy redundant prototype capacity, leading to head-class over-segmentation and tail-class absorption. To address this issue, we propose DRP, a nonparametric prototype inference framework for long-tailed deep clustering that rectifies prototype-capacity allocation bias by governing prototype birth, activation, repulsion and reallocation. DRP builds a semantic support measure via denoising consistency to stop dense regions from gaining unfair prototype birth advantages. Using this measure as the base distribution, it constructs a Dirichlet-process reference law and derives a Gibbs-tilted prototype prior minimizing KL divergence under a log-determinantal redundancy constraint. This prior controls active prototype count and suppresses redundant similar prototypes in head regions. For under-covered but stable semantic regions, cardinality-preserving prototype exchange reallocates redundant prototype slots. An identity-initialized residual adapter injects the corrected prototype structure back into the representation space; prototype-balanced feedback and conservative regularization limit representation drift. All components are integrated into a finite alternating MAP inference pipeline requiring no ground-truth labels, class frequencies or predefined class number. Extensive experiments validate DRP's state-of-the-art performance on long-tailed clustering benchmarks, strong generalization to balanced data, and ability to infer unknown class counts.

1 INTRODUCTION

Deep clustering aims to jointly learn discriminative representations and latent class partitions from unlabeled data, and has become a fundamental problem in unsupervised visual understanding (Xie et al., 2016; Li et al., 2021). In recent years, contrastive learning, self-supervised representation learning, and prototype-based learning have substantially improved deep clustering performance. However, most existing methods (Caron et al., 2019; Ronen et al., 2022; Huang et al., 2024; Zhan et al., 2020) are developed under balanced or approximately balanced class distributions. Real-world visual data, by contrast, often exhibit pronounced long-tailed structure, where a small number of head classes contain abundant samples while the majority of tail classes are represented by only limited observations. This setting becomes particularly challenging under fully unsupervised learning, because class labels, class frequencies, and even the true number of classes are unavailable. Consequently, unlike supervised long-tailed learning, the model cannot directly perform class-level rebalancing. Maintaining the separability of both head and tail classes under an unknown class structure therefore remains a central challenge in long-tailed deep clustering.

Existing long-tailed clustering methods mainly address this challenge by strengthening the training signal associated with tail samples. For example, ConMix adjusts the contributions of different samples through contrastive mixing in the representation space (Li & Jia, 2025); DGD enhances sparse regions through a density-gradient-guided diffusion process (Ye et al., 2026); and MiniClustering estimates unsupervised sample weights using micro-clusters obtained from over-clustering (Li et al., 2026). These methods primarily operate on sample weights, feature distributions, or training data. However, they require an additional specified number of clusters and pay relatively limited attention to how prototype capacity is ultimately allocated across latent classes during clustering. As shown in Figure 1(a), existing SOTA methods tend to activate multiple highly similar prototypes in dense head

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.

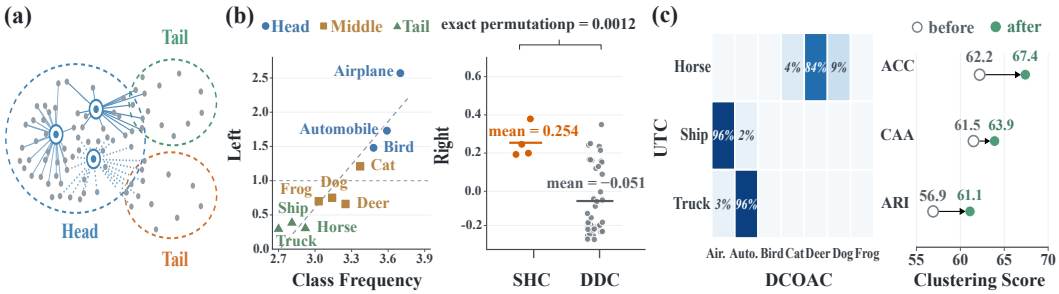


Figure 1: (a) Conceptual illustration: dense head-class regions activate multiple nearby prototypes, causing head-class over-segmentation; sparse tail-class regions lack dedicated prototypes, and their samples are absorbed by neighboring head-dominated clusters. (b) Left: effective prototype capacity correlates positively with class frequency. Right: prototypes within head classes are highly similar. SHC: Same Head-dominant Class; DDC: Different Dominant Classes. (c) Left: many tail samples are assigned to head regions. Right: redundant-prototype reallocation improves performance. UTC: Uncovered Tail Class. DCOAC: Dominant Class of the Assigned Cluster.

regions, whereas sparse tail regions struggle to acquire dedicated prototypes, resulting in head-class over-segmentation and tail-class absorption. Figure 1(b) further shows that the effective prototype capacity ultimately allocated to each class is significantly positively correlated with its class frequency, while prototypes within head classes exhibit high similarity. As illustrated in Figure 1(c), these observations reveal a causal chain beyond the conventional explanation of insufficient tail samples: because the clustering objective accumulates contributions over samples, fitting head regions yields greater overall gains, making local variations within these regions more likely to activate additional similar prototypes. These redundant prototypes further increase the attraction capacity of head regions, causing nearby tail samples to be progressively absorbed. Reallocating redundant prototypes from head classes, in turn, improves clustering quality. We refer to this phenomenon as *prototype-capacity allocation bias*. This observation suggests that the difficulty of long-tailed clustering arises not only from insufficient tail samples, but also from the redundant occupation of limited prototype capacity by head classes. Simply strengthening tail signals therefore does not structurally resolve the problem.

Motivated by this finding, we propose DRP, a nonparametric framework for long-tailed deep clustering to correct prototype-capacity allocation bias in prototype inference. Instead of letting sample density dominate prototype assignment, DRP jointly evaluates semantic stability, prototype redundancy and capacity utilization. Starting from an overcomplete prototype pool, it models prototype birth, activation, retention and reallocation: unstable candidates are suppressed, head-region similar prototypes incur higher retention costs, and capacity is diverted to stably supported yet under-covered regions. Without ground-truth labels, class frequencies or predefined class counts, DRP reduces head redundancy and allocates prototypes for sparse stable tail semantics. Specifically, we first construct a support measure from denoising consistency, where prototype birth eligibility depends on a support threshold instead of raw sample frequency. Built upon this measure, we develop a novel Gibbs-tilted DP prior under the log-determinantal redundancy constraint to penalize highly similar prototype configurations. A cardinality-preserving prototype exchange swaps redundant active prototypes with under-covered candidates for capacity reallocation. An identity-initialized residual adapter embeds corrected prototypes into feature space; prototype-balanced feedback and conservative regularization limit representation drift and collapse. These modules constitute an alternating MAP optimization pipeline. Experiments show DRP achieves SOTA results on long-tailed clustering benchmarks, generalizes well to balanced data, and infers unknown class counts.

Our main contributions are summarized as follows:

- We introduce a denoising-supported base measure for prototype inference. A one-sided support barrier suppresses unsupported candidate directions while avoiding additional birth preference for head regions once sufficient support is established.
- We construct a Dirichlet-process reference measure from denoised semantic base measure, derives the Gibbs-interacting prototype law via minimum-information KL projection under the log-determinantal redundancy constraint, and regulates prototype evolu-

tion with cardinality-preserving prototype exchange. DRP requires no ground-truth labels, class-frequency statistics or pre-defined class number.

- We design a unified MAP-style prototype-inference objective and a finite alternating optimization algorithm, together with conditional mechanism analyses. Experiments verify that DRP achieves SOTA results on multiple long-tailed clustering benchmarks, exhibits strong generalization to balanced data, and can infer unknown class counts.

2 Related Work

Unsupervised Long-Tailed Deep Clustering. Deep clustering jointly learns feature representations and cluster assignments from unlabeled data. Representative methods such as DEC (Xie et al., 2016) and DeepCluster (Caron et al., 2019) couple representation learning with iterative clustering or pseudo-label optimization. Later approaches improve clustering through view invariance, online assignment updates, and contrastive learning, including Invariant Information Clustering (Ji et al., 2019), Online Deep Clustering (Zhan et al., 2020), SwAV (Caron et al., 2021) and Contrastive Clustering (Li et al., 2021). Recent methods further exploit neighborhood consistency, prototype scattering, mutual-information constraints, sample stability, and semantic guidance to improve representation quality and assignment stability (Yu et al., 2023; Huang et al., 2023; Li et al., 2023; Qian, 2023; Li et al., 2025; Cai et al., 2023). However, most existing methods assume a predefined cluster cardinality, and their prototype allocation is still indirectly determined by sample-level objectives. Long-tailed deep clustering has recently attracted increasing attention. ConMix (Li & Jia, 2025) uses representation-level contrastive mixup to rebalance the contributions of head and tail samples. DGD (Ye et al., 2026) enhances sparse regions through density-gradient-guided diffusion, while MiniClustering (Li et al., 2026) estimates unsupervised sample weights from over-clustered mini-clusters. These methods mainly correct sample contributions, feature distributions, or training data. In contrast, our method directly performs support-aware prototype inference by controlling prototype generation, redundancy, and capacity reallocation across latent classes.

Nonparametric Models, Prototype Geometry, and Support Priors. Bayesian nonparametric models, including Dirichlet process mixtures, provide a probabilistic basis for clustering without fixing the number of clusters (Ferguson, 1973; Antoniak, 1974; Blei & Jordan, 2006). DeepDPM extends this idea to deep clustering through split–merge operations and dynamic structure adjustment (Ronen et al., 2022), while generalized category discovery relaxes the known-cardinality assumption in a partially labeled open-world setting (Vaze et al., 2022). Prototype geometry provides a complementary source of structural regularization: ProPos promotes prototype scattering and inter-cluster separation (Huang et al., 2023), and DivClust controls diversity across alternative clustering solutions (Metaxas et al., 2023). Determinantal point processes offer a general probabilistic framework for subset diversity (Kulesza & Taskar, 2012). Our method combines these ideas in a different way by using a Gibbs-form log-determinantal term as a prior over the active prototype set, while denoising consistency supplies a one-sided support criterion for prototype validity. In this way, our method jointly couples unknown-cardinality inference, redundancy suppression, and capacity-preserving prototype reallocation for long-tailed deep clustering.

3 Proposed Method

3.1 Problem Formulation and Framework Overview

Problem Formulation. Let $\mathcal{D} = \{x_i\}_{i=1}^{N}$ be an unlabeled image set sampled from an unknown, potentially long-tailed class distribution. Labels, class frequencies, and the true class count are unavailable during training. DRP jointly infers a partition and its active cardinality while reducing redundant head coverage and improving independent coverage of supported tail semantics.

Framework Overview. As illustrated in Figure 2(a), DRP employs a frozen encoder E_0 to establish a stable base representation and attaches an identity-initialized lightweight residual adapter:

$$u_i^{(v)} = \mathrm{norm}(E_0(t_v(x_i))), \qquad z_i^{(v)}(\phi) \quad = \mathrm{norm}\Big(u_i^{(v)} + \gamma A_\phi(u_i^{(v)})\Big), \qquad A_{\phi_0} \equiv 0. \tag{1}$$

Here, $u_i^{(v)}$ and $z_i^{(v)}(\phi)$ denote the base feature and adapted feature of the v-th augmented view, respectively. The fixed space $U = \{u_i\}$ is used to determine whether a candidate direction is supported

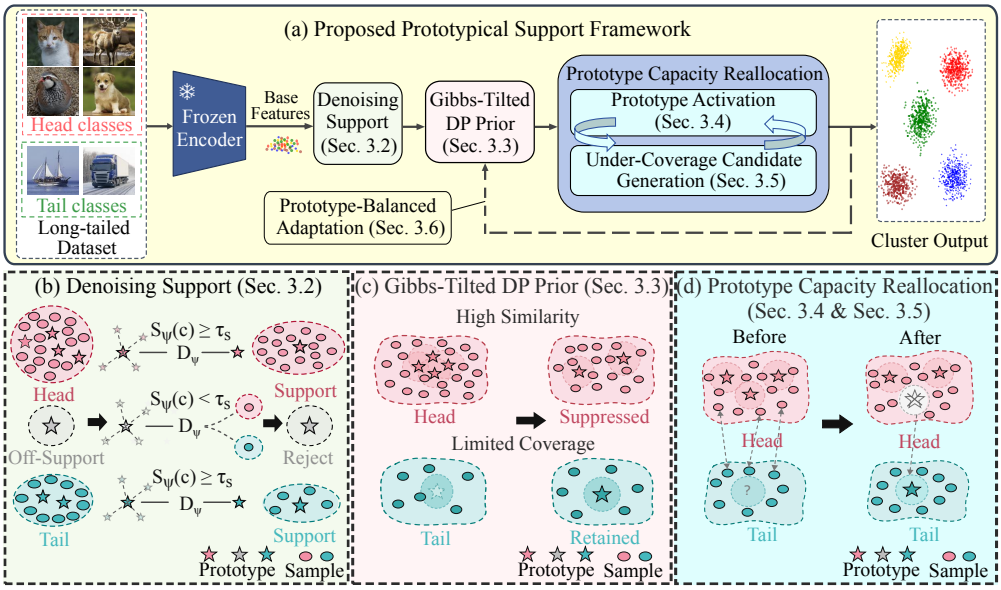


Figure 2: Overview of the proposed framework. (a) Given features extracted by a frozen encoder, the framework sequentially applies denoising-supported prototype selection, Gibbs repulsion, and prototype capacity reallocation, while a residual adapter refines the representations for the final cluster assignments. (b) The denoising-supported mechanism retains prototypes with sufficient local support from both head and tail classes and rejects off-support prototypes. (c) Gibbs repulsion suppresses redundant prototypes with high within-head similarity while preserving independent prototype coverage for tail classes. (d) Prototype reallocation identifies under-covered tail regions and reallocates redundant capacity from head classes to activate new tail prototypes, yielding more balanced prototype coverage.

by stable semantic evidence, whereas the learnable space $Z_\phi = \{z_i(\phi)\}$ receives feedback from the evolving clustering structure. Separating these two spaces prevents the support criterion from drifting together with the current partition. The model maintains soft responsibilities $Q = \{q_{im}\}$ and an active prototype set with unknown cardinality, $C = \{c_k\}_{k=1}^{K_+}$, and ultimately outputs the hardened sample partition together with the estimated number of clusters K_+. The following sections develop a coherent prototype-capacity correction pipeline. Sec. 3.2 constrains the birth locations of candidate prototypes using fixed denoising support; Sec. 3.3 lifts base measure to a DP reference law and derives its Gibbs-interacting configuration prior; Sec. 3.4 derives an optimizable structural objective from the resulting posterior; Sec. 3.5 performs capacity reallocation through under-coverage proposals and cardinality-preserving prototype exchange; and Sec. 3.6 further stabilizes the corrected representation using prototype-balanced feedback. Intermediate derivations, full probabilistic formulations, and optimization pseudocode are provided in Appendix A and Appendix B.

3.2 Denoising Support

Nonparametric clustering first requires specifying where new prototypes can emerge. In long-tailed settings, a suitable base measure should satisfy two requirements. First, it should be support-sensitive, assigning lower prior mass to candidate directions that deviate substantially from the structure of the observed representations. Second, it should be frequency-saturating: once a candidate direction receives a minimum level of representational support, its birth advantage should not continue to increase merely because more samples lie in its neighborhood. A standard uniform hyperspherical measure fails to provide support sensitivity, as it cannot distinguish semantically supported regions from directions induced by random noise or optimization drift. An empirical density measure, by contrast, fails to provide frequency saturation, since its prior mass keeps increasing with the local sample count, making dense head modes more likely to attract additional prototypes. To this end, we construct a support-aware base measure induced by denoising recovery consistency. We train a directional denoiser D_ψ on the fixed base features U, and measure the support of a candidate

direction by whether it can return to its neighborhood after a local perturbation:

$$s_\psi(c) = \mathbb{E}_\epsilon[\cos(D_\psi(\mathrm{norm}(c + \sigma\epsilon)), c)], \qquad b_\psi(c) = [\tau_s - s_\psi(c)]_+^2. \tag{2}$$

Here, $c \in \mathbb{S}^{d-1}$ denotes a candidate prototype direction, $\epsilon \sim \mathcal{N}(0, I_d)$ is a d-dimensional Gaussian perturbation, $\sigma > 0$ controls the perturbation strength, τ_s is the minimum support threshold, and $[a]_+ = \max(a, 0)$. $s_\psi(c)$ is the local denoising recovery consistency score, while $b_\psi(c)$ defines a one-sided support barrier.

Given the support-violation cost $b_\psi(c)$ produced by the denoiser, the goal of this module is to transform it into a probability distribution defined on the unit hypersphere $\mathbb{S}^{d-1}$, such that directions with sufficient semantic support are more likely to be selected as new prototypes. We refer to this distribution as the support-aware birth base measure, denoted by H_ψ. Rather than directly assuming a prototype density, we define the denoising-support base measure as the solution to the following KL-regularized variational problem:

$$H_\psi = \arg\min_{\nu \ll \mu_\mathbb{S}} \{\mathrm{KL}(\nu \| \mu_\mathbb{S}) + \lambda_s \mathbb{E}_{c\sim\nu}[b_\psi(c)]\}. \tag{3}$$

Here, $\mu_\mathbb{S}$ denotes the uniform probability measure on the unit hypersphere, and $\nu \ll \mu_\mathbb{S}$ indicates that the candidate distribution ν is absolutely continuous with respect to $\mu_\mathbb{S}$. The parameter $\lambda_s \geq 0$ controls the support-violation penalty. We use ν to denote a probability measure in order to distinguish it from the responsibility matrix $Q = \{q_{im}\}$. The unique solution to the strictly convex problem above is an exponential tilt of the uniform hyperspherical measure:

$$H_\psi(dc) = h_\psi(c)\mu_\mathbb{S}(dc), \qquad h_\psi(c) = \frac{\exp[-\lambda_s b_\psi(c)]}{Z_\psi}. \tag{4}$$

Here, $h_\psi = dH_\psi/d\mu_\mathbb{S}$ is the density of the base measure, and Z_ψ is the normalization constant ensuring that the measure integrates to one. Consequently, unsupported directions receive lower birth probability, whereas all directions exceeding the support threshold incur the same zero barrier cost. Details of denoiser training, the variational derivation, and the normalization proof are provided in Appendix A.1 and B.6.

3.3 Gibbs-Tilted Dirichlet-Process Prior

The denoising-based support mechanism introduced in Sec. 3.2 can eliminate individual spurious candidates, but it cannot prevent multiple plausible prototypes from redundantly explaining the same high-density region. This limitation follows from the ordinary Dirichlet-process construction itself: conditional on the induced partition, its distinct atoms are independent draws from H_ψ, so the prior contains no interaction that distinguishes a dispersed prototype configuration from a geometrically redundant one. Rather than appending a repulsion loss after specifying this model, we treat the DP-induced law as a reference distribution and derive the closest interacting law whose expected configuration redundancy is controlled. Let $\Pi_N = \{B_k\}_{k=1}^{K_+}$ be the partition induced by $G \sim \mathrm{DP}(\alpha, H_\psi)$. Conditional on Π_N, the distinct prototype atoms $C = (c_1, \ldots, c_{K_+})$ are independently drawn from H_ψ. The resulting finite-sample DP reference law P_0 has density

$$p_0(\Pi_N, C \mid \psi) = p_{\mathrm{CRP}}(\Pi_N \mid \alpha) \prod_{k=1}^{K_+} h_\psi(c_k). \tag{5}$$

This law provides adaptive cardinality and support-aware atom locations, but its product form leaves the joint geometry of the atoms unrestricted. For an active prototype configuration C, we define the spherical similarity kernel and its associated redundancy energy as

$$L_C[k, \ell] = \exp\{\rho(c_k^\top c_\ell - 1)\}, \qquad \bar{L}_C = \frac{L_C + \varepsilon I}{1 + \varepsilon}, \qquad \mathcal{E}_{\mathrm{rep}}(C) = -\log\det(\bar{L}_C). \tag{6}$$

$\rho > 0$ controls sensitivity to angular separation, $\varepsilon > 0$ ensures positive definiteness, and I_{K_+} is the $K_+ \times K_+$ identity matrix. Among all laws absolutely continuous with respect to P_0, we define

$$P_\mathrm{G} = \arg\min_{P \ll P_0} \{\mathrm{KL}(P \| P_0) + \lambda_g \mathbb{E}_P[\mathcal{E}_{\mathrm{rep}}(C)]\}. \tag{7}$$

Equivalently, P_G is the minimum-information modification of the DP reference law under an upper bound on expected redundancy, with $\lambda_g \geq 0$ acting as the Lagrange multiplier. The unique solution of equation 7 is the Gibbs change of measure

$$\frac{dP_\mathrm{G}}{dP_0}(\Pi_N, C) = \frac{\exp[-\lambda_g \mathcal{E}_{\mathrm{rep}}(C)]}{\mathcal{Z}_{G,N}} = \frac{\det(\bar{L}_C)^{\lambda_g}}{\mathcal{Z}_{G,N}}, \tag{8}$$

$$p_{\mathrm{G}}(\Pi_N, C \mid \psi) = \frac{1}{\mathcal{Z}_{G,N}} p_{\mathrm{CRP}}(\Pi_N \mid \alpha) \prod_{k=1}^{K_+} h_\psi(c_k) \det(\bar{L}_C)^{\lambda_g}. \tag{9}$$

Taking the negative log of this generative prior necessarily produces $\lambda_g \mathcal{E}_{\mathrm{rep}}(C)$, which is therefore inherited by the MAP objective in Sec. 3.4 rather than added as an external regularizer. Details of the variational identity, the generative construction, and well-posedness are provided in Appendix A.2 and Appendix B.7.

3.4 Prototype Activation

Sec. 3.2 and Sec. 3.3 specify the credible locations and the ensemble geometry of prototypes, respectively, but the trade-off between fitting gain and capacity cost still needs to be compared on a common scale. For a fixed adapter ϕ, the soft responsibility over active prototypes is defined as

$$q_{im} = \frac{\pi_m \exp(\kappa z_i(\phi)^\top \tilde{c}_m)}{\sum_{\ell \in \mathcal{I}_r} \pi_\ell \exp(\kappa z_i(\phi)^\top \tilde{c}_\ell)}, \qquad m \in \mathcal{I}_r. \tag{10}$$

Here, $\pi_m \geq 0$ denotes the mixture weight satisfying $\sum_{m \in \mathcal{I}_r} \pi_m = 1$, $\kappa > 0$ is the concentration parameter of the vMF likelihood, and q_{im} denotes the posterior responsibility that sample i is assigned to active prototype m, with $\sum_{m \in \mathcal{I}_r} q_{im} = 1$. For brevity, we write $z_i = z_i(\phi)$ in what follows.

Combining the vMF fitting term, the cardinality prior, the support-aware base measure, and Gibbs repulsion yields the following finite MAP surrogate:

$$\begin{aligned} \mathcal{F}_{\mathrm{struct}} = & -\sum_i \sum_{m \in \mathcal{I}_r} q_{im} \log \pi_m - \kappa \sum_i \sum_{m \in \mathcal{I}_r} q_{im} z_i^\top \tilde{c}_m + \sum_i \sum_{m \in \mathcal{I}_r} q_{im} \log q_{im} \\ & + \lambda_a^{\mathrm{eff}} \|r\|_0 + \lambda_s \sum_m r_m b_\psi(\tilde{c}_m) + \lambda_g \mathcal{E}_{\mathrm{rep}}(C_r). \end{aligned} \tag{11}$$

Here, $C_r = \tilde{c}_m \in \mathcal{I}_r$, and λ_a^{eff} denotes the effective activation cost after absorbing the normalization contribution of the base measure. The first three lines comprise the sample–prototype fitting terms and the responsibility entropy, while the remaining three terms respectively penalize unnecessary capacity, unsupported prototype locations, and redundant prototype configurations.

Let $G_{\mathrm{fit}}(m; \phi)$ denote the decrease in fitting energy obtained by temporarily activating candidate m and re-optimizing the local responsibilities, and let $\Delta_{\mathrm{rep}}(m; \phi)$ denote the corresponding increase in repulsive energy before and after activation. candidate m is independently activated only if

$$G_{\mathrm{fit}}(m; \phi) > \lambda_a^{\mathrm{eff}} + \lambda_s b_\psi(\tilde{c}_m) + \lambda_g \Delta_{\mathrm{rep}}(m; \phi). \tag{12}$$

Birth, pruning, and the subsequent swap operation therefore reduce the same structural objective, rather than acting as mutually independent heuristic rules. Details of the conditional posterior and the objective transformation are provided in Appendix A.3 and Appendix A.4.

3.5 Under-Coverage Candidate Generation

Sec. 3.3 uses the Gibbs repulsion term to release redundant prototype capacity from head regions. This section further specifies how the released capacity is reallocated. We use an absolute coverage residual to identify regions that are supported by the fixed representation but remain insufficiently explained by the current structure, and propose new candidates within their local neighborhoods:

$$d_i(C_r) = 1 - \max_{k \in \mathcal{I}_r} z_i(\phi)^\top \tilde{c}_k, \omega_i^{\mathrm{uc}} = [d_i(C_r) - \tau_u]_+ \mathbb{I}[s_\psi(u_i) \geq \tau_s], \tilde{c}_m^{\mathrm{prop}} = \mathrm{norm}\left(\sum_{i \in \mathcal{N}_m} \omega_i^{\mathrm{uc}} z_i(\phi) \right). \tag{13}$$

Here, $d_i(C_r)$ is the cosine coverage residual between sample i and its best-matching active prototype, τ_u is the under-coverage threshold, $\mathbb{I}[\cdot]$ denotes the indicator function, $\mathcal{N}_m$ is the local neighborhood associated with the m-th candidate, and ω_i^{uc} is the corresponding neighborhood aggregation weight. For an under-covered candidate m that fails the independent birth criterion, we allow it to replace an inefficient active prototype k:

$$r^{k \to m} = r - e_k + e_m, \qquad \Delta_{\mathrm{swap}}(k, m; \phi) = \mathcal{F}^\star_{\mathrm{struct}}(r^{k \to m}; \phi) - \mathcal{F}^\star_{\mathrm{struct}}(r; \phi), \tag{14}$$

Table 1: Clustering performance (%) at imbalance ratio = 5. Higher is better. Best results are bold.

Method	Venue	CIFAR-10-LT				CIFAR-20-LT				STL-10-LT			
		ACC	CAA	NMI	ARI	ACC	CAA	NMI	ARI	ACC	CAA	NMI	ARI
BYOL	NeurIPS'20	60.4	66.4	62.0	46.5	38.5	39.2	41.3	23.4	48.5	46.4	49.2	35.4
CoNR	NeurIPS'23	41.4	43.5	34.9	23.3	23.8	23.4	21.6	12.1	32.4	31.3	29.0	18.2
ProPos	TPAMI'23	51.4	59.2	52.2	34.1	40.7	39.3	42.6	26.4	37.2	39.3	34.5	21.8
DMICC	AAAI'23	40.6	42.5	36.9	25.4	25.2	23.1	20.6	7.5	45.9	45.6	45.3	35.5
SeCu	ICCV'23	56.1	59.0	54.5	40.3	34.8	33.7	35.8	19.9	51.6	50.8	46.0	39.1
LFSS	ICML'25	58.1	63.1	59.7	43.3	43.5	42.2	46.8	28.6	46.6	43.7	45.3	34.0
ConMix	ICLR'25	61.6	65.4	59.8	45.6	42.8	41.0	43.9	27.7	48.9	49.7	48.8	35.9
DGD	CVPR'26	66.2	68.4	61.2	55.6	43.8	41.5	45.8	28.3	51.3	50.4	50.8	38.4
MiniClustering	ICLR'26	74.3	72.6	**69.9**	62.1	45.6	42.3	47.5	30.7	54.7	52.7	54.4	42.8
DRP	**Ours**	**78.8**	**76.2**	**69.9**	**63.2**	**48.3**	**44.4**	**49.3**	**33.2**	**57.6**	**54.5**	**55.0**	**44.1**

Table 2: Clustering performance (%) at imbalance ratio = 10. Higher is better. Best results are bold.

Method	Venue	CIFAR-10-LT				CIFAR-20-LT				STL-10-LT			
		ACC	CAA	NMI	ARI	ACC	CAA	NMI	ARI	ACC	CAA	NMI	ARI
BYOL	NeurIPS'20	51.9	55.2	56.3	41.7	39.8	37.1	41.9	25.0	41.2	41.8	46.8	29.6
CoNR	NeurIPS'23	31.4	44.3	29.2	17.8	20.3	17.9	17.3	8.0	34.8	32.5	30.7	21.0
ProPos	TPAMI'23	46.1	49.3	52.5	34.2	36.8	33.6	40.1	22.5	35.6	38.2	37.2	23.9
DMICC	AAAI'23	36.6	39.5	36.8	25.9	24.7	21.9	20.7	10.1	41.3	41.9	38.7	30.3
SeCu	ICCV'23	51.5	57.0	54.8	39.6	32.0	30.7	33.3	17.6	48.2	49.6	50.0	36.8
LFSS	ICML'25	56.3	59.7	57.9	43.0	41.1	40.0	46.8	28.4	43.3	42.3	45.4	30.8
ConMix	ICLR'25	53.3	58.2	57.1	40.8	41.7	39.3	43.6	27.0	47.4	48.7	48.2	33.9
DGD	CVPR'26	56.7	59.2	58.1	47.2	42.6	40.3	46.2	28.0	50.4	50.8	49.0	35.4
MiniClustering	ICLR'26	64.6	61.4	63.9	56.7	44.3	41.5	47.2	30.4	52.0	**54.4**	**50.8**	39.1
DRP	**Ours**	**67.4**	**63.9**	**65.7**	**61.1**	**46.5**	**43.2**	**47.8**	**31.5**	**54.2**	53.8	50.4	**40.2**

where e_k and e_m are the standard basis vectors associated with the corresponding coordinates, and $\mathcal{F}^{\star}_{\text{struct}}$ denotes the profiled objective obtained after re-optimizing Q and π. Among all candidate exchanges, we accept only the best swap satisfying $\Delta_{\text{swap}} < 0$. Because $\|r^{k\to m}\|_0 = \|r\|_0$, the fixed cardinality cost cancels exactly in the swap comparison. The expanded form of the swap criterion and the details of candidate screening are provided in Appendix A.5.

3.6 Prototype-Balanced Adaptation

After capacity reallocation, updating the representation by summing over samples would again allow large prototypes to produce stronger gradients, potentially overwhelming the newly established small-scale coverage. Let $n_m = \sum_i q_{im}$ denote the responsibility mass of prototype m, and let $\mathcal{I}_s = \{m : r_m = 1, n_m \geq \tau_n\}$ denote the set of stable active prototypes. We first normalize the within-prototype contribution by its responsibility mass and then average equally across prototypes:

$$\mathcal{L}_{\text{pbal}}(\phi) = -\frac{1}{|\mathcal{I}_s|} \sum_{m\in\mathcal{I}_s} \frac{\sum_i \text{sg}(q_{im}) z_i(\phi)^\top \text{sg}(\tilde{c}_m)}{\text{sg}(n_m) + \epsilon_q}. \tag{15}$$

Here, $\text{sg}(\cdot)$ denotes the stop-gradient operator, and $\epsilon_q > 0$ prevents division by zero. To prevent the adaptation process from drifting away from the base semantics or collapsing across samples, we optimize the adapter with

$$\mathcal{L}_{\text{adapt}} = \mathcal{L}_{\text{pbal}} + \lambda_{\text{anc}}\mathcal{L}_{\text{anc}} + \lambda_{\text{inv}}\mathcal{L}_{\text{view}} + \lambda_{\text{var}}\mathcal{L}_{\text{var}} + \lambda_{\text{cov}}\mathcal{L}_{\text{cov}}. \tag{16}$$

Here, $\mathcal{L}_{\text{anc}}$ anchors the adapted representations to the base features, $\mathcal{L}_{\text{view}}$ enforces cross-view consistency, and $\mathcal{L}_{\text{var}}$ and $\mathcal{L}_{\text{cov}}$ respectively preserve feature variance and suppress dimensional redundancy. The corresponding λ terms are nonnegative weighting coefficients. The complete regularization terms, spherical prototype updates are provided in Appendix A.6.

4 Experiments

4.1 Experiment Settings

Datasets and Evaluation Metrics. We evaluate on long-tailed versions of CIFAR-10, CIFAR-20, STL-10, and Tiny ImageNet, as well as ImageNet-LT. Following standard long-tail construction protocols (Cao et al., 2019; Zhou et al., 2020), the imbalance ratio (IR) is the ratio between the

Table 3: Tiny ImageNet-LT clustering results (%) at an imbalance ratio of 10.

Method	Venue	ACC	CAA	NMI	ARI
BYOL	NeurIPS'20	16.1	14.0	34.2	8.6
CoNR	NeurIPS'23	8.8	8.2	28.2	3.9
ProPos	TPAMI'23	15.0	13.2	34.4	7.5
DMICC	AAAI'23	9.27	8.8	28.0	3.7
SeCu	ICCV'23	11.5	8.7	39.7	8.5
LFSS	ICML'25	22.9	18.4	28.7	14.0
ConMix	ICLR'25	16.5	14.5	34.9	8.3
DGD	CVPR'26	19.4	17.1	36.7	10.1
MiniClustering	ICLR'26	25.5	19.4	40.6	14.9
DRP	**Ours**	**27.1**	**22.3**	**43.4**	**17.3**

Table 4: ImageNet-LT clustering results (%) on the official long-tailed split.

Method	Venue	ACC	CAA	NMI	ARI
BYOL	NeurIPS'20	14.8	10.9	50.6	10.2
CoNR	NeurIPS'23	6.38	6.85	38.9	2.19
DMICC	AAAI'23	5.24	5.94	37.9	1.83
ConMix	ICLR'25	15.4	12.2	51.6	11.4
DGD	CVPR'26	15.2	12.5	51.9	11.4
MiniClustering	ICLR'26	15.7	12.8	**52.3**	11.7
DRP	**Ours**	**16.9**	**13.8**	51.9	**12.3**

Table 5: Balanced ImageNet clustering (%). Higher is better; "–" denotes unavailable results.

Method	Venue	ImageNet-50			ImageNet-100			ImageNet-200			ImageNet-1K		
		ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
SIC	AAAI'23	75.4	81.4	67.1	67.2	77.2	55.8	57.9	73.4	46.4	47.0	77.2	34.3
TEMI	BMVC'23	80.1	86.1	71.0	75.0	85.6	65.4	73.1	85.2	62.1	34.6	63.2	26.6
TAC	ICML'24	78.7	83.7	67.5	63.8	68.1	57.6	59.2	70.9	50.8	58.2	79.9	43.5
CLUDI	ICML'25	82.1	91.2	76.2	76.6	87.1	67.8	73.7	86.1	63.2	45.5	67.0	34.8
CAE	AAAI'26	–	–	–	–	–	–	–	–	–	53.1	79.3	39.3
DRP	**Ours**	**84.3**	**91.8**	**79.7**	**78.1**	**89.0**	**71.2**	**74.8**	**87.6**	**66.3**	**59.3**	**81.7**	**44.7**

largest and smallest class sizes. To evaluate the generalization ability on balanced data, we further use large-scale balanced datasets such as ImageNet-1K. Following MiniClustering (Li & Jia, 2025; Li et al., 2026), We report clustering accuracy (ACC), class-average accuracy (CAA), normalized mutual information (NMI), and adjusted rand index (ARI).

Implementation Details and Baselines. Following ConMix and MiniClustering (Li & Jia, 2025; Li et al., 2026), long-tailed experiments use a ResNet-18 encoder pretrained with BYOL (Grill et al., 2020); the base encoder is then frozen for DRP. Training schedules follow the corresponding baselines, with batch size 512 and random seed 42. Balanced ImageNet experiments use a frozen CLIP ViT-B/16 image encoder for all compared variants so that representation quality is controlled. We compare with BYOL (Grill et al., 2020), CoNR (Yu et al., 2023), ProPos (Huang et al., 2023), DMICC (Li et al., 2023), SeCu (Qian, 2023), LFSS (Li et al., 2025), ConMix (Li & Jia, 2025), DGD (Ye et al., 2026), and MiniClustering (Li et al., 2026). Balanced-data baselines are SIC (Cai et al., 2023), TEMI (Adaloglou et al., 2023), TAC (Li et al., 2024), CLUDI (Uziel et al., 2025), and CAE (Zhu et al., 2026).

4.2 Experiments Results.

Long-Tailed Clustering Quality. We conduct experiments on multiple long-tailed datasets and compare DRP with strong state-of-the-art baselines. The main results are summarized in Table 1–4. When the imbalance ratio is set to 5, DRP obtains the best or tied-best result in all 12 dataset. When the imbalance ratio is set to 10, DRP remains best in 10 of the 12 comparisons. The concurrent improvements in ACC, CAA, and ARI suggest that the gains are not confined to sample-rich classes: DRP improves overall assignment quality while producing more class-balanced and structurally consistent partitions under long-tailed distributions.

Generalization To Balanced Data. Furthermore, we evaluate the generalization ability of DRP on several large-scale balanced datasets, with the results reported in Table 5. DRP achieves the best result in all 4 dataset from ImageNet-50 to ImageNet-1K. Compared with the strongest competing result for each entry, its improvements range from 1.1 to 2.2 points in ACC, 0.6 to 1.9 points in NMI, and 1.2 to 3.5 points in ARI. This consistent advantage across different class scales shows that the prototype-allocation mechanism does not depend on skewed class frequencies and remains effective as the number of balanced classes increases.

Partition Quality With Estimated-K. We also report the partition quality when the true number of classes is unknown, as shown in Table 6. On the smaller benchmarks, the estimated cardinality differs from the target by at most three clusters and exactly recovers the 10 classes of STL-10-LT

Table 6: Comparison of the cluster number inferred by our framework versus the Ground Truth number of classes. "Balanced" denotes datasets without an imbalance ratio.

Method	CIFAR-10		CIFAR-20		STL-10		Tiny ImageNet	ImageNet-LT	ImageNet-50	ImageNet-100	ImageNet-200	ImageNet-1K
	IR = 5	IR = 10	IR = 5	IR = 10	IR = 5	IR = 10	IR = 10	IR = 256	Balanced	Balanced	Balanced	Balanced
Predicted	13.0	11.7	22.0	19.3	10.0	9.3	189.7	1020.0	52.6	96.2	210.0	916.4
Ground truth	10	10	20	20	10	10	200	1000	50	100	200	1000

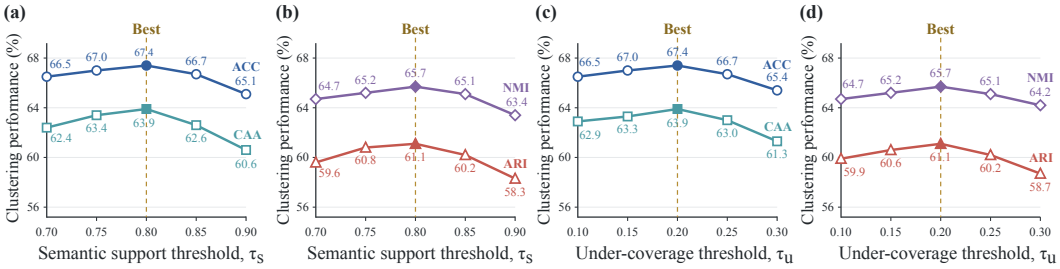


Figure 3: Sensitivity to the semantic-support threshold τ_s and the under-coverage threshold τ_u.

at IR$=5$. For datasets containing 50–1,000 classes, the relative deviation remains within 8.4%. The number of clusters inferred by DRP closely matches the ground-truth value, demonstrating its ability to infer an unknown number of classes.

4.3 Ablation and Sensitivity Analysis

Table 7: Component ablation.

Model	CIFAR-10-LT		Tiny ImageNet-LT		ImageNet-50	
	ACC	NMI	ACC	NMI	ACC	NMI
Baseline	60.4	59.8	21.8	37.9	78.7	85.1
w/o denoising support	66.1	64.7	25.9	41.7	82.9	89.7
w/o Gibbs repulsion	65.5	63.8	25.4	41.2	82.5	88.6
w/o under-coverage candidate	64.9	63.9	25.1	41.5	83.5	90.6
w/o prototype-balanced adaptation	65.8	65.1	25.8	42.2	80.9	87.2
DRP (full)	**67.4**	**65.7**	**27.1**	**43.4**	**84.3**	**91.8**

We conduct ablation studies to evaluate the effectiveness of the key components. As shown in Table 7, removing any component leads to a noticeable performance drop, confirming the necessity of each module. We further perform sensitivity analyses of two key hyperparameters: the fixed semantic-support threshold τ_s and the under-coverage threshold τ_u. As shown in Figure 3 (a) and (b), DRP remains stable under moderate values of the semantic-support threshold. As candidate directions lacking support are gradually filtered out, clustering performance improves and reaches its best observed value at $\tau_s = 0.80$. When the threshold is increased further, performance declines noticeably, with a more pronounced degradation in CAA. This indicates that overly strict support filtering preferentially harms independent prototype coverage in low-frequency semantic regions. As shown in Figure 3 (c) and (d), DRP maintains relatively stable performance over a moderate range of under-coverage thresholds. Smaller thresholds generate more candidate proposals, but support constraints and the energy-based acceptance criterion can partially suppress the adverse effects of invalid candidates. In contrast, excessively large thresholds suppress valid under-coverage candidates, resulting in more pronounced declines in CAA and ARI. Appendix C.3 additionally reports sensitivity results for $K_{\max}$ and $\kappa_{\max}$, visualizations of the learned representations.

5 Conclusion

We propose DRP, a nonparametric long-tailed deep clustering framework leveraging denoising support, repulsive point processes and conservative representation adaptation. Integrating a support-aware base measure, Gibbs repulsion and cardinality-preserving prototype exchange under a unified MAP objective, DRP suppresses head redundancy and enlarges tail coverage. A residual adapter receives balanced prototype feedback, with anti-collapse regularization limiting representation drift. Experiments confirm DRP improves clustering performance and infers unknown class counts.

AI USE STATEMENT

Large Language Models (LLMs) were used in this research only for writing optimization, grammar checking, and limited assistance with simple code debugging. For code debugging, the LLMs helped identify minor syntax errors, runtime errors, and simple implementation bugs. These tools also gave feedback on how clearly, consistently, and effectively the methodological and experimental descriptions were presented. No part of the theoretical contributions, experimental design, data analysis, or results was generated by LLMs.

ETHICS STATEMENT

This work studies unsupervised image clustering using established image benchmarks and pretrained representations. It does not involve recruiting human participants or collecting new personal data. Although DRP aims to improve the representation of low-frequency classes, better class-balanced clustering performance does not imply fairness across demographic groups. Biases in the datasets and pretrained encoders may persist in the resulting clusters. Applications involving people or sensitive data therefore require additional privacy safeguards, bias assessment, and human oversight.

REPRODUCIBILITY STATEMENT

The Method section presents the mathematical formulation and optimization objectives of DRP. Appendix A provides additional derivations, optimization details, and algorithm pseudocode, while Appendix B states the assumptions and proofs underlying the theoretical analyses. The experimental setup describes the datasets, evaluation metrics, encoder configurations, and baseline protocols. Appendix C documents the construction of the long-tailed splits, exact sample statistics, hyperparameter sensitivity analyses, and computational complexity. Appendix D.2 reports results across five random seeds to assess sensitivity to random initialization. We plan to publicly release the source code to facilitate independent reproduction of our results.

REFERENCES

Nikolas Adaloglou, Felix Michels, Hamza Kalisch, and Markus Kollmann. Exploring the limits of deep image clustering using pretrained models, 2023. URL https://arxiv.org/abs/2303.17896.

Charles E. Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/2958336.

David M. Blei and Michael I. Jordan. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121 – 143, 2006. doi: 10.1214/06-BA104. URL https://doi.org/10.1214/06-BA104.

Shaotian Cai, Liping Qiu, Xiaojun Chen, Qin Zhang, and Longteng Chen. Semantic-enhanced image clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 6869–6878, 2023.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/621461af90cadfdaf0e8d4cc25129f91-Paper.pdf.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features, 2019. URL https://arxiv.org/abs/1807.05520.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments, 2021. URL https://arxiv.org/abs/2006.09882.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Thomas S. Ferguson. A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/2958008.

Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels, 2020. URL https://arxiv.org/abs/2005.12320.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Dong Huang, Ding-Hua Chen, Xiangji Chen, Chang-Dong Wang, and Jian-Huang Lai. Deepclue: Enhanced deep clustering via multi-layer ensembles in neural networks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8(2):1582–1594, 2024. doi: 10.1109/TETCI.2024.3353598.

Zhizhong Huang, Jie Chen, Junping Zhang, and Hongming Shan. Learning representation for clustering via prototype scattering and positive sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7509–7524, 2023. doi: 10.1109/TPAMI.2022.3216454.

Xu Ji, João F. Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation, 2019. URL https://arxiv.org/abs/1807.06653.

Ziyu Jiang, Tianlong Chen, Bobak J Mortazavi, and Zhangyang Wang. Self-damaging contrastive learning. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4927–4939. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/jiang21a.html.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2-3):123–286, December 2012. ISSN 1935-8245. doi: 10.1561/2200000044. URL http://dx.doi.org/10.1561/2200000044.

Hongyu Li, Lefei Zhang, and Kehua Su. Dual mutual information constraints for discriminative clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 8571–8579, 2023.

Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. Contrastive clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 8547–8555, 2021.

Yunfan Li, Peng Hu, Dezhong Peng, Jiancheng Lv, Jianping Fan, and Xi Peng. Image clustering with external guidance. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 27890–27902. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/li24aa.html.

Zhixin Li and Yuheng Jia. Conmix: Contrastive mixup at representation level for long-tailed deep clustering. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=3lH8WT0fhu.

Zhixin Li, Yuheng Jia, Hui Liu, and Junhui Hou. Learning from sample stability for deep clustering. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 34904–34919. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/li25am.html.

Zhixin Li, Yuheng Jia, Guanliang Chen, Hui LIU, and Junhui Hou. Mini-cluster guided long-tailed deep clustering. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 79896–79916, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/813ed01a11220fcdeed10e0d7da2de25-Paper-Conference.pdf.

Ioannis Maniadis Metaxas, Georgios Tzimiropoulos, and Ioannis Patras. Divclust: Controlling diversity in deep clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3418–3428, June 2023.

Qi Qian. Stable cluster discrimination for deep clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16645–16654, October 2023.

Meitar Ronen, Shahaf E. Finder, and Oren Freifeld. Deepdpm: Deep clustering with an unknown number of clusters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9861–9870, June 2022.

Yaling Tao, Kentaro Takagi, and Kouta Nakata. Clustering-friendly representation learning via instance discrimination and feature decorrelation. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=e12NDM7wkEY.

Roy Uziel, Irit Chelly, Oren Freifeld, and Ari Pakman. Clustering via self-supervised diffusion. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 60711–60726. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/uziel25a.html.

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 (CVPR)*, pp. 7492–7501, June 2022.

Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 478–487, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/xieb16.html.

Xulun Ye, Yuanyuan Deng, and Kun Zhou. Dgd: Density gradient-guided diffusion for long-tailed clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 7654–7664, June 2026.

Chunlin Yu, Ye Shi, and Jingya Wang. Contextually affinitive neighborhood refinery for deep clustering. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 5778–5790. Curran Associates, Inc., 2023. doi: 10.52202/075280-0253. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/123cfe7d8b7702ac97aaf4468fc05fa5-Paper-Conference.pdf.

Xiaohang Zhan, Jiahao Xie, Ziwei Liu, Yew-Soon Ong, and Chen Change Loy. Online deep clustering for unsupervised representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6687–6696, 2020. doi: 10.1109/CVPR42600.2020.00672.

Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Xingyu Zhu, Beier Zhu, Yunfan Li, Junfeng Fang, Shuo Wang, Kesen Zhao, and Hanwang Zhang. Hierarchical semantic alignment for image clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 29177–29185, 2026.

APPENDIX CONTENTS

A Methodological and Optimization Details **15**

A.1 Denoising Support and Base-Measure Construction 15

A.2 DP Reference Distribution and Gibbs Tilting 16

A.3 Hyperspherical Likelihood and Unified Posterior 18

A.4 Unified Structural Objective and Prototype Activation 19

A.5 Prototype Capacity Reallocation 21

A.6 Adaptation Regularization and Spherical Prototype Updates 22

B Mechanism Analysis and Proofs **23**

B.1 Support Control 23

B.2 Repulsion Against Prototype Collapse 24

B.3 Coordinate-Wise Activation Criterion 24

B.4 Cardinality-Preserving Prototype Exchange 25

B.5 Representation Drift Bound under Anchoring 25

B.6 Variational Derivation of the Denoising-Supported Base Measure 26

B.7 Well-Posedness of the Gibbs Tilt 26

B.8 Remarks on the Finite-Cardinality Penalty 26

C Additional Experimental Details **27**

C.1 Long-Tailed Dataset Construction and Exact Sample Statistics 27

C.2 Visualization of Learned Representations 29

C.3 Hyperparameter Sensitivity Analysis 31

C.4 Computational Complexity Analysis 31

D Additional Experiments **32**

D.1 Additional Baselines and Comparison Results 32

D.2 Multi-Seed Experiments 32

D.3 Experiments on Different Backbone Networks 32

APPENDIX

A METHODOLOGICAL AND OPTIMIZATION DETAILS

This section provides additional methodological and optimization details omitted from the main text due to space constraints, complementing the description of the proposed framework in the main paper. The formal proposition and its proof are provided in Appendix B.

A.1 DENOISING SUPPORT AND BASE-MEASURE CONSTRUCTION

For a fixed noise scale $\sigma > 0$, define the perturbation operator

$$u_\sigma(u, \epsilon) = \mathrm{norm}(u + \sigma\epsilon), \qquad \epsilon \sim \mathcal{N}(0, I). \tag{17}$$

where $u \in \mathbb{S}^{d-1}$, $\epsilon \sim \mathcal{N}(0, I_d)$, and I_d denotes the d-dimensional identity matrix.

The denoiser is trained only on the fixed base features U:

$$\begin{aligned} \psi_0 &= \arg\min_{\psi} \mathbb{E}_{u\sim\widehat{P}_U,\epsilon} \left[1 - \cos(D_\psi(u_\sigma(u, \epsilon)), u)\right], \\ \widehat{P}_U &= N^{-1} \sum_i \delta_{u_i}. \end{aligned} \tag{18}$$

Here, $\widehat{P}_U$ is the empirical distribution of the base features, and δ_{u_i} denotes the Dirac measure concentrated at u_i. Under the spherical directional loss, the ideal denoiser satisfies

$$D^*(\tilde{u}) = \mathrm{norm}(\mathbb{E}[u \mid \tilde{u}]).$$

Therefore, whether a locally perturbed candidate can be mapped back to its original neighborhood provides a proxy for local representational support. After training, D_{ψ_0} is frozen, and we abbreviate the subscript as ψ in the main text.

To clarify the relation between Eq. equation 3 and Eq. equation 4 in the main text, let $q_\nu(c) = d\nu/d\mu_{\mathbb{S}}$ denote the Radon–Nikodym density of the candidate distribution; this quantity is unrelated to the sample responsibilities q_{im}. The objective can be written as

$$\begin{aligned} &\int_{\mathbb{S}^{d-1}} q_\nu(c) \log q_\nu(c)\, d\mu_{\mathbb{S}}(c) \\ &\quad + \lambda_s \int_{\mathbb{S}^{d-1}} q_\nu(c) b_\psi(c)\, d\mu_{\mathbb{S}}(c), \end{aligned} \tag{19}$$

subject to $q_\nu \geq 0$ and $\int q_\nu\, d\mu_{\mathbb{S}} = 1$. Introducing a Lagrange multiplier for the normalization constraint and taking the variation with respect to q_ν, the stationary condition is

$$\log q_\nu(c) + 1 + \lambda_s b_\psi(c) + \eta = 0, \tag{20}$$

which yields

$$q_\nu(c) \propto \exp[-\lambda_s b_\psi(c)].$$

The normalization constant is

$$Z_\psi = \int_{\mathbb{S}^{d-1}} \exp[-\lambda_s b_\psi(c)]\, \mu_{\mathbb{S}}(dc). \tag{21}$$

Since $b_\psi(c)$ is finite and nonnegative, we have $0 < Z_\psi \leq 1$; moreover, the strict convexity of the KL term guarantees uniqueness of the solution. This derivation gives a precise interpretation of Eq. equation 4: H_ψ is the probability measure that is closest, in the KL sense, to the uniform hyperspherical reference measure while simultaneously penalizing the expected violation of denoising support, rather than an exact estimate of the data density. Figure 4 shows that the denoising support can serve as valid evidence for the birth location of new prototypes. The one-sided barrier only ensures that no additional reward is granted once the support threshold is exceeded, thereby preventing prototypes from being continually pulled toward the highest-density head modes. A more complete variational derivation is provided in Sec. B.6.

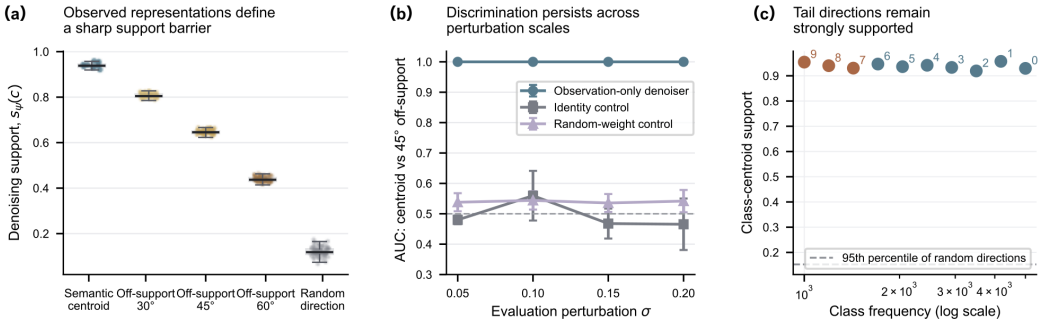


Figure 4: Denoising-support diagnostics. Denoising consistency is used as a one-sided birth constraint rather than as a continuously increasing reward for high-density directions. As shown in (a), the denoising support score decreases progressively as candidate directions deviate from the center of the observed representations, reaching the lowest level in random directions, indicating that this score captures the local compatibility between candidate directions and the observed representation structure. (b) further shows that this distinction between supported and unsupported regions remains stable across different perturbation scales. The results in (c) demonstrate that low-frequency class directions still obtain significantly higher support than random directions, and the support score does not exhibit a notable systematic decline with class frequency.

A.2 DP Reference Distribution and Gibbs Tilting

From H_ψ to the DP Reference Distribution. The measure H_ψ defined in Eq. equation 4 is now used as the base probability measure of a Dirichlet process:

$$G \sim \mathrm{DP}(\alpha, H_\psi), \qquad \theta_i \mid G \overset{\text{iid}}{\sim} G. \tag{22}$$

Here, $\alpha > 0$ is the concentration parameter, while H_ψ governs the locations on the sphere at which new atoms may arise. After marginalizing out G, ties among the latent atoms induce a random partition $\Pi_N = \{B_1, \ldots, B_{K_+}\}$ of the N samples, whose Chinese restaurant process (CRP) partition probability is

$$p_{\mathrm{CRP}}(\Pi_N \mid \alpha) = \frac{\Gamma(\alpha)\alpha^{K_+}}{\Gamma(N+\alpha)} \prod_{k=1}^{K_+} \Gamma(n_k). \tag{23}$$

Here, p_{CRP} denotes the partition probability induced by the Chinese restaurant process, $\Gamma(\cdot)$ is the Gamma function, and $n_k = |B_k|$ is the number of samples in the k-th partition block. Conditional on Π_N, its K_+ distinct atoms $C = (c_1, \ldots, c_{K_+})$ are mutually independent and identically distributed according to H_ψ. Together, the CRP partition and these distinct atoms define a finite-sample DP reference distribution P_0 over partition–prototype configurations.

To make explicit the configuration dimension that varies with K_+, let $\mathfrak{P}_{N,K}$ denote the set of partitions of $\{1, \ldots, N\}$ into K canonically ordered blocks. We then define the variable-dimensional configuration space as

$$\Omega_N = \bigsqcup_{K=1}^{N} \mathfrak{P}_{N,K} \times (\mathbb{S}^{d-1})^K. \tag{24}$$

Here, $\bigsqcup$ denotes the disjoint union of configuration spaces with different cardinalities. Accordingly, Ω_N simultaneously accommodates a variable number of partition blocks and their corresponding spherical prototype locations.

We define the following reference measure on Ω_N:

$$\nu_N(d\Pi, dC) = \sum_{K=1}^{N} \sum_{\Pi \in \mathfrak{P}_{N,K}} \delta_\Pi(d\Pi)\, \mu_{\mathbb{S}}^{\otimes K}(dC). \tag{25}$$

Here, ν_N is the reference measure on Ω_N, δ_Π is the point mass concentrated at partition Π, and $\mu_{\mathbb{S}}^{\otimes K}$ is the K-fold product of the uniform probability measure on the sphere. Moreover, $d\Pi$ and dC

denote the measure elements associated with the partition and the prototype configuration, respectively. Therefore, the density of the finite-sample DP reference distribution P_0 with respect to ν_N is

$$\begin{aligned} p_0(\Pi_N, C \mid \psi) &:= \frac{dP_0}{d\nu_N}(\Pi_N, C) \\ &= p_{\mathrm{CRP}}(\Pi_N \mid \alpha) \prod_{k=1}^{K_+} h_\psi(c_k). \end{aligned} \tag{26}$$

Here, P_0 denotes the finite-sample DP reference distribution jointly induced by the CRP partition probability and the independent base-measure atoms, while p_0 denotes its density with respect to ν_N.

Gibbs Measure Transformation. Adaptive cardinality alone cannot prevent multiple atoms from concentrating around the same head-class mode. In particular, when a head region already provides a large fitting gain, a variable-cardinality model may continue to activate multiple local prototypes to explain minor variations within the head region, rather than reserving capacity for sparse tail-semantic directions. To introduce repulsion among prototypes, we define the following hyperspherical kernel for $C = (c_1, \ldots, c_{K_+})$:

$$L_C[k, \ell] = \exp\{\rho(c_k^\top c_\ell - 1)\}, \qquad \bar{L}_C = \frac{L_C + \varepsilon I}{1 + \varepsilon}. \tag{27}$$

The corresponding repulsive energy is defined as

$$\mathcal{E}_{\mathrm{rep}}(C) = -\log\det(\bar{L}_C). \tag{28}$$

Here, $L_C \in \mathbb{R}^{K_+ \times K_+}$ is the similarity-kernel matrix of the active prototypes, $\rho > 0$ controls the interaction range of the kernel, I denotes the K_+-dimensional identity matrix, and $\varepsilon > 0$ is a jitter term introduced to ensure positive definiteness and numerical stability. Moreover, $\bar{L}_C$ is the normalized kernel matrix, $\det(\cdot)$ denotes the determinant, and $\mathcal{E}_{\mathrm{rep}}(C)$ is the repulsive energy of the prototype configuration.

When two prototype directions approach each other, the corresponding rows of the kernel matrix become nearly linearly dependent, causing the determinant to decrease and the repulsive energy to increase. Therefore, this energy suppresses geometric redundancy over the entire set of active prototypes rather than determining the number of clusters by itself. Its purpose is not to force all samples from a head class to collapse to a single point. Instead, it requires multiple nearby head prototypes to pay a higher configuration cost, thereby reducing the tendency of head regions to expand their explanatory capacity by duplicating local prototypes.

The interacting law is determined before posterior inference. Let $\mathcal{P}(P_0)$ be the set of probability measures on Ω_N that are absolutely continuous with respect to P_0. We seek the law that changes the DP reference distribution as little as possible while controlling the expected configuration energy:

$$P_{\mathrm{Gibbs}} = \operatorname*{arg\,min}_{P \in \mathcal{P}(P_0)} \left\{ \mathrm{KL}(P \| P_0) + \lambda_g \mathbb{E}_P[\mathcal{E}_{\mathrm{rep}}(C)] \right\}. \tag{29}$$

For an active redundancy budget ξ, this is the Lagrangian form of $\min_P \mathrm{KL}(P\|P_0)$ subject to $\mathbb{E}_P[\mathcal{E}_{\mathrm{rep}}(C)] \le \xi$. Thus, the DP law remains the reference semantics, and dependence is introduced only to the extent required by the set-level constraint.

The Gibbs variational identity gives the unique solution. Specifically, the prior density of P_{Gibbs} on Ω_N with respect to ν_N is

$$\begin{aligned} p^{\mathrm{prior}}_{\mathrm{Gibbs}}(\Pi_N, C \mid \psi) &:= \frac{dP_{\mathrm{Gibbs}}}{d\nu_N}(\Pi_N, C) \\ &= \frac{p_0(\Pi_N, C \mid \psi)}{\mathcal{Z}_{G,N}} \exp[-\lambda_g \mathcal{E}_{\mathrm{rep}}(C)]. \end{aligned} \tag{30}$$

Here, P_{Gibbs} is the prior probability measure obtained after imposing prototype repulsion, $p^{\mathrm{prior}}_{\mathrm{Gibbs}}$ is its density with respect to ν_N, $\lambda_g \ge 0$ controls the strength of repulsion, and $\mathcal{Z}_{G,N}$ is the Gibbs

normalizing constant (see Appendix B.7 for a proof of its well-posedness):

$$\begin{aligned}\mathcal{Z}_{G,N} &= \int_{\Omega_N} p_0(\Pi, C \mid \psi) \exp[-\lambda_g \mathcal{E}_{\mathrm{rep}}(C)]\, \nu_N(d\Pi, dC) \\ &= \mathbb{E}_{P_0}\left[\det(\bar{L}_C)^{\lambda_g}\right].\end{aligned} \tag{31}$$

For any $P \in \mathcal{P}(P_0)$, the objective in equation 29 satisfies

$$\begin{aligned}&\mathrm{KL}(P\|P_0) + \lambda_g \mathbb{E}_P[\mathcal{E}_{\mathrm{rep}}(C)] \\ &\quad = \mathrm{KL}(P\|P_{\mathrm{Gibbs}}) - \log \mathcal{Z}_{G,N}.\end{aligned} \tag{32}$$

Since relative entropy is nonnegative and vanishes only when its two arguments agree, equation 32 proves both optimality and uniqueness. Consequently, the Radon–Nikodym derivative of P_{Gibbs} with respect to P_0 is

$$\frac{dP_{\mathrm{Gibbs}}}{dP_0}(\Pi_N, C) = \frac{\det(\bar{L}_C)^{\lambda_g}}{\mathcal{Z}_{G,N}}. \tag{33}$$

Expanding p_0 gives the equivalent expression

$$\begin{aligned}p^{\mathrm{prior}}_{\mathrm{Gibbs}}(\Pi_N, C \mid \psi) = \frac{p_{\mathrm{CRP}}(\Pi_N \mid \alpha)}{\mathcal{Z}_{G,N}} \\ \times \prod_{k=1}^{K_+} h_\psi(c_k) \det(\bar{L}_C)^{\lambda_g}.\end{aligned} \tag{34}$$

Because the determinant factor jointly couples the prototype atoms, we refer to the resulting construction as a diffusion-inspired, denoising-supported, and Gibbs-tilted DP-type prototype prior. The CRP factor supplies the reference law over adaptive cardinalities and accommodates unequal cluster sizes, but is not claimed to correct long-tailed bias by itself. The denoising-supported base measure governs the validity of individual prototype directions, whereas Gibbs conditioning makes the probability of the entire configuration depend on its redundancy. Therefore, cardinality and geometry are coupled in the final prior even though their roles remain interpretable.

Together, these three components constrain head-semantic dominance. If an additional prototype within a head region yields only a limited improvement in data fitting, it must still pay both an activation cost and a redundancy cost. If the prototype is instead moved outside the head region toward an unsupported direction, it additionally incurs the support-barrier penalty.

A.3 Hyperspherical Likelihood and Unified Posterior

From the Partition to the Observed Features. Given a partition $\Pi_N = \{B_1, \ldots, B_{K_+}\}$, if $i \in B_k$, then the cluster assignment of sample i is $a_i = k$. Its normalized feature $z_i \in \mathbb{S}^{d-1}$ is assumed to be generated from the prototype direction c_k. We adopt the following von Mises–Fisher (vMF) conditional likelihood:

$$p(z_i(\phi) \mid a_i = k, c_k, \phi) = C_d(\kappa) \exp\left(\kappa z_i(\phi)^\top c_k\right). \tag{35}$$

Here, $p(\cdot \mid \cdot)$ denotes a conditional probability density, a_i is the partition-block index of sample i, $z_i(\phi)$ is its current adapted feature, and c_k is the unit mean direction of the k-th cluster. The concentration parameter $\kappa > 0$ controls how tightly the features are distributed around their corresponding prototype, while $C_d(\kappa)$ is the vMF normalizing constant that depends only on the feature dimension d and κ.

Assuming that the features are conditionally independent given (Π_N, C), the complete log-likelihood is

$$\begin{aligned}\log \mathcal{L}_{\mathrm{vMF}}(Z_\phi \mid \Pi_N, C) = N \log C_d(\kappa) \\ + \kappa \sum_{k=1}^{K_+} \sum_{i \in B_k} z_i(\phi)^\top c_k.\end{aligned} \tag{36}$$

Here, $\mathcal{L}_{\mathrm{vMF}}(Z_\phi \mid \Pi_N, C)$ denotes the joint vMF likelihood of all adapted features given the partition and prototype configuration, and the superscript $\top$ denotes vector transposition. This likelihood provides the data-fitting signal that encourages each sample feature to lie close to its assigned prototype.

From the Prior and Likelihood to the Unified Posterior. By Bayes' rule, multiplying the likelihood in Eq. equation 36 by the Gibbs-tilted prior in Eq. equation 30 yields the posterior density of (Π_N, C) with respect to the reference measure ν_N:

$$\begin{aligned} p^{\mathrm{post}}_{\mathrm{Gibbs}}(\Pi_N, C \mid Z_\phi, \psi) &= \frac{\mathcal{L}_{\mathrm{vMF}}(Z_\phi \mid \Pi_N, C)}{m_\psi(Z_\phi)} p^{\mathrm{prior}}_{\mathrm{Gibbs}}(\Pi_N, C \mid \psi), \\ m_\psi(Z_\phi) &= \int_{\Omega_N} \mathcal{L}_{\mathrm{vMF}}(Z_\phi \mid \Pi, C) \\ &\qquad \times p^{\mathrm{prior}}_{\mathrm{Gibbs}}(\Pi, C \mid \psi)\, \nu_N(d\Pi, dC). \end{aligned} \tag{37}$$

Here, the superscripts prior and post denote the prior and posterior densities, respectively. The term $m_\psi(Z_\phi)$ is the model evidence obtained, for fixed ϕ, by integrating over all partitions and prototype configurations.

Expanding the prior and omitting constants that are independent of (Π_N, C) gives the following conditional posterior:

$$\begin{aligned} p^{\mathrm{post}}_{\mathrm{Gibbs}}(\Pi_N, C \mid Z_\phi, \psi) \propto & \underbrace{\mathcal{L}_{\mathrm{vMF}}(Z_\phi \mid \Pi_N, C)}_{\text{representation–prototype fit}} \\ & \times \underbrace{p_{\mathrm{CRP}}(\Pi_N \mid \alpha)}_{\text{adaptive partition cardinality}} \\ & \times \underbrace{\prod_{k=1}^{K_+} h_\psi(c_k)}_{\text{prototype support validity}} \\ & \times \underbrace{\det(\bar{L}_C)^{\lambda_g}}_{\text{active-prototype non-redundancy}} . \end{aligned} \tag{38}$$

From the Posterior to the Clustering Output. Ideally, MAP inference would directly search over the variable-cardinality space Ω_N:

$$(\widehat{\Pi}_N, \widehat{C}) = \arg \max_{(\Pi_N, C) \in \Omega_N} p^{\mathrm{post}}_{\mathrm{Gibbs}}(\Pi_N, C \mid Z_\phi, \psi). \tag{39}$$

Here, the notation $\widehat{\cdot}$ denotes a MAP estimate, while $\widehat{\Pi}_N$ and $\widehat{C}$ are the predicted partition and its corresponding prototype configuration, respectively. The predicted number of clusters is therefore $\widehat{K} = |\widehat{\Pi}_N|$. If $i \in \widehat{B}_k$, the predicted cluster label of sample i is $\widehat{y}_i = k$.

Accordingly, the inferred prototypes $\widehat{C}$ serve as latent parameters that connect data fitting, support validity, and non-redundancy. The primary outputs of the method remain the predicted cluster assignments $(\widehat{y}_1, \ldots, \widehat{y}_N)$ and the estimated cluster number $\widehat{K}$. Moreover, Eq. equation 43 in the main text is a finite MAP surrogate for this conditional posterior, rather than an exact posterior sampling procedure for the Dirichlet process.

A.4 Unified Structural Objective and Prototype Activation

Unified Structural Objective. The MAP objective derived in Appendix A.3 is defined over the variable-cardinality space Ω_N, for which exact inference is intractable. We therefore convert the unified posterior into a tractable finite MAP objective, yielding a unified structural objective. This approximation consists of three main steps. First, an overcomplete prototype inventory is used to approximate the set of potential prototype locations that may be activated. Second, binary activity variables indicate which candidate prototypes are currently active. Third, soft responsibilities approximate the discrete partition, allowing sample assignment, prototype selection, and prototype updating to be optimized alternately.

Specifically, we maintain an overcomplete finite prototype inventory $\widetilde{C} = \{\tilde{c}_m\}_{m=1}^{M}$, and introduce a binary activity variable $r_m \in \{0, 1\}$ for each candidate prototype. Let

$$\begin{aligned} \mathcal{I}_r &= \{m : r_m = 1\}, \\ C_r &= \{\tilde{c}_m : m \in \mathcal{I}_r\}, \\ K_+ &= \|r\|_0. \end{aligned} \tag{40}$$

Here, M is the size of the candidate-prototype inventory and $\tilde{c}_m$ denotes its m-th candidate prototype. The tilde distinguishes inventory candidates from the prototypes in the ideal variable-cardinality model. Moreover, $r = (r_1, \ldots, r_M)$ is the binary activity vector, $\mathcal{I}_r$ is the index set of active candidates, C_r is the corresponding active prototype set, and $\|r\|_0$ denotes the number of nonzero entries in r. The mixture weight π_m represents the mixing proportion of the m-th candidate prototype and satisfies $\pi_m \geq 0$ and $\sum_m r_m \pi_m = 1$. We define the responsibilities of inactive components to be zero. For each $m \in \mathcal{I}_r$, the soft responsibility is

$$q_{im} = \frac{\pi_m \exp\big(\kappa z_i(\phi)^\top \tilde{c}_m\big)}{\sum_{\ell \in \mathcal{I}_r} \pi_\ell \exp\big(\kappa z_i(\phi)^\top \tilde{c}_\ell\big)}. \tag{41}$$

Here, ℓ indexes the active candidate prototypes in the denominator, and q_{im} denotes the posterior soft responsibility of candidate prototype m for sample i. For each sample, these responsibilities satisfy $\sum_{m \in \mathcal{I}_r} q_{im} = 1.$. When κ is fixed, the vMF normalizing constant is independent of the assignments and can therefore be omitted. The resulting assignment free energy is

$$\begin{aligned} \mathcal{E}_{\text{fit}} = & -\sum_i \sum_{m \in \mathcal{I}_r} q_{im} \left(\log \pi_m + \kappa z_i(\phi)^\top \tilde{c}_m\right) \\ & + \sum_i \sum_{m \in \mathcal{I}_r} q_{im} \log q_{im}. \end{aligned} \tag{42}$$

Here, $\mathcal{E}_{\text{fit}}$ denotes the sample–prototype fitting free energy, consisting of the negative expected log-likelihood and the responsibility entropy term. Accordingly, the objective under the finite approximation is

$$\begin{aligned} \mathcal{F}_{\text{struct}} = \mathcal{E}_{\text{fit}} + \lambda_a^{\text{eff}} \|r\|_0 + \lambda_s \sum_{m=1}^{M} r_m b_\psi(\tilde{c}_m) \\ + \lambda_g \mathcal{E}_{\text{rep}}(C_r). \end{aligned} \tag{43}$$

Here, $\mathcal{F}_{\text{struct}}$ is the finite structural objective for fixed ϕ, and $\lambda_a^{\text{eff}} \geq 0$ is the effective activation-cost coefficient. The superscript eff indicates that this coefficient absorbs cardinality-dependent constants introduced by the finite approximation.

The term $\lambda_a^{\text{eff}} \|r\|_0$ is a finite-complexity surrogate inspired by DP-means, rather than an exact CRP posterior-sampling term. The effective coefficient also absorbs cardinality-dependent constants, including $K_+ \log Z_\psi$. Equation equation 43 is therefore a finite MAP approximation to the conditional posterior for fixed ϕ. The capacity term suppresses unnecessary active components, the frozen support barrier penalizes directions that lack empirical support from Stage I, and the repulsive energy discourages near-duplicate prototypes.

Prototype Activation. For a candidate prototype m that is currently inactive, let r^{+m} denote the activity state obtained by temporarily activating m. Under this state, its local responsibilities and mixture weights are re-optimized. We define

$$G_{\text{fit}}(m; \phi) = \min_{Q, \pi} \mathcal{E}_{\text{fit}}(Q, \widetilde{C}, r, \pi) - \min_{Q, \pi} \mathcal{E}_{\text{fit}}(Q, \widetilde{C}, r^{+m}, \pi). \tag{44}$$

Here, the superscript $+m$ indicates a quantity recomputed after temporarily adding candidate m to the current active set. Accordingly, $G_{\text{fit}}(m; \phi)$ denotes the reduction in theoptimal local fitting free

energy achieved by activating this candidate. We further define $n_m^+ = \sum_i q_{im}^{+m}$ as the proposal mass of candidate m, where q_{im}^{+m} is the responsibility assigned to candidate m after its temporary activation. For a feasible candidate satisfying $n_m^+ \geq \tau_n$, the coordinate-wise update activates m if and only if

$$G_{\text{fit}}(m;\phi) > \lambda_a^{\text{eff}} + \lambda_s b_\psi(\tilde{c}_m) + \lambda_g \Delta_{\text{rep}}(m;\phi). \tag{45}$$

The repulsive-energy increment is defined as

$$\Delta_{\text{rep}}(m;\phi) = \mathcal{E}_{\text{rep}}\left(C_r \cup \{\tilde{c}_m\}\right) - \mathcal{E}_{\text{rep}}(C_r). \tag{46}$$

Here, $\Delta_{\text{rep}}(m;\phi)$ denotes the increase in repulsive energy caused by adding candidate m. Prototype pruning is performed using the corresponding reverse energy comparison. This criterion is a coordinate-wise decision rule for a fixed current representation parameter ϕ: a candidate prototype is retained only when its fitting gain exceeds the sum of its activation cost, support penalty, and incremental redundancy cost.

A.5 Prototype Capacity Reallocation

Supported Under-Coverage Candidates. Equation equation 43 determines which prototypes in a given inventory should be active, but it does not specify how the inventory acquires new candidates from regions that are insufficiently explained by the current prototypes. Because soft responsibilities are relative quantities, a sample that is far from all active prototypes may still receive a high responsibility under its nearest prototype. We therefore introduce an absolute coverage residual that does not depend on the mixture weights:

$$d_i(C_r) = 1 - \max_{k \in \mathcal{I}_r} z_i(\phi)^\top \tilde{c}_k, \qquad \omega_i^{\text{uc}} = [d_i(C_r) - \tau_u]_+ \mathbb{I}[s_\psi(u_i) \geq \tau_s]. \tag{47}$$

Here, $d_i(C_r)$ measures the absolute under-coverage of sample i relative to the current active prototype set C_r, and τ_u is the minimum under-coverage threshold. The superscript uc denotes under-coverage, $\mathbb{I}[\cdot]$ is the indicator function, and ω_i^{uc} is the sample weight that jointly enforces sufficient coverage residual and support under the frozen denoiser. We construct a local neighborhood $\mathcal{N}_m$ from samples satisfying $\omega_i^{\text{uc}} > 0$ and form the proposed candidate direction as

$$\tilde{c}_m^{\text{prop}} = \text{norm}\left(\sum_{i \in \mathcal{N}_m} \omega_i^{\text{uc}} z_i(\phi)\right). \tag{48}$$

Here, $\mathcal{N}_m$ denotes the local sample neighborhood used to construct the m-th proposal, the superscript prop denotes a proposed candidate, and $\tilde{c}_m^{\text{prop}}$ is the normalized under-coverage-weighted direction of this neighborhood. After being inserted into an inactive inventory slot, the candidate must still satisfy the support penalty $b_\psi(\tilde{c}_m)$, the proposal-mass condition $n_m^+ \geq \tau_n$, and the activation or swapping criterion described below, where τ_n is the minimum responsibility mass required for a candidate. Thus, the support criterion determines whether a candidate direction is semantically credible, whereas the coverage residual determines whether the current structure requires a prototype in that direction.

Redundancy-Aware Prototype Swapping. Coordinate-wise birth and pruning change the active cardinality separately. Consequently, deleting a redundant prototype does not automatically transfer the released capacity to a supported under-covered candidate. Moreover, the fixed activation cost in Eq. equation 12 may prevent a candidate supported by only a small number of samples from being born independently. To connect these two operations without introducing an additional frequency model, we fix ϕ and the candidate inventory and define the profiled objective over the activity state as

$$\mathcal{F}_{\text{struct}}^\star(r;\phi) = \min_{Q,\pi} \mathcal{F}_{\text{struct}}(Q, \widetilde{C}, r, \pi; \phi). \tag{49}$$

Here, the superscript $\star$ indicates that the responsibility matrix Q and mixture weights π have been minimized out for the given activity state r and adapter parameter ϕ. When $K_+ > 1$, for an active

prototype $k \in \mathcal{I}_r$ and an inactive supported under-coverage candidate $m \notin \mathcal{I}_r$, we consider the following one-out-one-in state:

$$\begin{aligned} r^{k\to m} &= r - e_k + e_m, \\ C_r^{k\to m} &= \big(C_r \setminus \{\tilde{c}_k\}\big) \cup \{\tilde{c}_m\}, \\ \Delta_{\mathrm{swap}}(k, m; \phi) &= \mathcal{F}^{\star}_{\mathrm{struct}}(r^{k\to m}; \phi) - \mathcal{F}^{\star}_{\mathrm{struct}}(r; \phi). \end{aligned} \tag{50}$$

Here, $e_k, e_m \in \mathbb{R}^M$ are the standard basis vectors associated with the k-th and m-th coordinates, respectively. The superscript $k \to m$ denotes a one-out-one-in swap that removes the active prototype k and introduces the candidate prototype m. Accordingly, $\Delta_{\mathrm{swap}}(k, m; \phi)$ denotes the change in the profiled structural objective induced by this swap. The source prototype is preferentially selected from active prototypes with a small reverse-pruning cost $\mathcal{F}^{\star}_{\mathrm{struct}}(r - e_k; \phi) - \mathcal{F}^{\star}_{\mathrm{struct}}(r; \phi)$, whereas the target prototype is selected from candidates generated by Eq. equation 48 that have not passed the independent birth criterion. In practice, we first shortlist source prototypes with the lowest removal costs, compare them with the under-coverage candidates, and accept the best swap satisfying $\Delta_{\mathrm{swap}}(k, m; \phi) < 0$. Because $\|r^{k\to m}\|_0 = \|r\|_0$, the swap increment can be expanded as

$$\Delta_{\mathrm{swap}}(k, m; \phi) = \Delta\mathcal{E}_{\mathrm{fit}}^{k\to m} + \lambda_s[b_\psi(\tilde{c}_m) - b_\psi(\tilde{c}_k)] + \lambda_g\big[\mathcal{E}_{\mathrm{rep}}(C_r^{k\to m}) - \mathcal{E}_{\mathrm{rep}}(C_r)\big]. \tag{51}$$

Here, $\Delta\mathcal{E}_{\mathrm{fit}}^{k\to m}$ denotes the change in fitting energy after the swap and the subsequent re-optimization of Q and π. The fixed-cardinality cost associated with λ_a^{eff} cancels exactly. Although this operation does not make the accumulated fitting gain invariant to sample frequency, it allows a supported under-covered candidate to reuse an inefficient active slot without paying the additional fixed cost of an independent birth. It therefore lowers the entry barrier for candidates supported by relatively few samples.

A.6 Adaptation Regularization and Spherical Prototype Updates

We first freeze ϕ and minimize Eq. equation 43 with respect to Q, π, r, and $\widetilde{C}$. We then stop the gradients through these structural variables and ψ and update only the adapter parameter ϕ. A conventional sample-summed vMF loss allows larger clusters to contribute proportionally larger gradients. To mitigate this imbalance, we first normalize the contribution of each stable active prototype by its responsibility mass and then average uniformly across prototypes:

$$\mathcal{L}_{\mathrm{pbal}}(\phi) = -\frac{1}{|\mathcal{I}_s|}\sum_{m\in\mathcal{I}_s}\frac{\sum_i \mathrm{sg}(q_{im}) z_i(\phi)^\top \mathrm{sg}(\tilde{c}_m)}{\mathrm{sg}(n_m) + \epsilon_q}, \tag{52}$$

where the subscript pbal denotes prototype-balanced learning, $\mathcal{I}_s = \{m : r_m = 1,\ n_m \geq \tau_n\}$ is the set of stable active prototypes satisfying the minimum responsibility-mass requirement, and $|\mathcal{I}_s|$ is its cardinality. The operator $\mathrm{sg}(\cdot)$ denotes stop-gradient, and $\epsilon_q > 0$ is a numerical stability constant that prevents division by zero. The minimum-mass threshold prevents extremely small spurious clusters from receiving disproportionately large weights.

To keep the representation adaptation conservative, we further introduce base-representation anchoring and cross-view consistency:

$$\mathcal{L}_{\mathrm{anc}} = \frac{1}{N}\sum_i \big[1 - z_i(\phi)^\top \mathrm{sg}(u_i)\big], \tag{53}$$

$$\mathcal{L}_{\mathrm{view}} = \frac{1}{N}\sum_i \Big[1 - z_i^{(1)}(\phi)^\top \mathrm{sg}\big(z_i^{(2)}(\bar{\phi})\big)\Big]. \tag{54}$$

Here, the subscripts anc and view denote base-representation anchoring and cross-view consistency, respectively. The superscripts (1) and (2) denote two randomly augmented views of the same sample, and $\bar{\phi}$ is an exponential moving average (EMA) copy of ϕ. Positive-pair consistency alone may

still admit a collapsed solution in which all samples are mapped to the same point. Let $H \in \mathbb{R}^{B\times d}$ denote the batch-centered feature matrix, where H here is a feature matrix and should not be confused with the support base measure H_ψ. The symbol B denotes the batch size, $H_{:j}$ denotes the j-th feature dimension of H, and $j \in \{1,\dots,d\}$. We define the stabilized standard deviation of the j-th dimension as $v_j = \sqrt{\mathrm{Var}(H_{:j}) + \epsilon_v}$, where $\mathrm{Var}(\cdot)$ denotes the within-batch variance and $\epsilon_v > 0$ is a numerical stability constant. We further define the within-batch feature covariance matrix as $\Sigma_H = \frac{1}{B-1} H^\top H$. The anti-collapse regularizers are then given by

$$\begin{aligned}\mathcal{L}_{\mathrm{var}} &= \frac{1}{d}\sum_{j=1}^{d}[\nu_0 - v_j]_+,\\ \mathcal{L}_{\mathrm{cov}} &= \frac{1}{d(d-1)}\sum_{j\neq k}\Sigma_H[j,k]^2.\end{aligned} \tag{55}$$

The variance term maintains a minimum level of variation along each feature dimension, whereas the covariance term reduces redundancy across different dimensions. The complete adapter objective is

$$\mathcal{L}_{\mathrm{adapt}} = \mathcal{L}_{\mathrm{pbal}} + \lambda_{\mathrm{anc}}\mathcal{L}_{\mathrm{anc}} + \lambda_{\mathrm{inv}}\mathcal{L}_{\mathrm{view}} + \lambda_{\mathrm{var}}\mathcal{L}_{\mathrm{var}} + \lambda_{\mathrm{cov}}\mathcal{L}_{\mathrm{cov}}.$$

Here, $\mathcal{L}_{\mathrm{adapt}}$ denotes the complete adapter objective, and $\lambda_{\mathrm{anc}}, \lambda_{\mathrm{inv}}, \lambda_{\mathrm{var}}, \lambda_{\mathrm{cov}} \geq 0$ weight the representation-anchoring, cross-view-invariance, variance-preservation, and covariance-suppression terms, respectively. The subscript inv denotes invariance. The log-determinant term regularizes only the prototype set and therefore cannot replace these sample-level anti-collapse regularizers.

Given Q, r, and π, let $g_m = \nabla_{\tilde{c}_m}\mathcal{F}_{\mathrm{struct}}(Q, \widetilde{C}, r, \pi; \phi)$.We first project this Euclidean gradient onto the tangent space of the unit sphere and then apply a retraction:

$$g_m^\perp = \left(I - \tilde{c}_m\tilde{c}_m^\top\right) g_m, \qquad \tilde{c}_m \leftarrow \mathrm{norm}\left(\tilde{c}_m - \eta g_m^\perp\right). \tag{56}$$

Here, g_m is the Euclidean gradient of the structural objective with respect to the candidate prototype $\tilde{c}_m$, and I is the d-dimensional identity matrix. The projection matrix $I - \tilde{c}_m\tilde{c}_m^\top$ maps the gradient onto the tangent space of the sphere at $\tilde{c}_m$, the superscript $\perp$ denotes the resulting tangent-space gradient, and $\eta > 0$ is the prototype update step size. The subsequent normalization implements a retraction back onto the unit sphere. The final estimate of the number of clusters and the predicted sample labels are read from the converged active set:

$$\widehat{K} = \|r^\star\|_0, \qquad \hat{y}_i = \underset{m:r_m^\star=1}{\arg\max}\, q_{im}. \tag{57}$$

Here, the superscript $\star$ denotes the activity state obtained after convergence of the alternating optimization, $\widehat{K}$ is the final number of active prototypes, and $\hat{y}_i$ is the predicted cluster label of sample i. Algorithm 1 summarizes the complete procedure.

B Mechanism Analysis and Proofs

This section formalizes the mechanisms summarized in the main text. We provide theoretical results for the denoising-supported base measure, Gibbs repulsion, conditional prototype activation, cardinality-preserving prototype exchange, and representation anchoring.

B.1 Support Control

Support Violation Bound. Let $\mathcal{F}(x) = \mathcal{R}(x) + \lambda_s\mathcal{E}_{\mathrm{sup}}(x)$, where $\mathcal{E}_{\mathrm{sup}}(x) = \sum_m r_m b_\psi(\tilde{c}_m)$. Suppose that $\mathcal{R}(x) \geq B$, and that there exists a feasible reference solution x^{ref} such that $\mathcal{R}(x^{\mathrm{ref}})$ is finite and $\mathcal{E}_{\mathrm{sup}}(x^{\mathrm{ref}}) = 0$. Then any global minimizer $x^\star$ satisfies

$$\mathcal{E}_{\mathrm{sup}}(x^\star) \leq \frac{\mathcal{R}(x^{\mathrm{ref}}) - B}{\lambda_s}. \tag{58}$$

Algorithm 1 DRP Structure Inference and Conservative Representation Adaptation with a Fixed Support Reference

Require: Unlabeled data $\mathcal{D}$, inventory size M, thresholds τ_s, τ_u, τ_n, and structure/adaptation coefficients

Obtain base features U; train D_{ψ_0} only on U and freeze it
Initialize $A_{\phi_0} \equiv 0$, the overcomplete prototype inventory $\widetilde{C}$, and activity state r
Fix $\phi = \phi_0$ and run structural warm-up
repeat
  Fix ϕ; update responsibilities Q, mixing weights π, and active prototypes
  Compute candidate support costs using the fixed D_{ψ_0}
  Propose supported under-coverage candidates via Eq. equation 47
  Accept births that can independently pay their cost via Eq. equation 12
  Accept the best cardinality-preserving swap via Eq. equation 50, and perform pruning
  Update active prototypes via Eq. equation 56
  Stop gradients for structural variables and ψ_0; update ϕ via Eq. equation 16
  Update EMA parameters $\bar{\phi}$
until active set and structural objective stabilize
return $\widehat{K} = \|r\|_0$ and labels $\hat{y}_i = \arg\max_{m:r_m=1} q_{im}$

Proof. By optimality,

$$\mathcal{F}(x^\star) \le \mathcal{F}(x^{\mathrm{ref}}) = \mathcal{R}(x^{\mathrm{ref}}), \tag{59}$$

where the equality follows from $\mathcal{E}_{\mathrm{sup}}(x^{\mathrm{ref}}) = 0$. Since $\mathcal{R}(x^\star) \ge B$, we obtain

$$B + \lambda_s \mathcal{E}_{\mathrm{sup}}(x^\star) \le \mathcal{R}(x^{\mathrm{ref}}). \tag{60}$$

Rearranging yields Eq. equation 58.

If the reference solution itself incurs a nonzero support violation, the same argument gives

$$\mathcal{E}_{\mathrm{sup}}(x^\star) \le \mathcal{E}_{\mathrm{sup}}(x^{\mathrm{ref}}) + \frac{\mathcal{R}(x^{\mathrm{ref}}) - B}{\lambda_s}. \tag{61}$$

B.2 Repulsion Against Prototype Collapse

Repulsion Against Prototype Collapse. For the kernel in Eq. equation 6 without jitter, two identical active prototypes induce identical kernel rows, implying $\det(L_C) = 0$ and hence $-\log\det(L_C) = +\infty$. When $\varepsilon > 0$, the energy remains finite; nevertheless, repeated or nearly linearly dependent prototype configurations still induce a smaller kernel volume. Thus, the practical repulsion term discourages prototype collapse, although it does not rule it out under finite ε and finite λ_g.

Proof. Suppose that there exist $i \ne j$ such that $c_i = c_j$. Then, for any active prototype c_ℓ,

$$L_C[i,\ell] = \exp\{\rho(c_i^\top c_\ell - 1)\} = L_C[j,\ell]. \tag{62}$$

Hence, the i-th and j-th rows of L_C are identical, so L_C is singular and $\det(L_C) = 0$. Consequently, the corresponding energy without jitter diverges.

When jitter is added, $\bar{L}_C$ remains positive definite and the energy is finite. However, nearly duplicated prototypes induce nearly dependent kernel rows before jitter is added, leading to small eigenvalues and a smaller determinant than configurations with more independent kernel features. Therefore, the log-determinant term induces an anti-collapse bias, but does not impose a hard exclusion of collapsed configurations for finite ε and finite λ_g.

B.3 Coordinate-Wise Activation Criterion

Coordinate-Wise Activation Criterion. Fix the current adapter ϕ and all variables except the activity state of candidate prototype m. Re-optimize the local responsibilities and mixture weights under $r_m = 0$ and $r_m = 1$, respectively. Then activating m strictly decreases Eq. equation 43 if and only if Eq. equation 12 holds. This is a conditional coordinate-wise criterion and does not imply global optimality.

Proof. Let $\mathcal{F}_0$ denote the minimized coordinate objective under $r_m = 0$, and let $\mathcal{F}_1$ denote the corresponding value obtained after re-optimizing the local responsibilities and mixture weights under $r_m = 1$. By definition,

$$\begin{aligned}\mathcal{F}_1 - \mathcal{F}_0 = & -G_{\text{fit}}(m;\phi) + \lambda_a^{\text{eff}} + \lambda_s b_\psi(\tilde{c}_m) \\ & + \lambda_g \Delta_{\text{rep}}(m;\phi).\end{aligned} \tag{63}$$

Activating the candidate strictly decreases the coordinate objective if and only if $\mathcal{F}_1 - \mathcal{F}_0 < 0$. The conclusion further assumes that the current ϕ is fixed and that $n_m^+ \geq \tau_n$. It is therefore only a conditional coordinate-wise result.

B.4 Cardinality-Preserving Prototype Exchange

Cardinality Preservation and Objective Descent under Prototype Exchange. Fix the current adapter ϕ and the candidate pool. For any active prototype $k \in \mathcal{I}_r$ and inactive candidate $m \notin \mathcal{I}_r$, define $r^{k\to m} = r - e_k + e_m$. Then the number of active prototypes is unchanged by the exchange. Moreover, if Eq. equation 50 satisfies $\Delta_{\text{swap}}(k, m; \phi) < 0$, the exchange strictly decreases the profiled value of Eq. equation 43 after re-optimizing Q and π.

Proof. Since $r_k = 1$ and $r_m = 0$, removing k and activating m gives

$$\|r^{k\to m}\|_0 = \|r\|_0. \tag{64}$$

Hence, the complexity term $\lambda_a^{\text{eff}}\|r\|_0$ is exactly unchanged by the exchange.

The difference in the remaining fitting, support, and repulsion terms is precisely Eq. equation 51. By definition, $\Delta_{\text{swap}}(k, m; \phi) < 0$ is equivalent to the profiled objective being strictly smaller after the exchange.

This result only guarantees a cardinality-preserving reallocation of prototype capacity under the same finite objective. It does not guarantee that the incoming candidate corresponds to a ground-truth tail class, nor does it remove the dependence of $\Delta\mathcal{E}_{\text{fit}}^{k\to m}$ on sample frequency.

B.5 Representation Drift Bound under Anchoring

Representation Drift Bound under Anchoring. After absorbing the constant factor relating Eq. equation 53 to squared Euclidean distance into the coefficient, define

$$\mathcal{J}(\phi) = \mathcal{R}_{\text{adapt}}(\phi) + \lambda_{\text{anc}} \sum_i \|z_i(\phi) - u_i\|_2^2. \tag{65}$$

If $\mathcal{R}_{\text{adapt}}(\phi) \geq B_\phi$ and the identity initialization ϕ_0 is feasible, then any global minimizer $\phi^\star$ satisfies

$$\sum_i \|z_i(\phi^\star) - u_i\|_2^2 \leq \frac{\mathcal{R}_{\text{adapt}}(\phi_0) - B_\phi}{\lambda_{\text{anc}}}. \tag{66}$$

This bound controls only the aggregate representation drift and does not guarantee that the direction of adaptation necessarily improves the underlying semantic structure. If the drift of every sample is at most ϵ, then the change in cosine similarity between any pair of samples is at most 2ϵ.

Proof. By the optimality of $\phi^\star$ and the feasibility of the identity initialization,

$$\mathcal{R}_{\text{adapt}}(\phi^\star) + \lambda_{\text{anc}} \sum_i \|z_i(\phi^\star) - u_i\|_2^2 \leq \mathcal{R}_{\text{adapt}}(\phi_0). \tag{67}$$

Using $\mathcal{R}_{\text{adapt}}(\phi^\star) \geq B_\phi$ immediately yields Eq. equation 66.

For unit vectors z_i, z_j, u_i, u_j, if $\|z_i - u_i\|_2, \|z_j - u_j\|_2 \leq \epsilon$, then

$$\begin{aligned}|z_i^\top z_j - u_i^\top u_j| &= |(z_i - u_i)^\top z_j + u_i^\top (z_j - u_j)| \\ &\leq \|z_i - u_i\|_2 \|z_j\|_2 + \|u_i\|_2 \|z_j - u_j\|_2 \\ &\leq 2\epsilon,\end{aligned} \tag{68}$$

where the first inequality follows from the Cauchy–Schwarz inequality.

B.6 Variational Derivation of the Denoising-Supported Base Measure

For any $Q \ll \mu_{\mathbb{S}}$, let $q = dQ/d\mu_{\mathbb{S}}$. The objective in Eq. equation 3 can be written as

$$\begin{aligned}\mathcal{J}(q) = &\int_{\mathbb{S}^{d-1}} q(c) \log q(c)\, d\mu_{\mathbb{S}}(c) \\ &+ \lambda_s \int_{\mathbb{S}^{d-1}} q(c) b_\psi(c)\, d\mu_{\mathbb{S}}(c),\end{aligned} \tag{69}$$

subject to

$$\int q\, d\mu_{\mathbb{S}} = 1. \tag{70}$$

Introducing a Lagrange multiplier η, the augmented functional is

$$\widetilde{\mathcal{J}}(q, \eta) = \mathcal{J}(q) + \eta \left(\int q\, d\mu_{\mathbb{S}} - 1 \right). \tag{71}$$

Taking the first variation with respect to q gives

$$\frac{\delta \widetilde{\mathcal{J}}}{\delta q(c)} = \log q(c) + 1 + \lambda_s b_\psi(c) + \eta. \tag{72}$$

Setting the variation to zero yields

$$q(c) \propto \exp[-\lambda_s b_\psi(c)]. \tag{73}$$

Applying the normalization constraint then gives Eqs. equation 4 and equation 21.

The entropy term $\int q(c) \log q(c)\, d\mu_{\mathbb{S}}(c)$ is strictly convex in q on the probability simplex, while the support term is linear in q. Therefore, the overall variational objective is strictly convex, and the stationary point above is the unique global minimizer.

B.7 Well-Posedness of the Gibbs Tilt

For any finite active configuration, the exponentiated inner-product kernel in Eq. equation 6 is positive semidefinite. After adding εI, it becomes positive definite. The subsequent normalization preserves a unit diagonal. Therefore, by Hadamard's inequality,

$$0 < \det(\bar{L}_C) \le \prod_{k=1}^{K_+} \bar{L}_C[k, k] = 1. \tag{74}$$

Since P_0 is a probability measure and $\lambda_g \ge 0$, it follows that

$$0 < \mathcal{Z}_{G,N} = \mathbb{E}_{P_0} \left[\det(\bar{L}_C)^{\lambda_g} \right] \le 1. \tag{75}$$

Thus, the normalizing constant is finite and strictly positive, and Eq. equation 33 defines a valid Radon–Nikodym derivative.

B.8 Remarks on the Finite-Cardinality Penalty

The normalized base measure contributes

$$-\sum_{k=1}^{K_+} \log h_\psi(c_k) = \lambda_s \sum_{k=1}^{K_+} b_\psi(c_k) + K_+ \log Z_\psi. \tag{76}$$

Therefore, when the cardinality is optimized jointly, the term $K_+ \log Z_\psi$ cannot in general be discarded. In Eq. equation 43, this term has already been absorbed into λ_a^{eff}.

The remaining linear complexity penalty admits a DP-means-style interpretation only under a small-variance asymptotic regime in which the likelihood concentration and the DP concentration parameter are scaled jointly in an appropriate manner. Outside this regime, we interpret the resulting objective as a finite MAP surrogate rather than as exact CRP inference.

C Additional Experimental Details

This section provides additional experimental details, including the construction of the long-tailed datasets and their exact sample statistics, visualizations of the learned representations, sensitivity analyses of hyperparameters such as $K_{\max}$ and $\kappa_{\max}$, and an analysis of computational complexity.

C.1 Long-Tailed Dataset Construction and Exact Sample Statistics

Data pool. We follow the experimental protocols of ConMix and MiniClustering to construct the long-tailed datasets. For CIFAR-10, CIFAR-20, and Tiny ImageNet, we use the original training set as the data pool from which the long-tailed splits are constructed. CIFAR-20 is obtained from the 20 coarse-grained superclasses of CIFAR-100, with each superclass treated as one clustering category. For STL-10, we merge the official training and test sets and use the combined set for training and clustering. For the other datasets, only the training split is used. For ImageNet-LT, we directly use its official long-tailed split without further modification.

Long-tailed split. We use the imbalance ratio to determine the number of samples retained for each class when constructing a long-tailed dataset. The number of samples per class follows an exponentially decaying distribution, where the imbalance ratio is defined as the ratio between the largest and smallest class sizes. Specifically, let a dataset contain C classes, where $N_{\max}$ denotes the target number of samples in the largest class after construction and $N_{\min}$ denotes that in the smallest class. The imbalance ratio is defined as

$$\mathrm{IR} = \frac{N_{\max}}{N_{\min}}.$$

We index the classes from head to tail as $i \in \{0, \dots, C-1\}$. The number of retained samples in the i-th class is

$$n_i = \begin{cases} \left\lfloor N_{\max} \mathrm{IR}^{-i/(C-1)} \right\rfloor, & i < C-1, \\ \lfloor N_{\max}/\mathrm{IR} \rfloor, & i = C-1. \end{cases} \tag{77}$$

Exact sample statistics. Following Eq. equation 77, we compute and report the exact number of samples used for training in each long-tailed split. Table 8 summarizes the number of classes, imbalance ratio, maximum and minimum class sizes, and total number of samples for each dataset. Figure 5 further reports the exact class-wise sample counts for CIFAR-10-LT, CIFAR-20-LT, and STL-10-LT. Since Tiny ImageNet-LT contains 200 classes, its class-wise sample counts are reported separately in Figure 6, with class indices grouped into blocks of 25. For ImageNet-LT, we directly use the official long-tailed split and therefore report only its overall statistics in Table 8.

Table 8: Overall statistics of the long-tailed datasets used in our experiments.

Dataset	C	IR	$N_{\max}$	$N_{\min}$	Total
CIFAR-10-LT	10	5	5,000	1,000	25,423
CIFAR-10-LT	10	10	5,000	500	20,431
CIFAR-20-LT	20	5	2,500	500	25,114
CIFAR-20-LT	20	10	2,500	250	19,954
STL-10-LT	10	5	1,300	260	6,607
STL-10-LT	10	10	1,300	130	5,309
Tiny ImageNet-LT	200	10	500	50	39,072
ImageNet-LT	1,000	256	1,280	5	115.8K

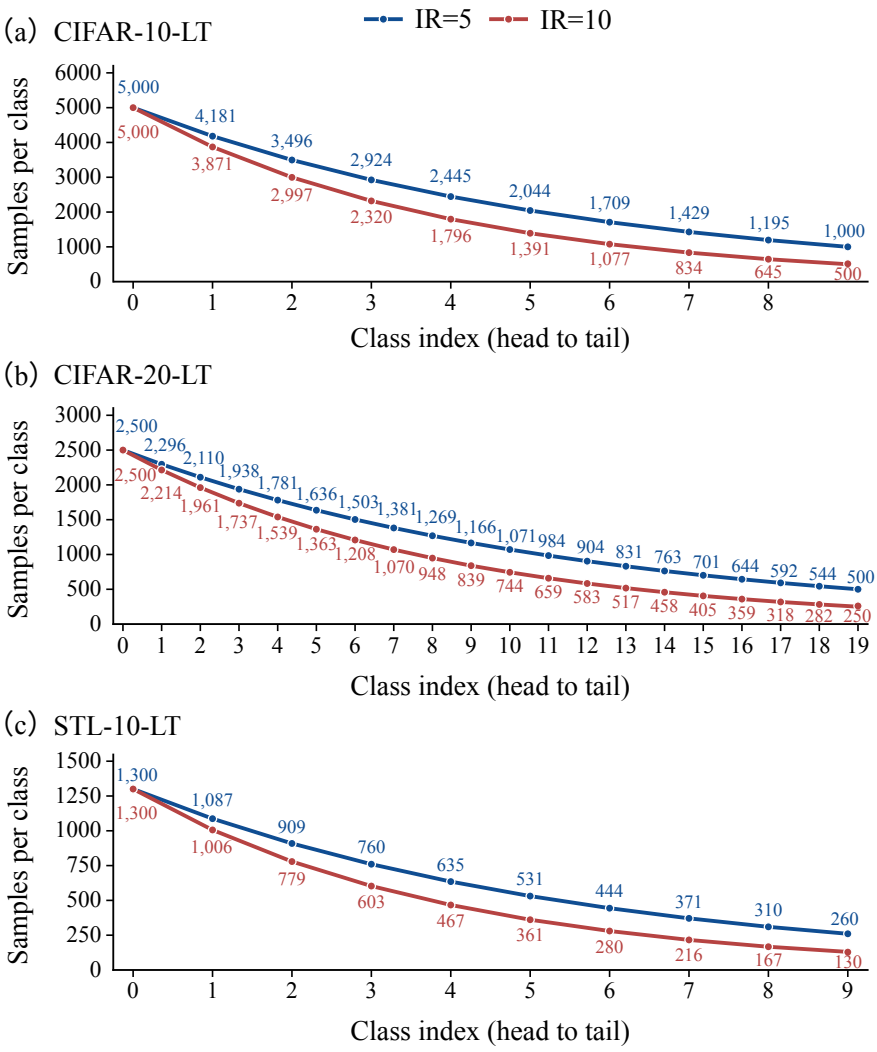


Figure 5: Exact class-wise sample counts for CIFAR-10-LT, CIFAR-20-LT, and STL-10-LT. Class indices are ordered from head to tail.

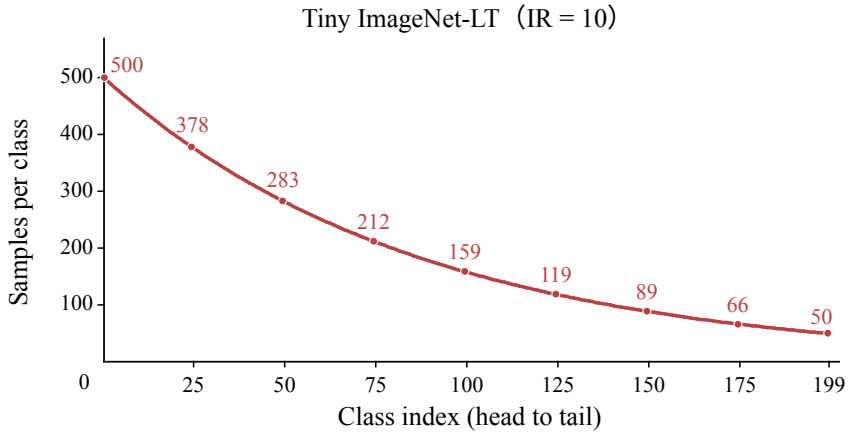


Figure 6: Exact class-wise sample counts for Tiny ImageNet-LT with imbalance ratio = 10.

C.2 Visualization of Learned Representations

We use t-distributed stochastic neighbor embedding (t-SNE) to qualitatively examine how the proposed method reshapes the learned representation space. Compared with the original feature space, the representations learned by our method exhibit more compact intra-class structures and clearer local separation between different classes. This behavior is consistently observed on both the long-tailed datasets (Figs. 7–10) and the balanced ImageNet datasets (Fig. 11), suggesting that the proposed method can recover well-separated clustering structures under both imbalanced and balanced data distributions.

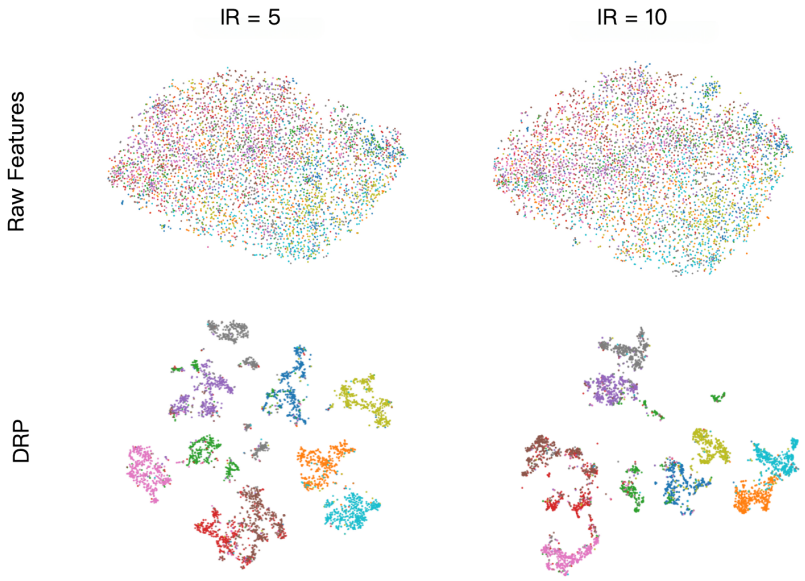


Figure 7: t-SNE visualization on CIFAR-10-LT.

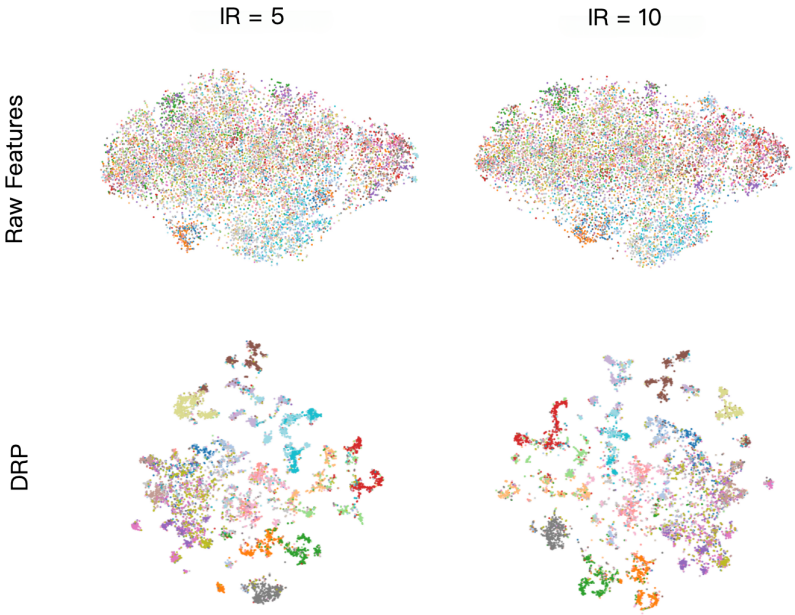


Figure 8: t-SNE visualization on CIFAR-20-LT.

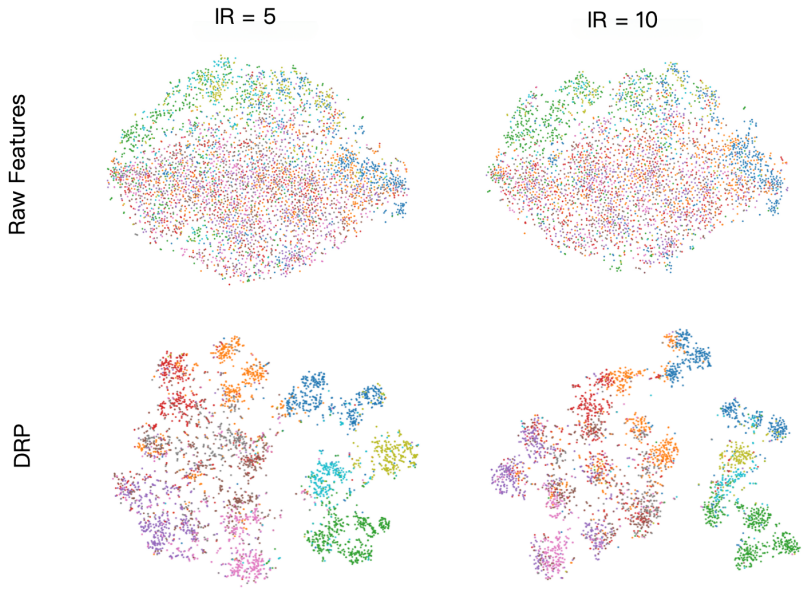


Figure 9: t-SNE visualization on STL-10-LT.

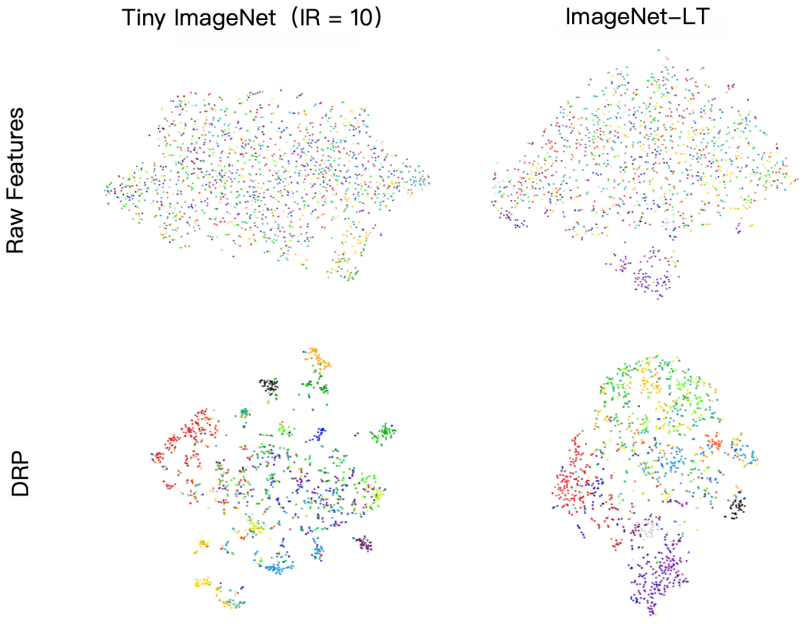


Figure 10: t-SNE visualization on Large-Scale Long-Tailed Datasets.

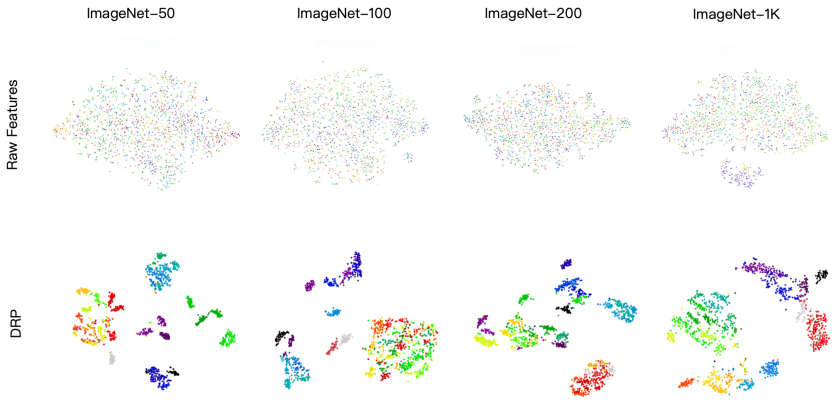


Figure 11: t-SNE visualizations on the balanced ImageNet datasets.

C.3 Hyperparameter Sensitivity Analysis

To further evaluate the robustness of unknown-category-number inference, we analyze the sensitivity of our method to the truncation level $K_{\max}$ and the maximum concentration parameter $\kappa_{\max}$. In addition to the standard clustering metrics, we report the estimated number of clusters $\widehat{K}$ to directly examine how these hyperparameters affect structural inference.

Table 9: Sensitivity to the truncation level $K_{\max}$.

$K_{\max}$	ACC	CAA	NMI	ARI	$\widehat{K}$
250	27.0	21.5	42.8	16.5	164.2
300	26.7	21.7	43.1	17.3	182.7
320	27.1	22.3	43.4	17.3	189.7
400	25.5	21.8	42.9	17.5	215.6

Table 10: Sensitivity to the maximum concentration parameter $\kappa_{\max}$.

$\kappa_{\max}$	ACC	CAA	NMI	ARI	$\widehat{K}$
115	26.7	21.9	43.6	17.6	196.0
120	26.5	21.9	43.4	17.3	195.2
150	27.1	22.3	43.4	17.3	189.7
200	26.1	21.6	43.0	17.4	188.4

C.4 Computational Complexity Analysis

Notation. Let N denote the number of samples, d the feature dimension, M the size of the candidate-prototype inventory, K_+ the number of active prototypes, T the number of alternating optimization rounds, and L the number of perturbations used to estimate denoising support. We use E_D for the number of denoiser-training passes and S for the maximum number of shortlisted birth or swap trials evaluated in one round. The costs of one encoder forward pass, one denoiser forward pass, and one adapter forward/backward pass per sample are denoted by C_E, C_D, and C_A, respectively. The latter includes the amortized cost of the batchwise variance–covariance regularization.

Time complexity. Extracting the frozen base features and training the support denoiser are one-time operations with cost $O(NC_E + E_D NC_D)$. Within each alternating round, updating the soft responsibilities, coverage residuals, mixture weights, and active prototypes costs $O(NK_+d)$. Evaluating the fixed-denoiser support for all candidates costs $O(MLC_D)$. Constructing the active-prototype

Table 11: Clustering performance (%) at imbalance ratio = 5. Best results are bold.

Method	Venue	CIFAR-10-LT				CIFAR-20-LT				STL-10-LT			
		ACC	CAA	NMI	ARI	ACC	CAA	NMI	ARI	ACC	CAA	NMI	ARI
SCAN	ECCV'20	55.6	60.3	54.0	39.3	38.8	39.7	42.6	25.1	48.8	49.8	46.2	34.9
SDCLR	ICML'21	44.1	50.5	43.4	38.0	40.2	38.7	40.3	23.8	35.8	37.4	34.4	19.3
CC	AAAI'21	25.2	20.9	18.3	1.47	16.0	12.0	18.0	0.0	41.5	30.0	46.2	22.4
IDFD	ICLR'20	56.7	63.0	51.8	37.1	31.2	30.7	30.5	16.2	41.6	41.2	39.6	24.3
DRP	**Ours**	**78.8**	**76.2**	**69.9**	**63.2**	**48.3**	**44.4**	**49.3**	**33.2**	**57.6**	**54.5**	**55.0**	**44.1**

kernel requires $O(K_+^2 d)$ operations, and its log-determinant can be computed by a Cholesky factorization in $O(K_+^3)$. The adapter update costs $O(NC_A)$. Exact re-evaluation of S shortlisted activation or exchange states contributes at most $O(SNK_+d)$ when their responsibilities are recomputed. The total time complexity is therefore

$$O\big(NC_E + E_D NC_D + T\left[NC_A + (S+1)NK_+d + MLC_D + K_+^2 d + K_+^3\right]\big). \tag{78}$$

In practice, the birth and exchange tests are applied only to a small shortlist, so S remains small. Since K_+, M, and S are much smaller than N in our experiments, the sample–prototype and adapter terms dominate, and the computational cost grows approximately linearly with the dataset size. The cubic term depends on the active prototype count rather than the number of samples.

Space complexity. DRP stores the base and adapted features, the candidate inventory, the responsibility matrix, and the active-prototype kernel. Their memory requirements are $O(Nd)$, $O(Md)$, $O(NK_+)$, and $O(K_+^2)$, respectively. With mini-batch size B, let A_A denote the adapter activation memory per sample and let P_D and P_A denote the numbers of denoiser and adapter parameters. The overall auxiliary space complexity is

$$O\big(N(d+K_+) + Md + K_+^2 + BA_A + P_D + P_A\big). \tag{79}$$

The frozen encoder parameters can be added as P_E when the shared backbone is included in the accounting. Thus, memory is linear in N and M, apart from the small $K_+ \times K_+$ prototype kernel.

Empirical wall-clock cost. We measured the downstream optimization by using one NVIDIA GeForce RTX 3080 GPU with 10 GB of memory and an Intel Core i7-10700K CPU. On CIFAR-10-LT with an imbalance ratio of 10 ($N = 20{,}431$), batch size 256, eight data-loading workers, and 200 training epochs, the run completed in 1,699.6 seconds (28.33 minutes). This corresponds to 8.50 seconds per epoch and an effective throughput of approximately 2,404 training samples per second. The measurement includes data loading, periodic evaluation, and final checkpoint serialization. It excludes the BYOL pretraining stage because the pretrained representation is shared by all downstream methods.

D Additional Experiments

D.1 Additional Baselines and Comparison Results

This subsection further supplements multiple baseline methods for performance comparison, including SCAN (Gansbeke et al., 2020), SDCLR (Jiang et al., 2021), CC (Li et al., 2021), and IDFD (Tao et al., 2021).

D.2 Multi-Seed Experiments

To evaluate robustness to random initialization, we repeat each experiment using five random seeds. Table 13 reports the mean and standard deviation across these five runs, complementing the single-run results presented in the main paper.

D.3 Experiments on Different Backbone Networks

In the main text, our proposed method uses BYOL (Grill et al., 2020) as the backbone. We further conduct experiments with SimCLR (Chen et al., 2020) and Moco (He et al., 2020) to validate the ef-

Table 12: Clustering performance (%) at imbalance ratio = 10. Best results are bold.

Method	Venue	CIFAR-10-LT				CIFAR-20-LT				STL-10-LT			
		ACC	CAA	NMI	ARI	ACC	CAA	NMI	ARI	ACC	CAA	NMI	ARI
SCAN	ECCV'20	52.0	50.1	59.4	43.8	37.0	32.4	37.8	23.1	47.1	45.4	47.5	34.2
SDCLR	ICML'21	38.9	44.3	42.5	26.5	37.8	35.9	39.6	22.9	34.6	37.9	32.2	17.3
CC	AAAI'21	40.6	27.5	43.9	18.8	19.9	14.3	21.9	1.1	43.0	35.3	44.7	25.4
IDFD	ICLR'20	47.5	54.9	48.4	33.1	28.7	27.2	28.6	15.1	38.6	34.7	36.8	22.8
DRP	**Ours**	**67.4**	**63.9**	**65.7**	**61.1**	**46.5**	**43.2**	**47.8**	**31.5**	**54.2**	**53.8**	**50.4**	**40.2**

Table 13: Clustering performance across five random seeds.

Metric	CIFAR-10-LT (IR = 10)	CIFAR-20-LT (IR = 10)	STL-10-LT (IR = 10)	ImageNet-50
ACC	68.4 ± 0.4	47.0 ± 0.2	54.2 ± 0.6	85.4 ± 0.2
CAA	64.2 ± 0.3	44.2 ± 0.6	53.4 ± 0.4	–
NMI	66.5 ± 0.6	47.6 ± 0.3	50.5 ± 0.4	92.0 ± 0.4
ARI	62.1 ± 0.3	31.9 ± 0.2	39.8 ± 0.2	79.8 ± 0.6

Table 14: Clustering performance in percentage (%) on CIFAR-10 and STL-10 (imbalance ratio = 10) based on various frameworks.

Datasets		CIFAR-10				STL-10			
Frameworks	Methods	ACC	CAA	NMI	ARI	ACC	CAA	NMI	ARI
BYOL	Baseline	51.9	55.2	56.3	41.7	41.2	41.8	46.8	29.6
	DRP	**67.4**	**63.9**	**65.7**	**61.1**	**54.2**	**53.8**	**50.4**	**40.2**
SimCLR	Baseline	43.4	47.9	45.8	30.4	38.1	38.8	37.5	21.6
	DRP	**54.5**	**58.9**	**58.3**	**43.1**	**50.9**	**51.2**	**48.8**	**38.0**
Moco	Baseline	37.9	41.1	38.1	26.3	30.5	30.6	21.1	12.8
	DRP	**49.2**	**53.8**	**53.0**	**36.6**	**44.5**	**43.5**	**38.3**	**27.8**

fectiveness of the method. As shown in Table 14, when applied to different pre-training frameworks, the method consistently maintains high performance, demonstrating its compatibility and robustness across various representation learning methods.