InfoSculpt: Rebalancing Discriminative and Discoverable Information for Generalized Category Discovery

Anonymous authors
Paper under double-blind review

Abstract

Generalized Category Discovery (GCD) poses a definitional tension: one representation space must simultaneously serve *discrimination* of known classes and *discovery* of unseen ones. We present evidence that a root cause of current GCD methods' failures on novel classes lies in unconstrained information allocation within the representation itself. Spectral analysis of a representative trained GCD model reveals two opposite pathologies coexisting in its novel-class features—over-compression into narrow manifolds, and under-compression that buries category cues in instance noise—while downstream remedies (debiasing heads, pseudo-label repair, decoupled branches) act at the classifier output and leave a residual gap. We then audit the allocation causally: through a conditional mutual information (CMI) variational bound, the trade-off is made an explicit knob, and under a matched-seed significance protocol, turning it toward discoverability improves *New*-class accuracy on 14/14 method–dataset pairs (+1.3 to +13.2 pp), significant in 13/13 held-out tests (Holm-adjusted $p \leq .008$) while old-class accuracy is preserved on average, and it outperforms single-term and static-balancing alternatives. Integrated into five SOTA methods across seven benchmarks, InfoSculpt raises All accuracy by +2.3 pp on average (up to +17.2 pp) and ranks best among all compared methods on five benchmarks, at 1.18× end-to-end cost with zero inference overhead.

1 Introduction

Deep learning has achieved great success in image classification under the "closed-world" assumption, where all test data belong to known classes. In real-world open-world scenarios Zhu et al. (2024); Geng et al. (2020), this assumption often fails due to costly annotations and emerging novel categories. Novel Category Discovery (NCD) Han et al. (2021) addresses unlabeled data from novel classes. Generalized Category Discovery (GCD) Vaze et al. (2022) further considers unlabeled data containing both known and novel classes, making it more realistic and challenging by requiring stricter disentanglement between categories. On the unlabeled side of GCD, ground-truth supervision is absent and the pseudo-labels generated for unlabeled data are unreliable, so the separability of latent categories depends strongly on the representation. Spectral analysis of ProtoGCD Ma et al. (2025) on CIFAR-100 reveals two opposite pathologies within its novel-class features (Fig. 1): a premature singular-value decay (*over-compression*) that traps novel features in narrow, low-rank manifolds, and an overly flat spectrum (*under-compression*) that buries category cues in instance-specific noise. Such misallocation underlies the distracted attention, scene-induced ambiguity, base-class bias, and old-class over-protectiveness reported by recent studies Xu et al. (2025); Peng et al. (2025); Liu et al. (2025a); Yu et al. (2026); yet current remedies patch each symptom at the classifier output Guo et al. (2025); the cause sits upstream, in the representation.

Both pathologies share one root: the competition between two kinds of information in the representation is unconstrained. Discrimination compresses labeled samples

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.

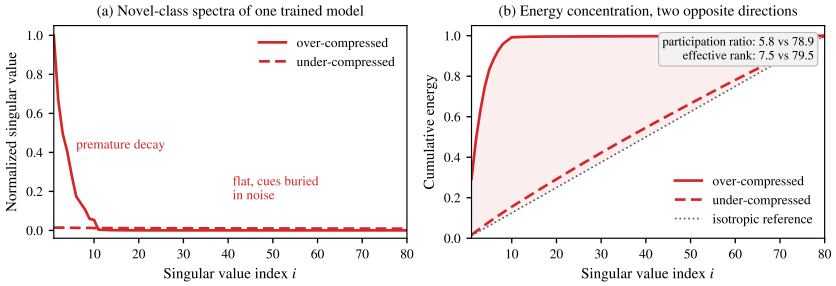

Figure 1: Allocation imbalance made visible. SVD Klema & Laub (1980) of last-layer features on CIFAR-100 novel-class data for ProtoGCD Ma et al. (2025). (a) Normalized spectra: a premature decay (*over-compression*) and a flat profile (*under-compression*) coexist within the same trained model. (b) The same spectra in cumulative form against an isotropic reference: energy either collapses into the leading directions or spreads evenly across all directions; participation ratio 5.8 vs. 78.9 (of 80). The allocation is unconstrained.

toward class structure, while discovery requires unlabeled samples to keep enough instance-level structure for latent categories to separate; over-compression and under-compression are the two extremes of this axis. The training objective balances the two ends only implicitly, so a trained model can land on either extreme (Fig. 1), and single-end corrections Liu et al. (2025b); Tang et al. (2025) likewise cover one end at a time. What GCD lacks is a representation-level mechanism that measures and controls this balance.

To this end, we propose **InfoSculpt**, which reformulates the old-new tension of GCD as an information-allocation problem and restores this balance. InfoSculpt quantifies the allocation through **Conditional Mutual Information (CMI)** Yang et al. (2024)—the input variability surviving in the representation beyond what the conditioning explains—and instantiates both ends from one principle: a **Category-level CMI** that contracts the spectral volume of class-conditional manifolds, compressing intra-class nuisance toward the discriminative end, and an **Instance-level CMI** that compresses augmentation orbits while preserving instance identity, protecting the discoverable end. Because the two ends inevitably conflict, a game-theoretic **Nash bargaining equilibrium** sets their trade-off adaptively in every batch.

InfoSculpt comes with a practical and rigorous surrogate. From the balanced Information Bottleneck objective Tishby et al. (2000) and its CMI reformulation, we derive a tractable Gaussian class-conditional variational bound: the Category-level term reduces to penalizing the log-determinant of empirical class covariances, and the Instance-level term to penalizing a covariance over augmentation orbits—two loss terms added to any GCD objective, with a 3×3 KKT fixed-point solved per batch and a single preference parameter q as the only hyperparameter. The training dynamics expose the allocation: the equilibrium prioritizes instance-level protection first (a *discovery phase*), category-level contraction second (a *sculpting phase*)—an ordering that emerges from gradient conflicts during training.

Main contributions are as follows:

- We identify unconstrained information allocation as a root cause of GCD's failures on novel classes, and make it observable through spectral analysis.
- We instantiate this allocation as two complementary CMI objectives, derived from a single variational bound and balanced by a Nash equilibrium, controlled by one preference parameter q.
- We validate the allocation knob causally under matched-seed significance tests, and show that it consistently improves five SOTA methods across seven benchmarks as a cheap post-training refinement.

2 Related Works

2.1 Generalized Category Discovery

GCD Vaze et al. (2022) learns from labeled known classes and unlabeled data containing both known and novel categories. Research has shifted from non-parametric clustering Fei et al. (2022) to parametric classifiers, with SimGCD Wen et al. (2023) establishing a strong baseline; subsequent work refines individual components, including dual-level adaptive pseudo-labeling (ProtoGCD Ma et al. (2025)), meta-visual prompts (AptGCD Zhang et al. (2025)), spatial prompts (SPTNet Wang et al. (2024b)), and mean-shift clustering (CMS Choi et al. (2024)). These task-driven designs leave the old-new competition to the training objective. A second line attacks this competition directly in the classifier: debiased heads against novel-class overconfidence (DeGCD Guo et al. (2025)), reciprocal pseudo-label repair (RLCD Liu et al. (2025a)), and discovery-memory decoupling in continual settings (DYDM Yu et al. (2026)) rebalance *predictions* while the representation's allocation remains unconstrained. Our work moves the fix upstream, into the representation itself.

2.2 Conditional Mutual Information

Closest to us are representation-level GCD proposals, each targeting a single geometric property: BIA Tang et al. (2026) restores isotropy at the collapse end, MTMC Tang et al. (2025) maximizes manifold capacity, gradient coordination Zheng et al. (2026) untangles supervised–unsupervised gradients, and SSR^2 He et al. (2026) optimizes rate reduction for multi-modal GCD. Information-theoretic objectives have appeared in adjacent roles: PIM Chiaroni et al. (2023) maximizes information at the prediction layer, and HiLo Wang et al. (2024a) minimizes mutual information for domain disentanglement. CMI itself Yang et al. (2024) has shaped representations across general learning Yang et al. (2024; 2025), federated class imbalance Hamidi et al. (2024), knowledge distillation Ye et al. (2024); Yang & Ye (2024), multi-view bottlenecks Li et al. (2022); Federici et al. (2020), continual GCD Zhao et al. (2025), and anomaly generalization Chhetri et al. (2025). InfoSculpt extends this line to the old-new allocation of GCD: both allocation ends derive from a single CMI variational bound, and the resulting balance is validated causally under matched-seed tests (Sec. 4.2). A mechanism-level comparison with the above methods is provided in Appendix Table 5, and a broader CMI review in Appendix C.2.

3 Methodology

3.1 Notations

Let $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} = \{1, \ldots, K\}$ be the input and category spaces. In GCD, the category space is partitioned into known classes $\mathcal{C}_{\text{old}}$ and novel classes $\mathcal{C}_{\text{new}}$, with $\mathcal{Y} = \mathcal{C}_{\text{old}} \cup \mathcal{C}_{\text{new}}$. We define $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ as random variables, where Y denotes the semantic category. Let $Z = f_\theta(X) \in \mathcal{Z} \subseteq \mathbb{R}^D$ be the latent representation produced by the encoder, and $\hat{Y} \in \Delta^{K-1}$ be the predicted soft-assignment. Note that $Y \to X \to Z \to \hat{Y}$ forms a Markov chain.

Let $\mathcal{D} = \mathcal{D}_l \cup \mathcal{D}_u$ be the dataset comprising labeled and unlabeled subsets, where $\mathcal{D}_l$ covers $\mathcal{C}_{\text{old}}$ and $\mathcal{D}_u$ is sampled from the mixture over $\mathcal{C}_{\text{old}} \cup \mathcal{C}_{\text{new}}$. A formal statement of the adopted GCD protocol and an illustrative overview are provided in Appendix B and Appendix Fig. 5. We define $P_\theta(\cdot|x) \in \Delta^{K-1}$ as the predictive distribution parameterized by θ. Let $\mathcal{B}, \mathcal{B}_l$ denote the mini-batches sampled from $\mathcal{D}$ and $\mathcal{D}_l$, respectively. We introduce $\omega, \omega' \sim \Omega$ as stochastic augmentation operators. To sculpt the latent manifold, we define $\mathbf{c}_k \in \mathcal{Z}$ as the empirical class prototype (or semantic anchor) for category k, which serves as a variational target in the CMI framework. Here $I(\cdot;\cdot|\cdot)$ represents the conditional mutual information that quantifies the information gap during the latent representation sculpting process.

3.2 Formulate GCD from Information Theory

Measuring the allocation: from IB to a tractable CMI objective. The allocation sketched in Sec. 1 needs a measurable quantity. The Information Bottleneck is the natural candidate: in GCD, the labeled subset covers $\mathcal{C}_{\text{old}}$ while the unlabeled subset is drawn from $\mathcal{C}_{\text{old}} \cup \mathcal{C}_{\text{new}}$, so the representation must balance discriminative compression on $\mathcal{D}_l$ against structural preservation on $\mathcal{D}_u$. To ensure a rigorous derivation, we depart from the following standard assumption in representation learning:

Assumption 1. *The latent representation Z is generated solely from the input X via a parametric encoder f_θ, such that Z is conditionally independent of the ground-truth label Y given X. This induces the Markov chain $Y \to X \to Z$.*

Remark 1. *This holds for standard architectures where f_θ takes only X as input, ensuring $P(Z \mid X, Y) = P(Z \mid X)$ regardless of the supervision used to update θ.*

Under this established Markov chain, the objective is to find an optimal representation Z that balances: (i) minimality, i.e., minimizing $I(X; Z)$ to discard information in X that is irrelevant to Y, and (ii) sufficiency, i.e., maximizing $I(Y; Z)$ to preserve predictive information about the labels. This trade-off is formalized by the IB Lagrangian Tishby et al. (2000):

$$\mathcal{L}_{\text{IB}} = I(X; Z) - \beta I(Y; Z), \tag{1}$$

where β controls the balance between compression and sufficiency. In GCD, this balance governs semantic sufficiency for known classes and structural preservation for novel classes.

Directly applying IB to modern neural networks faces two major challenges Zhuang et al. (2025). First, for deterministic networks, the representation Z is a function of the input X, causing $I(X; Z)$ to degenerate. Second, estimating the difference between two mutual information terms is difficult; errors in individual estimators for $I(X; Z)$ and $I(Y; Z)$ can compound, leading to unstable training dynamics. Recent theoretical advances address these issues by reconsidering the task-nuisance decomposition. Specifically, the conditional mutual information $I(X; Z|Y)$ serves as a tighter upper bound for nuisance information than the original $I(X; Z)$ during early stages of representation learning Tezuka & Namekawa (2021). In GCD, this conditional form aligns with compact class manifolds and category-discriminative variation.

Building on this conditional form, we consider a specific IB configuration to sculpt the latent space Z. We adopt the balanced regime ($\beta = 1$) to use CMI as a tractable surrogate with a clear analytical form. The parameter β regulates the compression-sufficiency trade-off, and setting $\beta = 1$ yields the CMI proxy $I(X; Z \mid Y)$ under the Markov chain $Y \to X \to Z$. The full derivation from the IB objective in Eq. (1) to this CMI form is provided in Appendix D.2. In GCD, this regime centers the optimization on nuisance suppression within each semantic category. Varying β explores different trade-off points, and the balanced CMI-based proxy offers an implementation-friendly framework for sculpting the latent space. Additional theoretical justifications are collected in Appendix D, with a focused discussion of the balanced regime in Appendix D.1.

Practical Estimator. We instantiate the balanced CMI proxy with a Gaussian class-conditional variational distribution $Q_\phi(z|y) = \mathcal{N}(z; \mu_y, \Sigma_y)$, which yields the empirical objective

$$\widehat{\mathcal{U}}_{\text{CMI}} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \left(\|f_\theta(x_i) - \mu_{y_i}\|^2_{\Sigma^{-1}_{y_i}} + \log |2\pi\Sigma_{y_i}| \right) + \mathcal{C}, \tag{2}$$

where $\|v\|^2_{\Sigma^{-1}} = v^\top \Sigma^{-1} v$ is the squared Mahalanobis distance, $\mathcal{C}$ absorbs the optimization-independent entropy term, and μ_y, Σ_y are the class-conditional empirical moments. The detailed derivation from the measure-theoretic CMI expansion to Eq. (2) is provided in Appendix D.3.

3.3 Information Space Sculpting via Dual-level CMI

We base our dual-constraint framework on the variational upper-bound estimator (2), unifying the covariance scaling into a single hyperparameter $\rho > 0$. The two competing kinds of

information of Sec. 1 now become two objectives: supervised semantics on $\mathcal{D}_l$ sets the discriminative end, and label-free consistency on $\mathcal{D}_u$ protects the discoverable end.

Category-Level Manifold Compression. To bridge the sample-wise upper-bound estimator and class-level spectral volume, we consider the optimal variational parameters for each category $k \in \{1, \ldots, K\}$. In GCD, this term acts on the supervised partition.

Proposition 1. *Given centered class-specific features $\mathbf{Z}_k \in \mathbb{R}^{d \times n_k}$ (with empirical mean $\boldsymbol{\mu}_k = \frac{1}{n_k} \sum_{i \in \mathcal{C}_k} \mathbf{z}_i = \mathbf{0}$) and the regularized variational covariance proxy $\Sigma_k = \mathbf{I} + \frac{\rho}{n_k} \mathbf{Z}_k \mathbf{Z}_k^\top$, minimizing the variational CMI upper-bound estimator $\widehat{\mathcal{U}}_{\mathrm{CMI}}$ from Eq. (2) yields the simplified class-level spectral loss $\mathcal{L}_{cat}$.*

Substituting $\boldsymbol{\mu}_{y_i} = \boldsymbol{\mu}_k$ and $\Sigma_{y_i} = \Sigma_k$ into Eq. (2) and using the regularized covariance parameterization under L_2-normalized features yields the coding-rate form adopted in InfoSculpt

$$\mathcal{L}_{\mathrm{cat}}^{(k)} \equiv \frac{1}{2} \log \det \left(\mathbf{I} + \frac{\rho}{n_k} \mathbf{Z}_k \mathbf{Z}_k^\top \right). \tag{3}$$

Summing over all K categories weighted by their empirical frequency n_k/n, we obtain:

$$\mathcal{L}_{\mathrm{cat}} = \sum_{k=1}^{K} \frac{n_k}{2n} \log \det \left(\mathbf{I} + \frac{\rho}{n_k} \mathbf{Z}_k \mathbf{Z}_k^\top \right). \tag{4}$$

The full derivation, including the Taylor expansion of the trace term and the gradient analysis that motivates the adopted coding-rate surrogate under L_2-normalized features, is deferred to Appendix D.4.

In practice, $\mathbf{Z}_k$ is computed within each mini-batch or via momentum-updated class prototypes, providing end-to-end gradients to contract the spectral volume of each class manifold.

Proposition 2 (Spectral view of prototype compactness)**.** *Let $\lambda_{k,1}, \ldots, \lambda_{k,r}$ be the non-zero eigenvalues of the empirical class covariance $\hat{\Sigma}_k = \frac{1}{n_k} \mathbf{Z}_k \mathbf{Z}_k^\top$. The category-level objective can be written as*

$$\mathcal{L}_{\mathrm{cat}}^{(k)} = \frac{1}{2} \sum_{j=1}^{r} \log(1 + \rho \lambda_{k,j}). \tag{5}$$

For isotropic covariance $\hat{\Sigma}_k = \sigma_k^2 \mathbf{I}$, this objective reduces to a monotone function of the within-class variance. For general anisotropic covariance, it controls the full eigenvalue spectrum and contracts the spectral volume of the class-wise manifold.

This proposition clarifies the relationship between InfoSculpt and prototype-based compactness. Prototype losses mainly optimize distances to class anchors. Eq. (5) optimizes the covariance spectrum induced by all samples in a class. The objective therefore captures both average dispersion and directional redundancy, aligning the optimization target with the singular-value failures observed in Fig. 1.

Instance-Level Geometric Invariance. To extend our framework to unlabeled data, we formulate a local instance-specific manifold $\mathcal{M}_{x_i}$ induced by stochastic transformations $\omega \sim \Omega$, which serve as a sufficient proxy for non-semantic nuisance factors Tishby et al. (2000); Chen et al. (2020). In GCD, this term is applied on $\mathcal{D}_u$. We then define the instance-level CMI as $I(\omega; Z|X)$, aiming to compress transformation-specific information and preserve the identity of x_i. By modeling the local latent distribution $P(Z|x_i)$ as a Gaussian variational proxy, we derive $\mathcal{L}_{\mathrm{inst}}$ as a tractable upper bound for the instance CMI. This objective balances point-wise consistency ($\mathcal{L}_{\mathrm{inst}}^1$) and local volume compression ($\mathcal{L}_{\mathrm{inst}}^2$):

$$\mathcal{L}_{\mathrm{inst}} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} [\underbrace{\frac{1}{2M} \sum_{j=1}^{M} \|z_{i,j} - \bar{z}_i\|^2}_{\mathcal{L}_{\mathrm{inst}}^1} + \underbrace{\frac{1}{2} \log \det \left(\mathbf{I} + \frac{\rho}{M} \mathbf{Z}_{\omega,i} \mathbf{Z}_{\omega,i}^\top \right)}_{\mathcal{L}_{\mathrm{inst}}^2}]. \tag{6}$$

where $\bar{z}_i = \frac{1}{M}\sum_{j=1}^{M} z_{i,j}$ is the empirical mean of M augmented views. Geometrically, $\mathcal{L}^1_{\text{inst}}$ acts as a centering force to enforce transformation-invariance. $\mathcal{L}^2_{\text{inst}}$ serves as a "spectral chisel" that penalizes the thickness of the local manifold along non-semantic dimensions. Minimizing (6) forces the latent representation to collapse toward a locally low-rank structure, achieving a state of geodesic invariance where features remain stable along transformation orbits. In GCD, this label-free compression complements the supervised category-level term on $\mathcal{D}_u$. The full derivation of Eq. (6) is provided in Appendix D.5.

Balancing Objectives via Nash-MTL. Applied as post-training refinement, Information Space Sculpting (ISS) must balance $\mathcal{L}_{\text{base}}$ with the CMI-based IB proxies $\mathcal{L}_{\text{cat}}$ and $\mathcal{L}_{\text{inst}}$. In GCD, their gradient scales can differ substantially, making standard weighting strategies (e.g., static or uncertainty-based Kendall et al. (2018)) prone to Pareto-suboptimal convergence or gradient domination. We therefore adopt a game-theoretic formulation based on Nash bargaining Navon et al. (2022).

We treat each objective as an independent player seeking a joint update direction:

$$\Delta\theta = \sum_{i\in\mathcal{T}} \alpha_i g_i, \qquad \mathcal{T} = \{\text{base}, \text{cat}, \text{inst}\}, \tag{7}$$

where $g_i = \nabla_\theta \mathcal{L}_i$ denotes the gradient of task i and $\boldsymbol{\alpha} \in \mathbb{R}^{|\mathcal{T}|}$ are non-negative weights lying on the probability simplex. The optimal weight vector $\boldsymbol{\alpha}^\star$ is defined as the unique Nash Bargaining Solution (NBS) obtained by maximizing the aggregate log-utility:

$$\boldsymbol{\alpha}^\star = \arg\max_{\boldsymbol{\alpha}} \sum_{i\in\mathcal{T}} \log\left(g_i^\top \Delta\theta\right), \quad \text{s.t. } \alpha_i \geq 0,\ \sum_{i\in\mathcal{T}} \alpha_i = 1, \tag{8}$$

where feasibility of the solution guarantees $g_i^\top \Delta\theta > 0$ for all i. The logarithmic objective serves as a barrier function, preventing any single task from being disproportionately penalized.

By applying the Karush–Kuhn–Tucker (KKT) conditions Gordon & Tibshirani (2012), the optimal weights are characterized by the nonlinear system

$$G^\top G\boldsymbol{\alpha} = \boldsymbol{\alpha}^{-1}, \tag{9}$$

where $G = [g_{\text{base}}, g_{\text{cat}}, g_{\text{inst}}] \in \mathbb{R}^{d\times 3}$ is the matrix of task gradients and $\boldsymbol{\alpha}^{-1}$ denotes the element-wise reciprocal of $\boldsymbol{\alpha}$. The objective in Eq. (8) is strictly concave over the simplex, ensuring the existence and uniqueness of the global optimum under the positive semi-definite matrix $G^\top G$. Given the small dimensionality of the system, we solve Eq. (9) efficiently using the iterative projection method proposed in Navon et al. (2022), which converges rapidly in practice.

This Nash-MTL strategy offers two key guarantees: (i) *Pareto optimality*, which identifies a balanced update direction across objectives; and (ii) *scale invariance*, which normalizes gradients across heterogeneous magnitudes. Together, these properties enable ISS to navigate the information-theoretic frontier between representational minimality and semantic sufficiency. CMI minimization encourages manifold compression, and InfoSculpt maintains representation richness through a high-rank pre-trained initialization and adaptive Nash-MTL balancing with the discriminative base loss ($\mathcal{L}_{\text{base}}$). This coordinated optimization keeps the latent space informative and precisely sculpted for GCD. A formal analysis of sufficiency anchoring and optimization stability is provided in Appendix D.6.

4 Experiments

We evaluate InfoSculpt along the chain of Sec. 1: the gap that motivates refinement (Sec. 4.1), the causal effect of the allocation knob (Sec. 4.2), and the contribution of each component (Sec. 4.3).

Datasets. We evaluate InfoSculpt on seven benchmark datasets commonly used in GCD, including CIFAR-10/100 Krizhevsky (2009), ImageNet-100 Tian et al. (2020), CUB-200-2011 Wah et al. (2011), Stanford Cars Krause et al. (2013), FGVC-Aircraft Maji et al. (2013),

Table 1: Comparison of InfoSculpt as a post-training refinement module with SOTA approaches on CIFAR-10, CIFAR-100, and ImageNet-100.

Models	CIFAR-10			CIFAR-100			ImageNet-100		
	All	Old	New	All	Old	New	All	Old	New
GCD (CVPR'22)	91.5	97.9	88.2	73.0	76.2	66.5	74.1	89.8	66.3
GCD + Ours	96.9,±0.0	95.1,±0.2	96.6,±0.0	80.4,±0.2	80.4,±0.0	76.8,±0.4	81.3,±0.6	90.9,±0.9	76.7,±1.3
PIM (ICLR'23)	94.7	97.4	93.3	78.3	84.2	66.5	83.1	95.3	77.0
DCCL (CVPR'23)	96.3	96.5	96.9	75.3	76.8	70.2	80.5	90.5	76.2
GPC (ICCV'23)	92.2	**98.2**	89.1	77.9	**85.0**	63.0	76.9	94.3	71.0
SimGCD (ICCV'23)	97.1	95.1	98.1	80.1	81.2	77.8	83.0	93.1	77.9
SimGCD + Ours	97.4,±0.0	95.3,±0.2	97.4,±0.0	80.3,±0.1	82.3,±0.0	78.0,±0.4	84.2,±0.6	93.0,±0.9	80.1,±1.2
LegoGCD (CVPR'24)	97.1	94.3	98.5	81.8	81.4	82.5	86.3	94.5	82.1
ActiveGCD (CVPR'24)	93.2	94.6	92.8	71.3	75.7	66.8	83.3	90.2	76.5
MTMC (arXiv'25)	-	-	-	80.2	81.5	77.5	86.7	93.1	83.6
ProtoGCD (TPAMI'25)	97.3	95.3	98.2	81.9	82.9	80.0	84.0	92.2	79.9
ProtoGCD + Ours	97.5,±0.0	95.1,±0.2	**98.7**,±0.0	82.0,±0.2	83.5,±0.0	78.9,±0.4	85.2,±0.6	93.7,±0.8	79.1,±1.3
FAIR (AAAI'26)	97.6	-	98.1	84.2	-	86.6	86.0	-	81.7
LEAH (AAAI'26)	97.5	-	-	82.1	-	-	84.6	-	-
RLCD (ICML'25)	97.4	96.4	97.9	83.4	84.2	81.9	86.9	94.2	83.2
RLCD + Ours	97.5,±0.0	95.6,±0.2	**98.7**,±0.0	83.5,±0.2	84.7,±0.1	82.1,±0.3	**87.1**,±0.6	94.2,±0.9	84.3,±1.3
AptGCD (CVPR'25)	98.7	95.8	97.3	85.5	81.8	**82.8**	84.3	**95.4**	87.8
AptGCD + Ours	**98.7**,±0.0	96.0,±0.2	97.2,±0.0	**85.7**,±0.2	82.0,±0.0	82.7,±0.4	84.5,±0.6	95.3,±0.9	**88.0**,±1.3

Table 2: Comparison of InfoSculpt as a post-training refinement module with SOTA approaches on CUB, Stanford Cars, FGVC-Aircraft and Herbarium 19.

Models	CUB			Stanford Cars			FGVC-Aircraft			Herbarium 19		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New
GCD (CVPR'22)	51.3	56.6	48.7	39.0	57.6	29.9	45.0	41.1	46.9	35.4	51.0	27.0
GCD + Ours	60.7,±0.1	66.9,±0.4	56.1,±0.2	53.3,±0.2	70.9,±0.4	47.1,±0.6	53.3,±0.4	57.3,±0.6	50.3,±1.0	44.7,±0.2	57.2,±0.4	38.3,±0.4
PIM (ICLR'23)	62.7	75.7	56.2	43.1	66.9	31.6	-	-	-	42.3	56.1	34.8
DCCL (CVPR'23)	63.5	60.8	64.9	43.1	55.7	36.2	-	-	-	-	-	-
GPC (ICCV'23)	55.4	58.2	53.1	42.8	59.2	32.8	46.3	42.5	47.9	-	-	-
SimGCD (ICCV'23)	60.3	65.6	57.7	53.8	71.9	45.0	54.2	59.1	51.8	44.0	58.0	36.4
SimGCD + Ours	61.5,±0.1	66.8,±0.4	57.0,±0.2	55.0,±0.2	73.1,±0.4	45.2,±0.6	54.4,±0.4	59.3,±0.6	52.1,±1.0	48.2,±0.2	60.2,±0.4	36.4,±0.4
LegoGCD (CVPR'24)	63.8	71.9	59.8	57.3	75.7	48.4	55.0	61.5	51.7	45.1	57.4	38.4
ActiveGCD (CVPR'24)	66.6	66.5	66.7	48.4	57.7	39.3	53.7	51.5	56.0	-	-	-
MTMC (arXiv'25)	62.1	65.8	60.3	52.3	70.0	43.7	55.1	58.9	53.1	45.6	57.8	39.0
Hyp-GCD (CVPR'25)	61.0	67.0	58.0	50.8	60.9	45.8	48.2	43.6	50.5	-	-	-
ProtoGCD (TPAMI'25)	65.2	68.5	60.5	53.8	73.7	44.2	56.8	62.5	53.9	44.5	59.4	36.5
ProtoGCD + Ours	66.3,±0.1	71.6,±0.4	60.6,±0.1	59.3,±0.1	78.6,±0.4	44.5,±0.7	56.9,±0.4	64.6,±0.6	53.5,±0.9	**48.6**,±0.2	**62.5**,±0.4	36.8,±0.4
FAIR (AAAI'26)	**73.7**	-	73.0	63.6	-	60.4	-	-	-	49.1	-	43.8
LEAH (AAAI'26)	69.4	-	-	65.5	-	-	58.3	-	-	-	-	-
RLCD (ICML'25)	70.0	79.1	65.4	64.9	79.3	58.0	60.6	62.2	59.8	46.4	61.2	38.4
RLCD + Ours	70.1,±0.1	**79.1**,±0.4	65.5,±0.2	**66.0**,±0.2	79.0,±0.4	59.1,±0.6	**60.7**,±0.4	63.2,±0.6	59.5,±1.0	47.5,±0.2	62.3,±0.4	40.5,±0.3
AptGCD (CVPR'25)	69.2	74.3	**70.3**	53.6	79.7	62.1	59.0	65.2	61.1	42.3	60.6	48.4
AptGCD + Ours	**70.3**,±0.1	75.4,±0.4	70.0,±0.2	58.7,±0.2	**79.8**,±0.5	**62.1**,±0.7	60.1,±0.4	**65.3**,±0.6	**62.1**,±1.0	44.4,±0.2	61.7,±0.4	**48.4**,±0.4

and Herbarium 19 Tan et al. (2019). Further experimental details are provided in Appendix E, including dataset splits and the evaluation protocol in Appendix E.1 and Appendix Table 6, implementation details in Appendix E.2, the step-by-step training procedure in Appendix Algorithm 1, and limitations and future directions in Appendix E.3.

4.1 Observation: downstream remedies leave a residual gap

We compare InfoSculpt against SOTA approaches, including earlier methods (GCD, PIM Chiaroni et al. (2023), DCCL Pu et al. (2023), GPC Zhao et al. (2023), SimGCD), more recent ones (LegoGCD Cao et al. (2024), ActiveGCD Ma et al. (2024), MTMC Tang et al. (2025), Hyp-GCD Liu et al. (2025b), ProtoGCD Ma et al. (2025), RLCD Liu et al. (2025a), AptGCD Zhang et al. (2025)), and the latest 2026 methods FAIR Bian et al. (2026) and LEAH Duan et al. (2026). Before any refinement, novel classes lag known classes by 11.1 pp on average across the 35 method–dataset pairs of Tables 1 and 2, and by over 20 pp on Stanford Cars for every base method. InfoSculpt is integrated as post-training refinement into five of them (GCD, SimGCD, ProtoGCD, RLCD, AptGCD). It raises All accuracy by +2.3 pp on average (up to +17.2 pp) and ranks best among all compared methods on five of the seven benchmarks: CIFAR-10 (98.7%, with AptGCD), CIFAR-100 (85.7%, with AptGCD), ImageNet-100 (87.1%, with RLCD), Stanford Cars (59.3%, with ProtoGCD), and FGVC-Aircraft (56.9%, with ProtoGCD). On fine-grained data, RLCD + InfoSculpt reaches 66.0% on Stanford Cars, and ProtoGCD + InfoSculpt improves Herbarium 19 by 4.1%. The refinement adds 50 epochs on a converged base model (1.18× end-to-end training, Appendix F) with zero inference-time and memory overhead.

Table 3: Ablation study of *InfoSculpt* loss components across benchmarks.

Components			Stanford Cars			Herbarium 19		
L_{base}	L_{cat}	L_{inst}	All	Old	New	All	Old	New
✓			53.8	73.7	44.2	44.5	59.4	36.5
✓	✓		55.1	75.3	43.3	45.4	60.7	36.3
✓		✓	53.9	73.0	**45.5**	43.2	56.7	**36.9**
✓	✓	✓	**59.3**	**78.6**	44.0	**48.6**	**62.5**	36.8

Table 4: Comparison of different multi-task balancing strategies on GCD task.

Strategy	Stanford Cars				Herbarium 19			
	All	Old	New	Time	All	Old	New	Time
Equal Weighting	50.8	69.5	37.6	**1.00**×	40.1	53.8	28.4	**1.00**×
Uncertainty Weighting Kendall et al. (2018)	53.4	72.1	39.2	1.02×	42.7	56.2	30.5	1.01×
GradNorm Chen et al. (2018)	56.1	75.3	41.5	1.15×	45.2	59.4	33.1	1.12×
Nash-MTL (Ours)	**59.3**	**78.6**	**44.0**	1.06×	**48.6**	**62.5**	**36.8**	1.05×

4.2 Improving Novel-Category Discovery with an Oriented Preference

The symmetric Nash objective (q=1) balances known-class stability against novel-class gains. To serve the discovery goal directly, we introduce a preference parameter on the label-independent Instance-CMI utility, $\log u_{\text{base}} + \log u_{\text{cat}} + q \log u_{\text{inst}}$, with task weights still solved adaptively every batch. Using only old-class validation data, we pre-specify $q \in \{1, 1.25, 1.5, 2\}$ and select the largest q meeting a 0.5-pp *Old*-validation non-inferiority rule on RLCD/CIFAR-100; the resulting $q^\star$=1.75 is frozen across methods and datasets, and the calibration pair is excluded from confirmatory tests. Under a five matched-seed protocol (paired checkpoints, splits, and environment; Holm-corrected tests; Appendix G), $q^\star$ improves *New* accuracy over matched baselines on all 14 method–dataset pairs by +1.3 to +13.2 pp (+6.07 on average), with *New* superiority significant in 13/13 held-out pairs (Holm-adjusted $p \leq .008$) and *Old* non-inferiority holding in 11/13. Turning the preference toward discovery also narrows this gap: the difference between known- and novel-class accuracy shrinks in all 14 pairs, by 5.1 pp on average. Modest novel-class gains under q=1 thus reflect the symmetric preference of the default setting.

4.3 Elimination: single-end and static alternatives fail

We assess the two eliminations in turn: each *InfoSculpt* loss component (Table 3) and the Nash-MTL balancer against static alternatives (Table 4), using ProtoGCD as the baseline (L_{base}). The two CMI terms are complementary: L_{cat} improves *Old* via intra-class compactness and L_{inst} benefits *New* via transformation-invariant features; combining them yields clear gains across benchmarks.

Fig. 2 presents a spectral case study that decomposes L_{inst} into an invariance term (L^1_{inst}) and a local volume regularizer (L^2_{inst}). For known categories (top row), L_{cat} suppresses redundant noise and enforces intra-class compactness; for novel categories similar to known ones (middle row), L^1_{inst} counteracts dimensional collapse by preserving the high-rank structure needed for category separation; for out-of-distribution novel categories (bottom row), L^2_{inst} prunes long-tail spectral noise to induce a structured, low-rank manifold. The middle and bottom rows correspond to the two pathologies of Fig. 1: collapse and buried cues, each addressed by the corresponding term above. Together they keep the latent space minimal for known classes yet expressive for novel discovery.

Table 4 compares Nash-MTL against common multi-task weighting strategies: it achieves the best accuracy across metrics—ruling out the static-weighting alternative in our audit—at near-instantaneous per-iteration cost, since the low-dimensional KKT system adds only two auxiliary objectives to the base method.

5 Why Does Allocation Balancing Work?

The training dynamics show the mechanism at work. Appendix Fig. 3 (Left) traces the adaptive weights $\boldsymbol{\alpha}^\star$ on Stanford Cars during the 50-epoch refinement phase: starting from a pre-trained ProtoGCD backbone, Nash-MTL first prioritizes α_{inst} (a **Discovery Phase** enforcing instance-level invariance and preventing collapse for novel categories), then shifts to α_{cat} (a **Sculpting Phase** of intra-class compression), before reaching a **Nash Equilibrium**—an ordering that emerges from Nash-MTL utilities and gradient conflicts.

Appendix Fig. 3 (Right) plots the representation trajectory over 250 epochs (old/new-class accuracy as proxies for compression and generalization): refinement activates once ProtoGCD

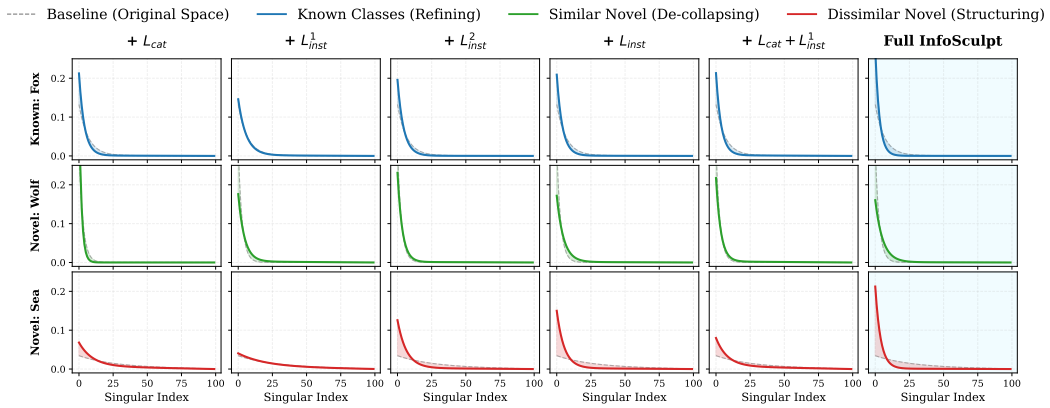

Figure 2: **Spectral Analysis of Latent Space Sculpting.** We analyze three representative categories from the CIFAR-100 dataset to illustrate the fine-grained effects of L_{cat} and L_{inst}.

converges at epoch 200 and immediately moves from a redundant, suboptimal region toward the "Minimal Sufficient Representation" regime.

5.1 Visualization

Appendix Fig. 4 shows CIFAR-10 t-SNE of feature and logit spaces: in the feature space (a–c), GCD+InfoSculpt yields clearly more compact and separated clusters than DINO or GCD; in the logit space (d–f), it produces highly compact, well-separated logits. Dual-level CMI thus structures the latent space at both levels.

6 Conclusion

We introduce InfoSculpt, a simple and effective post-training refinement for generalized category discovery, which recognizes known classes and discovers novel ones from unlabeled data containing both. InfoSculpt targets a root limitation of current GCD methods: training leaves the allocation between discriminative and discoverable information unconstrained, so novel-class features drift toward either over-compression or under-compression, and existing remedies act at the classifier output, downstream of this allocation. InfoSculpt restores the allocation from a single variational bound: a category-level CMI term contracts class-conditional manifolds for discrimination, an instance-level CMI term compresses augmentation orbits to preserve discoverable structure, and a Nash equilibrium balances the two under a single preference parameter. Integrated into five SOTA methods, InfoSculpt consistently improves novel-class accuracy under matched-seed significance tests while preserving old-class performance, raising overall accuracy across seven benchmarks at modest training cost and zero inference overhead. By making the old-new information allocation explicit and controllable, InfoSculpt offers a practical step toward reliable open-world category discovery.

Reproducibility statement

All datasets used in this paper (CIFAR-10/100, ImageNet-100, CUB, Stanford Cars, FGVC-Aircraft, and Herbarium 19) are publicly available, and all compared GCD methods are built on their official implementations. Full implementation details, hyperparameter configurations, derivation of the theoretical results, and additional experimental protocols are provided in the technical appendices. The complete source code and training scripts for InfoSculpt will be released upon publication.

AI USAGE DISCLOSURE

We used LLM-based assistants for language editing, figure preparation, and code maintenance under the authors' direction. All experiments, analyses, and claims were designed and verified by the authors, who take full responsibility for this work.

REFERENCES

Alessandro Achille and Stefano Soatto. Information dropout: Learning optimal representations through noisy computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2897–2905, 2018a. doi: 10.1109/TPAMI.2017.2784440.

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(50):1–34, 2018b.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Yuwei Bian, Shidong Wang, Yazhou Yao, and Haofeng Zhang. Foundation-adaptive integrated refinement for generalized category discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2453–2461, 2026.

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

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pp. 794–803. PMLR, 2018.

Rajeeb Thapa Chhetri, Zhixiong Chen, and Saurab Thapa. Latent sculpting for zero-shot generalization: A manifold learning approach to out-of-distribution anomaly detection. *arXiv preprint arXiv:2512.22179*, 2025.

Florent Chiaroni, Jose Dolz, Ziko Imtiaz Masud, Amar Mitiche, and Ismail Ben Ayed. Parametric information maximization for generalized category discovery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1729–1739, 2023.

Sua Choi, Dahyun Kang, and Minsu Cho. Contrastive mean-shift learning for generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23094–23104, 2024.

Yu Duan, Junzhi He, Zhanxuan Hu, Mengda Ji, Rong Wang, and Quanxue Gao. Learning intrinsic hierarchy for generalized category discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20950–20958, 2026.

Mhairi Dunion, Trevor McInroe, Kevin Luck, Josiah Hanna, and Stefano Albrecht. Conditional mutual information for disentangled representations in reinforcement learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 80111–80129. Curran Associates, Inc., 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/fd750154df5f199f94df897975621306-Paper-Conference.pdf`.

Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck, 2020. URL `https://arxiv.org/abs/2002.07017`.

Yixin Fei, Zhongkai Zhao, Siwei Yang, and Bingchen Zhao. Xcon: Learning with experts for fine-grained category discovery. *arXiv preprint arXiv:2208.01898*, 2022.

Chuanxing Geng, Sheng-jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3614–3631, 2020.

Geoff Gordon and Ryan Tibshirani. Karush-kuhn-tucker conditions. *Optimization*, 10(725/36):725, 2012.

Pengcheng Guo, Yonghong Song, and Boyu Wang. Towards debiased generalized category discovery. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 5271–5279, 2025.

Shayan Mohajer Hamidi, Renhao Tan, Linfeng Ye, and En-Hui Yang. Fed-it: Addressing class imbalance in federated learning through an information-theoretic lens. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pp. 1848–1853. IEEE, 2024.

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

Wei He, Xianghan Meng, Zhiyuan Huang, Xianbiao Qi, Rong Xiao, and Chun-Guang Li. Multi-modal representation learning via semi-supervised rate reduction for generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 39637–39646, 2026.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7482–7491, 2018.

Virginia Klema and Alan Laub. The singular value decomposition: Its computation and some applications. *IEEE Transactions on automatic control*, 25(2):164–176, 1980.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.

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

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

Bo Li, Yifei Shen, Yezhen Wang, Wenzhen Zhu, Dongsheng Li, Kurt Keutzer, and Han Zhao. Invariant information bottleneck for domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7399–7407, 2022.

Duo Liu, Zhiquan Tan, Linglan Zhao, Zhongqiang Zhang, Xiangzhong Fang, and Weiran Huang. Generalized category discovery via reciprocal learning and class-wise distribution regularization. *arXiv preprint arXiv:2506.02334*, 2025a.

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

Shijie Ma, Fei Zhu, Zhun Zhong, Xu-Yao Zhang, and Cheng-Lin Liu. Active generalized category discovery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16890–16900, 2024.

Shijie Ma, Fei Zhu, Xu-Yao Zhang, and Cheng-Lin Liu. Protogcd: Unified and unbiased prototype learning for generalized category discovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

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

Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-task learning as a bargaining game. *arXiv preprint arXiv:2202.01017*, 2022.

Zhengyuan Peng, Jinpeng Ma, Zhimin Sun, Ran Yi, Haichuan Song, Xin Tan, and Lizhuang Ma. Mos: Modeling object-scene associations in generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15118–15128, 2025.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5171–5180. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/poole19a.html.

Nan Pu, Zhun Zhong, and Nicu Sebe. Dynamic conceptional contrastive learning for generalized category discovery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7579–7588, 2023.

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

Luyao Tang, Kunze Huang, Chaoqi Chen, and Cheng Chen. Generalized category discovery via token manifold capacity learning. *arXiv preprint arXiv:2505.14044*, 2025.

Luyao Tang, Kunze Huang, Chaoqi Chen, and Cheng Chen. Bures-isotropy alignment: Manifold learning of generalized category discovery. In *International Conference on Learning Representations (ICLR)*, 2026.

Taro Tezuka and Shizuma Namekawa. Information bottleneck analysis by a conditional mutual information bound. *Entropy*, 23(8):974, 2021.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pp. 776–794. Springer, 2020.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pp. 1–5. Ieee, 2015.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

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

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Hongjun Wang, Sagar Vaze, and Kai Han. Hilo: A learning framework for generalized category discovery robust to domain shifts. *arXiv preprint arXiv:2408.04591*, 2024a.

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

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

Qiyu Xu, Zhanxuan Hu, Yu Duan, Ercheng Pei, and Yonghang Tai. A hidden stumbling block in generalized category discovery: Distracted attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

En-hui Yang and Linfeng Ye. Markov knowledge distillation: Make nasty teachers trained by self-undermining knowledge distillation fully distillable. In *European Conference on Computer Vision*, pp. 154–171. Springer, 2024.

En-Hui Yang, Shayan Mohajer Hamidi, Linfeng Ye, Renhao Tan, and Beverly Yang. Conditional mutual information constrained deep learning: Framework and preliminary results. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pp. 569–574. IEEE, 2024.

En-Hui Yang, Shayan Mohajer Hamidi, Linfeng Ye, Renhao Tan, and Beverly Yang. Conditional mutual information constrained deep learning for classification. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

Linfeng Ye, Shayan Mohajer Hamidi, Renhao Tan, and En-Hui Yang. Bayes conditional distribution estimation for knowledge distillation based on conditional mutual information. *arXiv preprint arXiv:2401.08732*, 2024.

Jiawei Yu, Zijian Gao, Xingxing Zhang, Xuan Liu, Huaimin Wang, and Kele Xu. Decouple your discovery and memory in continual generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25022–25031, 2026.

Wei Zhang, Baopeng Zhang, Zhu Teng, Wenxin Luo, Junnan Zou, and Jianping Fan. Less attention is more: Prompt transformer for generalized category discovery. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 30322–30331, 2025.

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

Zihao Zhao, Xiao Li, Zhonghao Chang, and Ningge Hu. Multi-view contrastive learning with maximal mutual information for continual generalized category discovery. *Expert Systems with Applications*, 266:125994, 2025.

Haiyang Zheng, Nan Pu, Yaqi Cai, Teng Long, Wenjing Li, Nicu Sebe, and Zhun Zhong. The devil is in gradient entanglement: Energy-aware gradient coordinator for robust generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3563–3573, 2026.

Man Zhou, Keyu Yan, Jie Huang, Zihe Yang, Xueyang Fu, and Feng Zhao. Mutual information-driven pan-sharpening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1798–1808, June 2022.

Fei Zhu, Shijie Ma, Zhen Cheng, Xu-Yao Zhang, Zhaoxiang Zhang, and Cheng-Lin Liu. Open-world machine learning: A review and new outlooks. *arXiv e-prints*, pp. arXiv–2403, 2024.

Tianqu Zhuang, Hongyao Yu, Yixiang Qiu, Hao Fang, Bin Chen, and Shu-Tao Xia. Stealthy shield defense: A conditional mutual information-based approach against black-box model inversion attacks. In *The Thirteenth International Conference on Learning Representations*, 2025.

Technical appendices and supplementary material

A Overview

The supplementary material is organized as follows:

- Section B formulates the baseline GCD setting used throughout the paper.
- Section C extends the related-work discussion.
  - (i) Subsection C.1 presents a mechanism-level differentiation across representative GCD paradigms.
  - (ii) Subsection C.2 reviews conditional mutual information and related information-theoretic formulations.
- Section D provides additional theoretical justifications and derivations.
  - (i) Subsection D.1 explains the balanced regime with $\beta = 1$.
  - (ii) Subsection D.2 derives the transition from IB to CMI.
  - (iii) Subsection D.3 derives the practical estimator used in the main text.
  - (iv) Subsection D.4 presents the derivation of the category-level CMI surrogate.
  - (v) Subsection D.5 presents the derivation of the instance-level CMI objective.
  - (vi) Subsection D.6 analyzes representation stability.
- Section E details the experimental setup.
  - (i) Subsection E.1 summarizes the datasets and evaluation protocol.
  - (ii) Subsection E.2 provides implementation details.
  - (iii) Subsection E.3 discusses limitations and future directions.
- Section F profiles the computational cost of the refinement.
- Section G reports the full five-seed results of the new-oriented preference.

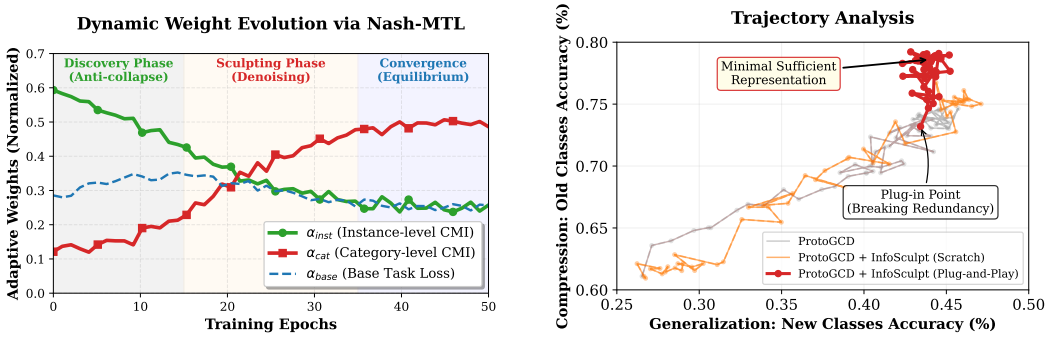

Figure 3: Training dynamics analysis on Stanford Cars. (Left) Evolution of adaptive weights $\boldsymbol{\alpha}^\star$ via Nash-MTL during the 50-epoch refinement phase. (Right) Information plane trajectory comparing different training paradigms.

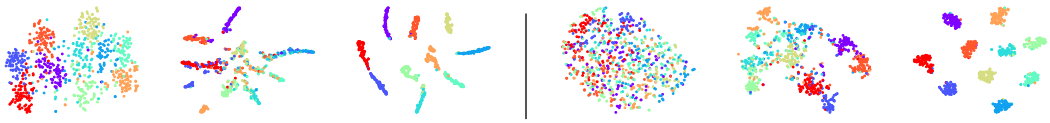

Figure 4: t-SNE visualizations on CIFAR-10. (a,d) DINO; (b,e) GCD; (c,f) GCD+InfoSculpt. (a–c) features and (d–f) logits.

B Baseline Problem Formulation for GCD

Formally, we consider a partially-labeled dataset $\mathcal{D} = \mathcal{D}_l \cup \mathcal{D}_u$. The labeled subset, denoted by $\mathcal{D}_l = \{(\boldsymbol{x}_i^l, y_i^l)\}_{i=1}^n \subset \mathcal{X}_l \times \mathcal{Y}_l$, comprises samples from a set of known, classes $\mathcal{C}_{\text{old}}$ (i.e.,

Table 5: Mechanistic comparison of representative GCD paradigms and InfoSculpt.

Aspect	**Prototype-based GCD (e.g., ProtoGCD)**	**Cross-view GCD (e.g., SimGCD, DINO)**	**InfoSculpt (ours)**
Optimization objective	Enforces sample alignment to semantic prototypes	Promotes agreement between augmented-view representations	Minimizes class-conditional spectral volume via a log-det CMI objective
Optimization scale	Local sample–prototype relations	Paired augmented-view relations	Class-manifold spectra with rank- and volume-aware regularization
Theoretical principle	Metric learning with prototype assignment	Contrastive/self-distillation learning linked to InfoNCE	Balanced IB reformulated as a tractable CMI variational upper bound
Benefit for GCD	Semantic anchors for category assignment	View-invariant features for robust recognition	Spectrum-aware nuisance compression with category-discriminative structure preserved for novel-class discovery

$\mathcal{Y}_l = \mathcal{C}_{\text{old}}$). The unlabeled subset, $\mathcal{D}_u = \{(\boldsymbol{x}_j^u)\}_{j=1}^m \subset \mathcal{X}_u$, contains samples whose underlying label space $\mathcal{Y}_u$ spans both the old classes $\mathcal{C}_{\text{old}}$ and a set of novel classes $\mathcal{C}_{\text{new}}$, such that $\mathcal{Y}_u = \mathcal{C}_{\text{old}} \cup \mathcal{C}_{\text{new}}$.

The objective of GCD is to use the supervision from $\mathcal{D}_l$ to simultaneously cluster the novel-class samples from $\mathcal{C}_{\text{new}}$ and classify the old-class samples from $\mathcal{C}_{\text{old}}$ that are present in $\mathcal{D}_u$. The number of old classes, $K_{\text{old}} = |\mathcal{C}_{\text{old}}|$, is directly inferred from $\mathcal{D}_l$. The number of novel classes $K_{\text{new}} = |\mathcal{C}_{\text{new}}|$, is assumed to be known a priori. The total number of classes is $K = K_{\text{old}} + K_{\text{new}}$. In this protocol, semantic labels are observed on the known partition, while the unlabeled pool is drawn from both known and novel classes. Figure 5 illustrates this setting by showing that the unlabeled data contains both old and novel categories.

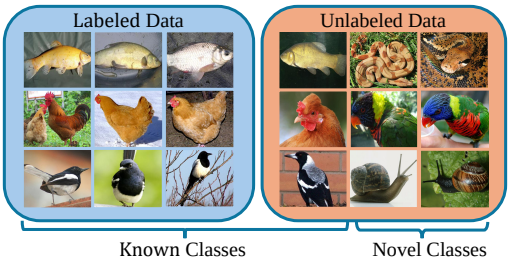

Figure 5: Generalized Category Discovery (GCD). Unlike traditional settings, GCD assumes the unlabeled data is a mixture of known and novel classes, which poses a core challenge for model learning.

C Extended Related Work

C.1 Mechanism-Level Differentiation Across GCD Paradigms

C.2 Extended Review of Conditional Mutual Information

Conditional Mutual Information (CMI) Yang et al. (2024) has become a useful tool for shaping feature spaces with label-aware information constraints. Yang et al. Yang et al. (2024) introduce CMI-constrained deep learning as a general framework, and their later classification study Yang et al. (2025) shows that CMI minimization can tighten class-conditional clusters and improve prediction quality. Hamidi et al. Hamidi et al. (2024) apply the same lens to federated learning and design Fed-IT to alleviate class imbalance across clients. Ye et al. Ye et al. (2024) study Bayes conditional distribution estimation for knowledge distillation under

a CMI objective, while Yang and Ye Yang & Ye (2024) propose Markov knowledge distillation to improve transfer from teachers that are difficult to distill. Dunion et al. Dunion et al. (2023) use CMI to learn disentangled representations in reinforcement learning, and Zhuang et al. Zhuang et al. (2025) formulate a CMI-based defense against black-box model inversion attacks. Related information-theoretic control also appears in pan-sharpening, where Zhou et al. Zhou et al. (2022) use mutual information to guide cross-modal fusion.

The theoretical foundation of InfoSculpt is closely related to variational information-theoretic learning. Tishby et al. Tishby et al. (2000) introduce the Information Bottleneck (IB) principle to characterize representations through compression and prediction sufficiency, and Tishby and Zaslavsky Tishby & Zaslavsky (2015) connect this principle to deep neural networks. Alemi et al. Alemi et al. (2016) develop the Variational Information Bottleneck for tractable training, while Achille and Soatto Achille & Soatto (2018a) propose Information Dropout to learn optimal representations through noisy computation. Achille and Soatto Achille & Soatto (2018b) further analyze how invariance and disentanglement emerge in deep representations under information constraints. Poole et al. Poole et al. (2019) provide a systematic analysis of variational MI bounds and show how estimator behavior depends on the chosen variational family. These studies motivate our use of a Gaussian class-conditional proxy for CMI, which yields a computable upper bound with an efficient spectral parameterization.

CMI and related information objectives have also been used in open-world and robustness-oriented settings. Li et al. Li et al. (2022) propose the Invariant Information Bottleneck for domain generalization and use information constraints to filter pseudo-invariant features. Federici et al. Federici et al. (2020) formulate a multi-view information bottleneck that extracts robust factors shared across views. Zhao et al. Zhao et al. (2025) maximize mutual information in continual GCD to encourage class diversity throughout incremental discovery. Chhetri et al. Chhetri et al. (2025) introduce latent sculpting for zero-shot anomaly generalization and show how geometric compression can organize manifolds for better transfer. These studies demonstrate that conditional and variational information objectives offer a flexible language for representation shaping across tasks. Building on this line of work, InfoSculpt introduces a dual-CMI framework tailored to the unique challenges of GCD, leveraging information-theoretic sculpting to ensure that the representation remains both discriminative for known classes and sufficiently expressive for novel category discovery.

D Additional Theoretical Justification

D.1 Theoretical Justification for the Balanced Regime ($\beta = 1$)

In our formulation of InfoSculpt, we set the IB trade-off parameter to $\beta = 1$. While β is theoretically a tunable hyperparameter, choosing this *balanced regime* is supported by the information decomposition and the GCD protocol.

Optimal Decomposition of Nuisance Information. According to the *nuisance decomposition* of information Achille & Soatto (2018b); Tezuka & Namekawa (2021), under the Markov chain $Y \to X \to Z$, the mutual information admits the exact decomposition

$$I(X; Z) = I(Y; Z) + I(X; Z \mid Y). \tag{10}$$

Substituting this identity into the standard IB objective

$$\mathcal{L}_{\text{IB}} = I(X; Z) - \beta I(Y; Z), \tag{11}$$

yields

$$\mathcal{L}_{\text{IB}} = (1 - \beta) I(Y; Z) + I(X; Z \mid Y). \tag{12}$$

When $\beta = 1$, the first term vanishes and the objective reduces to minimizing the conditional mutual information $I(X; Z \mid Y)$. In GCD, this focus aligns with compact class manifolds and category-discriminative variation.

Avoiding Estimation Error Accumulation. From an optimization perspective, evaluating $\mathcal{L}_{\text{IB}}$ for $\beta \neq 1$ requires separately estimating $I(X; Z)$ and $I(Y; Z)$, each involving

high-dimensional density estimation (e.g., via MINE or CLUB). This leads to accumulated estimator variance,

$$\mathrm{Var}(\widehat{\mathcal{L}}_{\mathrm{IB}}) \approx \mathrm{Var}(\widehat{I}_X) + \beta^2 \mathrm{Var}(\widehat{I}_Y), \tag{13}$$

which often produces unstable gradients and representation collapse in deep networks. In contrast, the $\beta = 1$ regime collapses the objective into a single CMI term. In InfoSculpt, we approximate this term with a Gaussian variational proxy, which leads to the log-determinant form in Proposition 1. This formulation is significantly more stable and avoids compounding estimation errors from multiple mutual information estimators.

D.2 Derivation from IB to CMI

We provide the detailed derivation of the balanced-regime proxy used in the main paper. Starting from the Information Bottleneck Lagrangian,

$$\mathcal{L}_{\mathrm{IB}} = I(X;Z) - \beta I(Y;Z), \tag{14}$$

setting $\beta = 1$ gives

$$\begin{aligned} \mathcal{L}_{\mathrm{IB}}\big|_{\beta=1} &= I(X;Z) - I(Y;Z) && (15)\\ &= [H(Z) - H(Z|X)] - [H(Z) - H(Z|Y)] && (16)\\ &= H(Z|Y) - H(Z|X). && (17) \end{aligned}$$

The second line follows from the standard mutual-information identity $I(A;B) = H(B) - H(B|A)$.

Under the Markov chain $Y \to X \to Z$, the encoder output Z depends on the label Y only through the input X, so the conditional distribution satisfies $P(Z \mid X, Y) = P(Z \mid X)$. This implies

$$H(Z|X) = H(Z|X,Y). \tag{18}$$

Substituting this relation into the previous expression yields

$$\begin{aligned} \mathcal{L}_{\mathrm{IB}}\big|_{\beta=1} &= H(Z|Y) - H(Z|X,Y) && (19)\\ &= I(X;Z \mid Y). && (20) \end{aligned}$$

Therefore, in the balanced regime, minimizing the IB objective is equivalent to minimizing the conditional mutual information $I(X;Z \mid Y)$. This form directly characterizes nuisance variation within each semantic category and provides the tractable information-theoretic proxy adopted by InfoSculpt.

On the role of the Gaussian class-conditional assumption. Following common practice in variational information estimation, our upper bound invokes a Gaussian class-conditional form. We make its status explicit in two regimes. *Exact regime:* when class-conditional features are approximately Gaussian after the projection head and the log-det terms are well-conditioned, the bound is exact up to estimation error of the covariances. *Surrogate regime:* for non-Gaussian features, the Gaussian form serves as a computable surrogate whose minimization still drives the two empirical quantities the bound controls—class-conditional spectral volume and augmentation-orbit information—in the intended direction; the spectral analyses in the main text (Fig. 2) verify directly that these quantities move as predicted.

D.3 Derivation of the Practical Estimator

We provide the detailed derivation of the practical estimator used in the main paper. Starting from the standard measure-theoretic expansion of conditional mutual information,

$$I(X;Z|Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} \int_{\mathcal{Z}} P(x,y,z) \log \frac{P(z|x,y)}{P(z|y)} dz dy dx. \tag{21}$$

By the Markov property $Y \to X \to Z$, the embedding Z is conditionally independent of the label Y given the input X, which implies $P(z|x,y) = P(z|x)$. Substituting this relation into Eq. (21) gives

$$I(X;Z|Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} P(x,y) \int_{\mathcal{Z}} P(z|x) \log \frac{P(z|x)}{P(z|y)} dz dy dx. \tag{22}$$

To express the category-level marginal $P(z|y)$, we apply the law of total probability over the input space $\mathcal{X}$:

$$P(z|y) = \int_{\mathcal{X}} P(z|x,y)P(x|y)dx \tag{23}$$

$$= \int_{\mathcal{X}} P(z|x)P(x|y)dx \tag{24}$$

$$= \mathbb{E}_{X|Y=y}\left[P(z|X)\right]. \tag{25}$$

Given a labeled dataset $\mathcal{D}_l = \{(x_i, y_i)\}_{i=1}^n$, let $\mathcal{D}_y = \{(x_i, y_i) \in \mathcal{D}_l : y_i = y\}$ denote the subset of samples belonging to category y, with $|\mathcal{D}_y|$ its cardinality. Using the empirical class-conditional distribution $P(x|y) \approx \frac{1}{|\mathcal{D}_y|}\sum_{j:y_j=y}\delta(x - x_j)$, we obtain the empirical estimator

$$\bar{P}_\theta(z|y) = \frac{1}{|\mathcal{D}_y|}\sum_{j:y_j=y} P_\theta(z|x_j). \tag{26}$$

We then approximate the joint distribution $P(x, y)$ with the empirical distribution over the labeled set, $P(x, y) \approx \frac{1}{n}\sum_{i=1}^n \delta(x - x_i)\delta(y - y_i)$. Substituting this approximation together with Eq. (26) into Eq. (22) yields

$$I(X;Z|Y) \approx \int_{\mathcal{X}}\int_{\mathcal{Y}} \left(\frac{1}{n}\sum_{i=1}^n \delta(x - x_i)\delta(y - y_i)\right) \int_{\mathcal{Z}} P_\theta(z|x)\log\frac{P_\theta(z|x)}{\bar{P}_\theta(z|y)}dzdydx \tag{27}$$

$$= \frac{1}{n}\sum_{i=1}^n \int_{\mathcal{Z}} P_\theta(z|x_i)\log\frac{P_\theta(z|x_i)}{\bar{P}_\theta(z|y_i)}dz. \tag{28}$$

For a deterministic encoder f_θ, the conditional distribution $P_\theta(z|x_i)$ collapses to $\delta(z - f_\theta(x_i))$, which makes $H(Z|X)$ an optimization-independent constant. Following the Variational Information Bottleneck (VIB) framework Alemi et al. (2016); Achille & Soatto (2018a) and the analysis of variational MI bounds Poole et al. (2019), we introduce a tractable class-conditional variational proxy $Q_\phi(z|y)$ for $P(z|y)$. Since $\mathrm{KL}(P(z|y)\|Q_\phi(z|y)) \geq 0$, replacing the intractable marginal $P(z|y)$ with $Q_\phi(z|y)$ yields the upper bound

$$I(X;Z|Y) = \mathbb{E}_{P(x,y)}\left[\mathrm{KL}\left(P_\theta(z|x)\|P(z|y)\right)\right] \tag{29}$$

$$\leq \mathbb{E}_{P(x,y)}\left[\mathrm{KL}\left(P_\theta(z|x)\|Q_\phi(z|y)\right)\right]. \tag{30}$$

We choose a Gaussian proxy $Q_\phi(z|y) = \mathcal{N}(z; \mu_y, \Sigma_y)$, which provides an efficient spectral parameterization of the class-wise manifold. Substituting $z_i = f_\theta(x_i)$ into Eq. (30) recovers the empirical objective used in the main paper:

$$\widehat{\mathcal{U}}_{\mathrm{CMI}} = \frac{1}{n}\sum_{i=1}^n \frac{1}{2}\left(\|f_\theta(x_i) - \mu_{y_i}\|^2_{\Sigma_{y_i}^{-1}} + \log|2\pi\Sigma_{y_i}|\right) + \mathcal{C}, \tag{31}$$

This expression corresponds to Eq. (2) in the main paper, where $\|v\|^2_{\Sigma^{-1}} = v^\top\Sigma^{-1}v$ is the squared Mahalanobis distance, $\mathcal{C}$ absorbs the optimization-independent entropy term, and μ_y, Σ_y are the class-conditional empirical moments.

D.4 Derivation of the Category-Level CMI Surrogate

We provide an analytical account of the transition from the variational CMI objective to the coding-rate surrogate used in the main paper. The derivation characterizes the roles of the trace and log-determinant terms under the regularized covariance parameterization and motivates the practical category-level objective adopted in InfoSculpt.

Starting from the variational CMI formulation derived in the methodology, the full optimization objective for category k is given by:

$$\mathcal{L}^{(k)}_{cat} = \frac{1}{2\rho}\,\mathrm{tr}\left(\mathbf{I} - (\mathbf{I} + \rho\hat{\Sigma}_k)^{-1}\right) + \frac{1}{2}\log\det\left(\mathbf{I} + \rho\hat{\Sigma}_k\right), \tag{32}$$

where $\hat{\Sigma}_k = \frac{1}{n_k}\mathbf{Z}_k\mathbf{Z}_k^\top$ is the empirical covariance matrix. Since each feature vector z_i is L_2-normalized ($\|z_i\|_2 = 1$), the trace of the covariance matrix remains constant:

$$\operatorname{tr}(\hat{\Sigma}_k) = \operatorname{tr}\left(\frac{1}{n_k}\sum_{i\in\mathcal{C}_k} z_i z_i^\top\right) = \frac{1}{n_k}\sum_{i\in\mathcal{C}_k}\|z_i\|_2^2 = 1. \tag{33}$$

To analyze the contribution of each term to the sculpting of the latent space, we perform a Taylor expansion of the trace term and the log-determinant term with respect to $\rho\hat{\Sigma}_k$:

$$\begin{aligned}
\mathcal{L}_{cat}^{(k)} &= \frac{1}{2\rho}\operatorname{tr}\left(\mathbf{I} - (\mathbf{I} - \rho\hat{\Sigma}_k + \rho^2\hat{\Sigma}_k^2 - O(\rho^3\hat{\Sigma}_k^3))\right) + \frac{1}{2}\operatorname{tr}\left(\rho\hat{\Sigma}_k - \frac{\rho^2}{2}\hat{\Sigma}_k^2 + O(\rho^3\hat{\Sigma}_k^3)\right) \\
&= \frac{1}{2\rho}\operatorname{tr}\left(\rho\hat{\Sigma}_k - \rho^2\hat{\Sigma}_k^2 + O(\rho^3\hat{\Sigma}_k^3)\right) + \frac{1}{2}\operatorname{tr}\left(\rho\hat{\Sigma}_k - \frac{\rho^2}{2}\hat{\Sigma}_k^2 + O(\rho^3\hat{\Sigma}_k^3)\right) \\
&= \underbrace{\left(\frac{1}{2}\operatorname{tr}(\hat{\Sigma}_k) + \frac{\rho}{2}\operatorname{tr}(\hat{\Sigma}_k)\right)}_{\text{Constant w.r.t. } \mathbf{Z}_k \text{ due to } L_2 \text{ norm}} - \underbrace{\left(\frac{\rho}{2}\operatorname{tr}(\hat{\Sigma}_k^2) + \frac{\rho^2}{4}\operatorname{tr}(\hat{\Sigma}_k^2)\right)}_{\text{Sculpting Terms (Manifold Contraction)}} + \dots
\end{aligned} \tag{34}$$

When calculating the gradient with respect to the feature matrix $\mathbf{Z}_k$, we observe from Eq. 33 that $\nabla_{\mathbf{Z}_k}\operatorname{tr}(\hat{\Sigma}_k) = 0$. Consequently, the gradient of the objective is dominated by the higher-order terms:

$$\begin{aligned}
\nabla_{\mathbf{Z}_k}\mathcal{L}_{cat}^{(k)} &= \nabla_{\mathbf{Z}_k}\left[\frac{1}{2\rho}\operatorname{tr}\left(\mathbf{I} - (\mathbf{I} + \rho\hat{\Sigma}_k)^{-1}\right)\right] + \nabla_{\mathbf{Z}_k}\left[\frac{1}{2}\log\det\left(\mathbf{I} + \rho\hat{\Sigma}_k\right)\right] \\
&= \frac{1}{2}(\mathbf{I} + \rho\hat{\Sigma}_k)^{-2}\frac{\partial\hat{\Sigma}_k}{\partial\mathbf{Z}_k} + \frac{\rho}{2}(\mathbf{I} + \rho\hat{\Sigma}_k)^{-1}\frac{\partial\hat{\Sigma}_k}{\partial\mathbf{Z}_k} \\
&\approx \underbrace{\frac{1+\rho}{2}\nabla_{\mathbf{Z}_k}\operatorname{tr}(\hat{\Sigma}_k)}_{=0} - \frac{2\rho+\rho^2}{2}\nabla_{\mathbf{Z}_k}\operatorname{tr}(\hat{\Sigma}_k^2) + O(\rho^3).
\end{aligned} \tag{35}$$

Under the regularized covariance parameterization, the second-order contraction effect $\frac{\rho^2}{2}\operatorname{tr}(\hat{\Sigma}_k^2)$ from the log-determinant term grows more prominently than the corresponding $\rho\operatorname{tr}(\hat{\Sigma}_k^2)$ contribution from the trace term. Given the L_2 normalization constraint, the trace term behaves as a smooth correction in the optimization landscape with a relatively small gradient.

This leads to the coding-rate surrogate used as the category-level objective:

$$\mathcal{L}_{cat}^{(k)} \equiv \frac{1}{2}\log\det\left(\mathbf{I} + \frac{\rho}{n_k}\mathbf{Z}_k\mathbf{Z}_k^\top\right). \tag{36}$$

This surrogate preserves the essential dynamics of manifold volume compression while improving numerical stability during training by focusing the optimization on the discriminative spectral dimensions of the feature space.

D.5 Derivation of Instance-Level CMI ($\mathcal{L}_{\text{INST}}$)

To extend the CMI-based Information Space Sculpting to unlabeled data, we treat each instance x_i as a unique latent "category", and its stochastic augmentations as samples drawn from this category. In GCD, this construction is applied on $\mathcal{D}_u$. Our objective is to minimize the instance-level CMI $I(\omega; Z \mid X_i)$, where ω denotes the nuisance transformation.

Following the variational CMI estimator derived in Eq. (2), for an instance x_i with M augmented views $\{z_{i,j}\}_{j=1}^M$, the local CMI can be approximated using a Gaussian variational proxy $Q(z \mid x_i) = \mathcal{N}(z; \bar{z}_i, \Sigma_i)$, where $\bar{z}_i = \frac{1}{M}\sum_{j=1}^M z_{i,j}$ denotes the empirical mean of the augmentation orbit. The estimator becomes

$$\hat{I}(\omega; Z \mid X_i) \approx \frac{1}{M} \sum_{j=1}^{M} \frac{1}{2} \left(\|z_{i,j} - \bar{z}_i\|^2_{\Sigma_i^{-1}} + \log\det(2\pi\Sigma_i) \right) + \mathcal{C}. \tag{37}$$

Following the derivation in **Proposition 1**, we adopt the optimal regularized covariance for the local manifold $\mathcal{M}_{x_i}$,

$$\Sigma_i = \mathbf{I} + \frac{\rho}{M} \mathbf{Z}_{\omega,i} \mathbf{Z}_{\omega,i}^\top, \tag{38}$$

where $\mathbf{Z}_{\omega,i} = [z_{i,1} - \bar{z}_i, \ldots, z_{i,M} - \bar{z}_i] \in \mathbb{R}^{d \times M}$ is the centered feature matrix of augmentations. Substituting Σ_i into Eq. (37), the quadratic term corresponds to the trace of the empirical covariance weighted by the precision matrix. For L_2-normalized features, as shown in Eq. (3) and Eq. (4) in the main paper, minimizing this term mainly enforces point-wise consistency toward the centroid $\bar{z}_i$. This yields the invariance component

$$\mathcal{L}^1_{\text{inst}} = \frac{1}{2M} \sum_{j=1}^{M} \|z_{i,j} - \bar{z}_i\|^2. \tag{39}$$

Meanwhile, the log-determinant term in Eq. (37) measures the coding rate (i.e., spectral volume) of the augmentation-induced manifold. Substituting the covariance structure above and ignoring constants leads to the local volume compression term

$$\mathcal{L}^2_{\text{inst}} = \frac{1}{2} \log\det \left(\mathbf{I} + \frac{\rho}{M} \mathbf{Z}_{\omega,i} \mathbf{Z}_{\omega,i}^\top \right). \tag{40}$$

Aggregating both components across instances i in a mini-batch $\mathcal{B}$ yields the final tractable objective in Eq. (6) of the main paper

$$\mathcal{L}_{\text{inst}} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \left[\frac{1}{2M} \sum_{j=1}^{M} \|z_{i,j} - \bar{z}_i\|^2 + \frac{1}{2} \log\det \left(\mathbf{I} + \frac{\rho}{M} \mathbf{Z}_{\omega,i} \mathbf{Z}_{\omega,i}^\top \right) \right]. \tag{41}$$

This formulation ensures that InfoSculpt simultaneously enforces transformation invariance via $\mathcal{L}^1_{\text{inst}}$ while sculpting the latent space by suppressing non-semantic directions within each instance's augmentation orbit through $\mathcal{L}^2_{\text{inst}}$.

D.6 Theoretical Analysis on Representation Stability

A fundamental challenge in IB and CMI based optimization is the risk of *representation collapse*. Specifically, in our framework, a trivial way to minimize the sculpting objective $I(X; Z \mid Y)$ is to allow both $I(X; Z)$ and $I(Y; Z)$ to vanish, leading to a degenerate solution where the latent space carries no information. In this section, we provide a rigorous justification for the stability of **InfoSculpt**, demonstrating how the interplay between the task-specific base loss and the Nash-MTL optimization strategy prevents collapse and ensures a semantically structured latent space.

D.6.1 Sufficiency Anchoring via a Variational Lower Bound

The total optimization objective of InfoSculpt is formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{base}} + \mathcal{L}_{\text{sculpt}}, \tag{42}$$

where $\mathcal{L}_{\text{base}}$ is the discriminative task loss (e.g., cross-entropy) and $\mathcal{L}_{\text{sculpt}}$ is the proposed CMI-based sculpting objective. We first show that $\mathcal{L}_{\text{base}}$ serves as an anchor that prevents the loss of class-defining information.

Proposition 3 (Sufficiency Anchoring)**.** *Minimizing the base task loss $\mathcal{L}_{base}$ is equivalent to maximizing a variational lower bound on the predictive sufficiency term $I(Y;Z)$.*

Proof. Let $q_\theta(y \mid z)$ be a variational approximation to the true posterior $p(y \mid z)$. Following the definition of mutual information between the label Y and representation Z:

$$I(Y;Z) = H(Y) - H(Y \mid Z) \tag{43}$$
$$= H(Y) + \mathbb{E}_{p(y,z)}[\log p(y \mid z)]. \tag{44}$$

By introducing the variational distribution $q_\theta(y \mid z)$, we can express $I(Y;Z)$ as:

$$I(Y;Z) = H(Y) + \mathbb{E}_{p(z)}\Big[D_{\mathrm{KL}}\big(p(y \mid z) \,\|\, q_\theta(y \mid z)\big) + \mathbb{E}_{p(y|z)}[\log q_\theta(y \mid z)]\Big] \tag{45}$$
$$\geq H(Y) + \mathbb{E}_{p(y,z)}[\log q_\theta(y \mid z)], \tag{46}$$

where the inequality holds due to the non-negativity of the Kullback-Leibler (KL) divergence. For a discriminative model, the base loss $\mathcal{L}_{\text{base}}$ corresponds to the negative log-likelihood:

$$\mathcal{L}_{\text{base}} = -\mathbb{E}_{p(y,z)}[\log q_\theta(y \mid z)]. \tag{47}$$

Since the label entropy $H(Y)$ is constant for a given dataset, minimizing $\mathcal{L}_{\text{base}}$ effectively maximizes the lower bound on $I(Y;Z)$. This guarantees that Z is anchored to retain sufficient semantic information to distinguish categories, thereby preventing the collapse of $I(Y;Z)$ during CMI minimization. □

D.6.2 Pareto-Optimality via Nash Bargaining

While $\mathcal{L}_{\text{base}}$ provides an informational anchor, the gradients of $\mathcal{L}_{\text{base}}$ (sufficiency) and $\mathcal{L}_{\text{sculpt}}$ (compression) may still conflict in direction or magnitude. To address this, we define the respective gradients as $g_b = \nabla_\theta \mathcal{L}_{\text{base}}$ and $g_s = \nabla_\theta \mathcal{L}_{\text{sculpt}}$. InfoSculpt adopts the Nash Bargaining Solution (NBS) to navigate this multi-objective trade-off.

Proposition 4 (Optimization Stability)**.** *The Nash-MTL update direction $\Delta\theta$ ensures that improvements in representational minimality (compression) do not occur at the expense of sufficiency (predictive power).*

Proof. The Nash Bargaining update $\Delta\theta$ is the solution to the aggregate log-utility maximization problem:

$$\max_{\Delta\theta} \sum_{i \in \{b,s\}} \log(g_i^\top \Delta\theta), \quad \text{subject to feasibility constraints.} \tag{48}$$

Applying the Karush–Kuhn–Tucker (KKT) conditions, the optimal update is a weighted combination:

$$\Delta\theta = \alpha_b g_b + \alpha_s g_s, \quad \text{where } \alpha_b, \alpha_s > 0. \tag{49}$$

The feasibility of the Nash solution strictly requires $g_b^\top \Delta\theta > 0$ and $g_s^\top \Delta\theta > 0$. The condition $g_b^\top \Delta\theta > 0$ ensures that every parameter update has a strictly positive projection onto the sufficiency gradient. This prevents the optimization trajectory from decreasing the mutual information $I(Y;Z)$ that the base loss aims to preserve. Any component of the compression gradient g_s that conflicts with the sufficiency anchor g_b is effectively moderated, ensuring a Pareto-optimal balance. □

D.6.3 Derivation of the Nash-MTL Fixed-Point System

We derive the fixed-point system solved by our Nash-MTL weight solver (Eq. (9) in the main paper). Let $G = [g_{\text{base}}, g_{\text{cat}}, g_{\text{inst}}]$ collect the three task gradients, and let the joint update be $\Delta\theta = G\alpha$ with weights α on the probability simplex ($\alpha_i > 0$, $\sum_i \alpha_i = 1$). Task i improves by the utility $u_i(\alpha) = g_i^\top \Delta\theta = (G^\top G\alpha)_i$. Nash-MTL maximizes the aggregate log-utility $\sum_i \log u_i(\alpha)$, which prevents any task's utility from collapsing to zero:

$$\max_{\alpha} \sum_i \log\left[(G^\top G\alpha)_i\right] \quad \text{s.t.} \quad \alpha_i > 0,\ \mathbf{1}^\top \alpha = 1. \tag{50}$$

Introducing the simplex multiplier λ and the non-negativity multipliers $\mu_i \geq 0$, stationarity of the Lagrangian yields, for every interior solution ($\mu_i = 0$):

$$\frac{(G^\top G)_{i,:}\,\alpha}{(G^\top G\alpha)_i} = \lambda, \quad \forall\, i, \tag{51}$$

i.e., each task's utility is inversely proportional to its weight, $\alpha_i\, u_i(\alpha) = 1/\lambda$. Because α is normalized on the simplex, this common positive constant can be absorbed by the normalization step of the iterative projection solver, giving the normalized fixed point

$$(G^\top G\,\alpha)_i = \alpha_i^{-1} \quad \Longleftrightarrow \quad G^\top G\,\alpha = \alpha^{-1}, \tag{52}$$

where α^{-1} denotes the element-wise inverse. Eq. (52) is exactly the system solved by the iterative projection method in each batch. In practice we solve it with 20 fixed-point iterations, which we empirically find sufficient for convergence (Appendix F reports the solver cost).

D.6.4 Geometric Volume Preservation in Latent Space

Beyond the information-theoretic perspective, the category-level sculpting loss $\mathcal{L}_{\text{cat}}$ provides an explicit geometric mechanism to resist collapse. By viewing the latent space as a collection of class-specific manifolds, we can analyze the effective volume $\mathcal{V}(\mathcal{Z})$.

The contrastive nature of the cross-entropy loss (embedded in $\mathcal{L}_{\text{base}}$) and the structure of our category-level prototypes $\mathbf{c}_i$ enforce a minimum inter-class repulsion:

$$\forall i \neq j, \quad \mathbb{E}\,\|\mathbf{c}_i - \mathbf{c}_j\|^2 \geq \delta, \tag{53}$$

for some margin $\delta > 0$. This geometric constraint prevents the global collapse of the latent space to a single point (Dirac measure). Consequently, the minimization of CMI in InfoSculpt is achieved via *intra-class compression* (i.e., $\text{Var}(Z \mid Y) \to 0$), complementing global variance reduction. This allows the model to "sculpt" a low-rank, compact manifold for each category while maintaining a high-rank, discriminative global structure.

E Experimental Setup

E.1 Datasets and Evaluation Protocol

Datasets. The effectiveness of InfoSculpt is validated on 7 benchmark datasets, consistent with prior work like SimGCD, including CIFAR-10/100 Krizhevsky (2009), ImageNet-100 Tian et al. (2020), as well as fine-grained classification datasets CUB-200-2011 Wah et al. (2011), Stanford Cars Krause et al. (2013), and FGVC-Aircraft Maji et al. (2013). To further assess the robustness of InfoSculpt, we include one more challenging large-scale datasets: Herbarium 19 Tan et al. (2019).

Each dataset is adhered to the standard GCD protocol established in SimGCD. Specifically, we form the labeled training set, $\mathcal{D}^l$, by sub-sampling 50% of the images from the known classes. The remaining images from these known classes, combined with all images from the novel classes, constitute the unlabeled set, $\mathcal{D}^u$. A detailed overview of the dataset splits is in Table 6.

Evaluation Protocol. During training, the entire dataset $\mathcal{D} = \mathcal{D}^l \cup \mathcal{D}^u$ is utilized. For evaluation, InfoSculpt measure performance using the standard unsupervised clustering accuracy (ACC) metric Vaze et al. (2022), computed on all samples in the unlabeled set $\mathcal{D}^u$. Given the ground-truth labels $y_i^\star$ and the model's predicted cluster assignments $\hat{y}_i$, ACC is defined as:

$$\text{ACC} = \frac{1}{M}\sum_{i=1}^{M} \mathbb{1}\left(y_i^\star = \text{map}(\hat{y}_i)\right), \tag{54}$$

where $M = |\mathcal{D}^u|$ is the total number of unlabeled samples, $\mathbb{1}(\cdot)$ is the indicator function, and $\text{map}(\cdot)$ is the optimal permutation mapping function. This mapping is determined using the Hungarian algorithm Kuhn (1955) to find the best one-to-one alignment between the predicted cluster indices and the ground-truth class labels.

Table 6: Overview of the datasets used in the experiments. We list the specific number of labeled and unlabeled images ($\mathcal{D}^l$, $\mathcal{D}^u$) and their corresponding class assignments ("Old" and "New").

Dataset	Labeled $\mathcal{D}^l$		Unlabeled $\mathcal{D}^u$	
	Images	**Old**	**Images**	**New**
CIFAR-10	12.5k	5	37.5k	10
CIFAR-100	20.0k	80	30.0k	100
ImageNet-100	31.9k	50	95.3k	100
CUB-200-2011	1.5k	100	4.5k	200
Stanford Cars	2.0k	98	6.1k	196
FGVC-Aircraft	1.7k	50	5.0k	100
Herbarium 19	8.9k	341	25.4k	683

E.2 Implementation Details

To complement the brief implementation summary in the main text, we provide additional details to ensure the reproducibility of InfoSculpt.

Optimization and Hyperparameters. InfoSculpt is integrated as a late-stage refinement module. For all base models, we follow a two-phase protocol. First, the base models are reproduced using their official implementations and hyperparameter configurations, including optimizer, learning rate schedules, and data augmentation strategies, until convergence. We then load the pre-trained weights and incorporate the dual CMI objectives (L_{cat} and L_{inst}) for refinement over 50 epochs. During this phase, we adopt the Adam optimizer with a reduced learning rate (1×10^{-5} for the backbone and 1×10^{-4} for the classification heads) to preserve the structure of the pre-trained representations. The scaling parameter ρ in Eq. (4) and Eq. (6) of the main paper, which controls the regularization scale of the feature manifold, is fixed to 0.5 across all seven datasets as a shared refinement setting. For the instance-level CMI objective, we set the number of augmentation views to $M = 2$ (two stochastic views per image), balancing computational efficiency and geometric invariance.

Nash-MTL Solver. To mitigate gradient conflicts between the base objective $\mathcal{L}_{base}$ and the proposed CMI proxies, we employ the Nash-MTL strategy to adaptively update the task weights α_{base}, α_{cat}, and α_{inst} at each training iteration. The KKT system in Eq. (9) of the main paper is solved using an iterative projection method with initial weights $\alpha_i = 1/|\mathcal{T}|$. Since the number of objectives is small ($|\mathcal{T}| = 3$), the additional computational overhead is negligible.

Hardware and Environment. All experiments are conducted on a single NVIDIA V100 GPU. The framework is implemented with CUDA 11.8. To ensure reproducibility and fair comparison, the Python and PyTorch versions, as well as the random seeds, follow the settings used in each corresponding baseline for both the baseline reproductions and the InfoSculpt post-training refinement experiments.

E.3 Limitation and Future Work

Limitation:

- **Performance Saturation on Generic Datasets:** InfoSculpt exhibits marginal gains on generic benchmarks like CIFAR-10/100 and ImageNet-100. We attribute this to the **high intrinsic separability** of these datasets, where SOTA baselines (e.g., AptGCD) already achieve near-saturated performance. Since InfoSculpt operates by refining the manifold geometry via information-theoretic constraints, its utility is naturally **bounded** when the initial representation space is already sufficiently discriminative. Its true efficacy is more pronounced in **fine-grained discovery**

Algorithm 1 InfoSculpt: Post-Training Refinement

Require: Dataset $\mathcal{D} = \mathcal{D}_l \cup \mathcal{D}_u$, base model f_θ, hyperparameter ρ, learning rate η, fine-tuning epochs $E_{ft} = 50$.
Ensure: Refined model parameters θ.
{**Stage I: Baseline Training (Training from Scratch)**}
Train f_θ using official settings of the base model until convergence.
Let $\mathcal{L}_{base}$ be the original loss function.
{**Stage II: Information Space Sculpting (Post-training Refinement)**}
Load pre-trained weights from Stage I.
for epoch $= 1$ **to** E_{ft} **do**
  for mini-batch $\mathcal{B} \subset \mathcal{D}$ **do**
    {*1. Multi-view Forward Pass*}
    Generate two augmented views $\{z_{i,1}, z_{i,2}\}$ for each image $x_i \in \mathcal{B}$.
    {*2. Loss Computation*}
    Compute $\mathcal{L}_{base}$ using the official base model objective.
    Compute $\mathcal{L}_{cat}$ via spectral volume of labeled manifolds.
    Compute $\mathcal{L}_{inst}$ via instance-level geometric invariance.
    {*3. Gradient Balancing via Nash-MTL*}
    Compute task-specific gradients: $g_{base} = \nabla_\theta \mathcal{L}_{base}$, $g_{cat} = \nabla_\theta \mathcal{L}_{cat}$, $g_{inst} = \nabla_\theta \mathcal{L}_{inst}$.
    Construct the gradient matrix $G = [g_{base}, g_{cat}, g_{inst}]$.
    Solve the KKT system $G^\top G \boldsymbol{\alpha} = \boldsymbol{\alpha}^{-1}$ using iterative projection to find optimal weights $\boldsymbol{\alpha}^\star = [\alpha_{base}, \alpha_{cat}, \alpha_{inst}]$.
    {*4. Parameter Update*}
    Compute joint update direction: $\Delta\theta = \sum_{i \in \{base,cat,inst\}} \alpha_i g_i$.
    $\theta \leftarrow \theta - \eta \cdot \Delta\theta$ (via Optimizer step).
  end for
end for
return θ

(Table 2), where inter-class boundaries are ambiguous and require explicit manifold "sculpting."

- **Post-Training Refinement Cost:** InfoSculpt is a method-level post-training refinement framework: it introduces two auxiliary objectives and Nash-MTL balancing, which add measurable training-time overhead (Appendix F) and require a 50-epoch refinement phase. While the inference path remains unchanged, this refinement cost should be budgeted in deployment.
- **Fixed Information Trade-off:** The current framework assumes a **balanced regime** ($\beta = 1$) in the IB Lagrangian. While empirically robust, fixing β may not be optimal across the entire Pareto frontier. Datasets with extreme label noise or heavy-tailed distributions might require a more flexible trade-off—either more aggressive compression ($\beta > 1$) to prune noise or enhanced sufficiency ($\beta < 1$) to preserve subtle category-defining signals.

Future Work:

- **Adaptive Sculpting:** A promising extension is to introduce **dynamic β scheduling**, treating the trade-off as a learnable parameter or a function of training progress. This would allow the model to prioritize "information discovery" in early stages and "manifold refinement" in later stages.
- **Algorithmic Efficiency:** While the computational overhead of the log-determinant and Nash-MTL computations is currently manageable, future work will focus on **low-rank approximations** and gradient-free optimization to further reduce the memory wall for large-scale, high-resolution discovery tasks.
- **Cross-Domain Generalization:** Applying InfoSculpt to open-world object detection or semantic segmentation remains an open challenge, where the complexity of spatial signals may require a multi-scale formulation of the CMI surrogate.

F Computational Cost Analysis

We isolate the overhead introduced by InfoSculpt itself with component-wise CUDA profiling. The base GCD training path is kept fixed while toggling $\mathcal{L}_{\text{cat}}$, $\mathcal{L}_{\text{inst}}$, and Nash/KKT balancing. Profiling uses CIFAR-10, batch size 128, two stochastic views, a ViT-B/16 DINO backbone, a fixed batch preloaded on GPU, CUDA event timing, three single-GPU sequential runs, 50 warmup steps, and 500 profiled steps per run; step time is reported as mean±std.

Table 7: Component-wise training-time profiling on CIFAR-10 (relative to the base training path). Peak GPU memory is 553.8 ± 0.0 MB for both the base path and the full refinement path.

Variant	sec/batch	Time ratio vs. base
Base	0.0224±0.0002	1.000×
Base + $\mathcal{L}_{\text{cat}}$	0.0324±0.0005	1.446±0.034×
Base + $\mathcal{L}_{\text{inst}}$	0.0244±0.0013	1.089±0.066×
Base + $\mathcal{L}_{\text{cat}}$ + $\mathcal{L}_{\text{inst}}$	0.0335±0.0016	1.493±0.057×
Base + full (Nash/KKT)	0.0428±0.0005	1.911±0.022×

As shown in Table 7, the two CMI objectives add modest step-level overhead, and the full refinement path uses 1.911× CUDA step time relative to the base training path. Because refinement runs for only 50 epochs on top of a converged base model, this step-level overhead translates into a bounded end-to-end cost: with the refinement occupying the final 50 epochs of the 250-epoch schedule (200 base epochs followed by 50 refinement epochs), a full run amounts to $200 + 50 \times 1.911 \approx 296$ baseline-equivalent epochs, i.e., ≈1.18× the compute of a plain 250-epoch base run. For comparison, training InfoSculpt from scratch would cost 1.911× end-to-end, so post-training refinement is also the computationally economical integration mode. The KKT fixed-point solver itself is lightweight (≈1.18 ms per batch for three objectives and 20 iterations, i.e., ≈5.3% of the base step time); the remaining overhead mainly comes from the additional task-gradient computation required to estimate the Nash weights. Peak GPU memory is unchanged. After refinement, InfoSculpt keeps the same backbone, classification head, and inference path as the base method, resulting in zero extra inference-time overhead.

G Full Results of the New-Oriented Refinement Preference ($q^\star$)

We report the complete five matched-seed results (All/Old/New, %, mean±std) for the new-oriented preference $q^\star$=1.75 described in Sec. 4.2 of the main paper. All RLCD/AptGCD baselines are reproduced in our environment with matched checkpoints, splits, seeds, and hardware, so every comparison is paired; these reproduced baselines may differ from the values originally reported in their papers, which are used only in the main-paper literature comparison. $q^\star$ is selected on RLCD/CIFAR-100 only via the *Old*-validation non-inferiority rule (marked †, excluded from all confirmatory counts) and frozen elsewhere. Relative to matched bases, $q^\star$ improves *New* on all 14 pairs by +1.3 to +13.2 pp (+6.07 pp on average); on the 13 held-out pairs, *New* superiority is significant in 13/13 (Holm-adjusted $p \leq .008$) and *Old* non-inferiority under the pre-specified 0.5-pp margin holds in 11/13.

Table 8: Five-seed All/Old/New accuracy (%, mean±std) under the new-oriented preference: (top) generic benchmarks; (bottom) fine-grained benchmarks. †: RLCD/CIFAR-100 is used only for Old-only calibration and excluded from the 13 held-out confirmatory tests.

Models	CIFAR-10			CIFAR-100			ImageNet-100		
	All	Old	New	All	Old	New	All	Old	New
RLCD	95.1±0.1	94.1±0.2	95.6±0.2	81.3±0.1†	82.1±0.2†	79.8±0.2†	84.7±0.2	92.1±0.2	81.0±0.2
RLCD + InfoSculpt ($q^\star$)	98.0±0.1	94.7±0.3	**99.6±0.2**	84.0±0.4†	83.8±0.3†	**84.4±0.7**†	88.9±1.1	93.1±0.9	**86.8±1.4**
AptGCD	96.2±0.5	94.4±0.5	97.1±0.6	81.3±0.1	80.0±0.2	84.1±0.3	86.0±0.2	93.7±0.3	82.2±0.3
AptGCD + InfoSculpt ($q^\star$)	97.1±0.1	94.4±0.2	**98.5±0.2**	82.1±0.2	80.5±0.1	**85.4±0.5**	91.9±1.3	93.8±1.0	**90.9±1.5**

Models	CUB			Cars			Aircraft			Herbarium 19		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New
RLCD	67.8±0.1	76.8±0.3	63.3±0.3	62.8±0.2	77.1±0.2	55.9±0.2	58.5±0.1	60.0±0.3	57.8±0.1	44.2±0.1	59.1±0.3	36.1±0.2
RLCD + InfoSculpt ($q^\star$)	71.4±0.5	78.1±0.5	**68.0±0.5**	67.1±0.5	78.0±0.5	**61.8±0.6**	61.9±0.8	62.3±0.7	**61.7±1.0**	49.5±0.6	61.5±0.5	**43.1±0.7**
AptGCD	69.1±0.4	72.6±0.4	67.3±0.4	60.4±0.3	78.1±0.4	51.8±0.4	59.3±0.3	63.6±0.3	57.1±0.6	47.1±0.3	58.9±0.4	40.7±0.4
AptGCD + InfoSculpt ($q^\star$)	73.3±0.4	73.9±0.5	**73.0±0.4**	69.3±0.5	78.3±0.6	**65.0±0.5**	64.7±0.8	63.8±0.7	**65.2±0.8**	54.4±0.5	60.1±0.5	**51.3±0.6**