ANCHORED RESIDUAL INTEGRATION WITH STRUCTURE-REGULARIZED ADAPTATION FOR GENERALIZED CATEGORY DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Generalized category discovery recognizes known and novel categories in mixed unlabeled data using annotations from known classes. In fine-grained settings, shared semantic structure can make category assignments ambiguous, while distinguishing cues appear as small deviations in the representation. We identify a geometric effect in cosine classification: relative to a known-semantic anchor, residual directional similarity is weighted by the product of the sample and prototype residual magnitudes. This can weaken the contribution of residual evidence to category decisions. We propose ARISE, a two-stage framework combining structure-regularized adaptation with selective anchored residual integration. The first stage adapts pretrained features under known-class supervision while regularizing local relations with a momentum reference encoder. The second constructs candidate-specific anchors for novel prototypes and uses square-root reweighting to moderate residual-magnitude attenuation. A novelty-and-ambiguity gate and candidate-weighted centering regulate the correction, which is designed to support ambiguous assignments in joint known–novel prediction. Across six benchmarks, ARISE achieves an average overall accuracy of 71.9% and the highest overall accuracy among the compared methods on all four fine-grained benchmarks.

1 INTRODUCTION

Conventional visual recognition typically assumes a closed-set setting in which training and test data share the same predefined label space (Krizhevsky et al., 2012; He et al., 2016). Real-world unlabeled data, however, often contain both previously seen categories and categories absent from labeled training data. Generalized Category Discovery (GCD) addresses this setting by leveraging annotations from known categories to recognize known classes while discovering novel ones in a mixed unlabeled set (Vaze et al., 2022). Although supervised learning provides useful semantic structure, shared semantics can make category assignments ambiguous in fine-grained settings, either between known and novel candidates or among closely related novel categories. Distinguishing evidence may remain encoded as small deviations from shared semantic structure, yet contribute weakly to the final score.

Recent GCD methods improve category discovery through parametric classification, self-distillation, distribution regularization, and fine-grained representation learning (Wen et al., 2023; Rastegar et al., 2025). We revisit this problem through the geometry of cosine classification. Relative to a semantic reference induced by known categories, the cosine similarity between a sample and a novel-class prototype decomposes into an anchor-aligned component and an orthogonal residual component. The former captures shared semantics, whereas the latter captures candidate-specific deviation. Crucially, the residual contribution is scaled by the residual magnitudes of both the sample and the prototype. Thus, even when residual directions contain useful discriminative information, their influence can remain weak when both representations are strongly aligned with the shared semantic reference. We refer to this effect as *residual-magnitude attenuation*. This geometry, illustrated in Fig. 1, suggests that difficult category decisions may arise not because discriminative evidence is absent, but because cosine scoring under-expresses evidence already encoded in the representation.

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

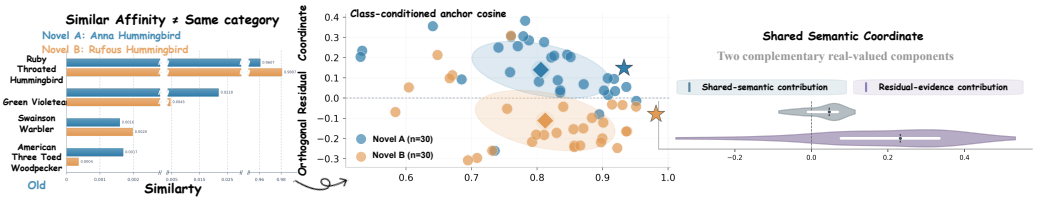


Figure 1: Motivation for ARISE. Similar affinities to known categories can leave the assignments of closely related novel categories ambiguous. Orthogonal residual geometry provides complementary candidate evidence, motivating selective residual reweighting. The illustrated case concerns novel–novel ambiguity.

Based on this analysis, we introduce **ARISE**, a two-stage framework coupling structure-regularized adaptation with anchored residual integration. In the first stage, we adapt a DINO-initialized encoder (Caron et al., 2021) using labeled known-category data. In addition to known-class supervision, ARISE regularizes local relational changes using class-balanced relation batches and an exponential-moving-average reference encoder.

In the discovery stage, ARISE explicitly strengthens residual evidence in ambiguous decisions. Each novel prototype is associated with a semantic anchor obtained by softly aggregating known prototypes. We decompose sample–prototype similarity into anchor-aligned and orthogonal residual components, and apply square-root reweighting to alleviate residual-magnitude attenuation. A novelty-and-ambiguity gate determines where correction is applied, while candidate-weighted centering regulates shared shifts in novel scores. Although only novel candidates receive explicit corrections, the corrected scores participate in joint competition between known and novel categories during learning and inference.

Our contributions are threefold:

- We analyze how cosine classification weights residual directional agreement relative to known-semantic anchors, showing that the product of residual magnitudes can suppress candidate-specific evidence.
- We propose ARISE, combining structure-regularized adaptation with anchored residual reweighting. Candidate-specific anchors, selective gating, and centering regulate novel-candidate corrections during learning and inference.
- ARISE achieves the highest overall accuracy among the compared methods on four fine-grained benchmarks and the best average overall accuracy across six benchmarks.

2 Related Work

2.1 Representation Learning and Adaptation

Learning transferable visual representations is central to modern recognition systems. Early work showed that deep features gradually shift from generic to task-specific representations, and that transferability tends to decrease as higher layers become increasingly specialized to the source task (Yosinski et al., 2014). Subsequent studies further examined how pretrained representations transfer across downstream recognition problems (Kornblith et al., 2019). More recently, self-supervised approaches such as MoCo and DINO have demonstrated that semantically meaningful and transferable visual structures can emerge without category-level supervision (He et al., 2020; Caron et al., 2021), making pretrained encoders a strong basis for downstream category discovery.

Adapting such representations to a target task, however, introduces a trade-off between task specialization and retention of pretrained structure. L2-SP addresses this issue by regularizing model parameters toward their pretrained initialization (LI et al., 2018), while DELTA instead constrains feature responses inherited from the pretrained network (Li et al., 2019). More recent analyses show that full fine-tuning can substantially distort high-quality pretrained features and weaken their generalization under distribution shift (Kumar et al., 2022). Momentum-based reference models, as used in Mean Teacher and MoCo, additionally provide slowly evolving targets that can stabilize

representation learning (Tarvainen & Valpola, 2017; He et al., 2020). ARISE focuses on preserving transferable relational structure during known-category adaptation. We regularize local neighborhood relations using class-balanced relation batches.

2.2 Generalized Category Discovery

Generalized category discovery (GCD) (Vaze et al., 2022) relaxes the assumption that labeled and unlabeled data have disjoint class sets, allowing unlabeled data to contain both known and novel classes. Subsequent work refines learning objectives and clustering strategies to learn more discriminative representations and uncover latent category structure in unlabeled data (Pu et al., 2023; Choi et al., 2024; Rastegar et al., 2025).

Parametric methods use a learnable classifier to jointly predict known and novel classes. Following UNO's (Fini et al., 2021) unified classification approach, SimGCD (Wen et al., 2023) employs self-distillation and entropy regularization to mitigate noisy pseudo-labels and imbalanced predictions. Subsequent methods focus on improving supervision and reducing bias toward known classes. DebGCD (Liu & Han, 2025) introduces an auxiliary debiasing classifier and semantic distribution guidance, while RLCD (Liu et al., 2025a) combines mutual learning across branches with class-level distribution regularization. RPC (Xu et al., 2026) enforces consistency in sample-to-known-prototype relations to support knowledge transfer and novel class discovery.

A complementary line of research improves GCD through semantic information (Bian et al., 2026), representation structure, and optimization. For semantic and local representation enhancement, ATCG (Han et al., 2026) generates textual concepts for unlabeled samples by drawing analogies from knowledge in labeled data, then fuses these concepts with visual features. CoGe-GCD (Tang et al., 2026) combines compositional perception with geometric calibration. To improve feature-space geometry, HypCD (Liu et al., 2025b) learns hierarchy-aware representations and classifiers in hyperbolic space. At the optimization level, EAGC (Zheng et al., 2026) addresses gradient entanglement through anchor-based gradient alignment and energy-aware elastic projection. This reduces interference with novel class learning while preserving discrimination among known classes. ARISE couples this score-level correction with structure-regularized adaptation.

3 Method

3.1 Problem Setup and Overview

Generalized category discovery uses a labeled dataset $\mathcal{D}_l = \{(x_i, y_i)\}$ from known classes $\mathcal{Y}_o$ and an unlabeled dataset $\mathcal{D}_u = \{x_i\}$ containing both known and novel classes $\mathcal{Y}_n$, with $\mathcal{Y}_o \cap \mathcal{Y}_n = \varnothing$. We assume known class counts $K_o = |\mathcal{Y}_o|$ and $K_n = |\mathcal{Y}_n|$, and denote their sum by K.

The encoder f_θ produces unit features $h_i = \mathrm{Normalize}(f_\theta(x_i))$. Each class has a trainable prototype w_k, normalized as $c_k = \mathrm{Normalize}(w_k)$. The base classifier computes cosine scores $u_{ik} = h_i^\top c_k$ and logits $z_{ik} = u_{ik}/\tau$, where τ is the temperature.

ARISE combines two sequential stages (Fig. 2). *Geometry-regularized known-class adaptation* learns from known labels while constraining changes in local feature relations. *anchored residual integration* decomposes cosine scores relative to candidate-specific known-semantic anchors and selectively lifts residual contributions attenuated by their magnitudes. Together, the two stages address representation adaptation and the selective use of residual evidence in ambiguous category decisions.

3.2 Geometry-Regularized Known-Class Adaptation

We initialize a student encoder f_θ and an EMA reference encoder f_ϕ from the same DINO-pretrained weights. At each optimization step, we sample two labeled known-class batches: a standard classification batch $\mathcal{B}_c$ and a class-balanced relation batch $\mathcal{B}_g$. For $\mathcal{B}_g$, we sample P classes and M instances per class; we use $P = 16$ and $M = 4$ by default.

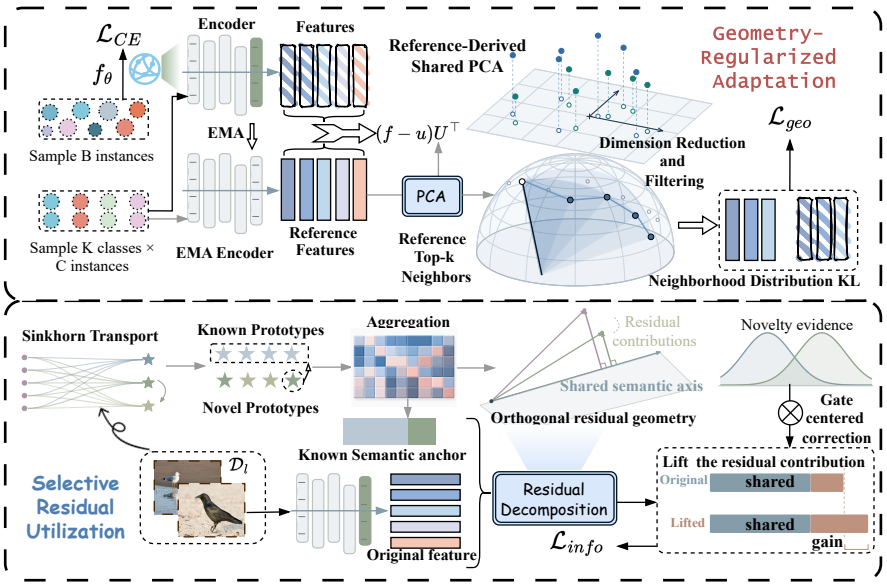


Figure 2: Overview of ARISE. Top: geometry-regularized adaptation combines known-class supervision with local relational distillation in shared, reference-derived PCA coordinates. Bottom: anchored residual integration constructs candidate-specific semantic anchors for novel prototypes and selectively reweights orthogonal residual contributions for joint category prediction.

Let v_k denote the classifier parameter for known class $k \in \mathcal{Y}_o$, and define $\bar{v}_k = \text{Normalize}(v_k)$. For $h_i = \text{Normalize}(f_\theta(x_i))$, the classification objective is

$$\mathcal{L}_{\text{CE,old}} = -\frac{1}{|\mathcal{B}_c|} \sum_{i \in \mathcal{B}_c} \log \frac{\exp(h_i^\top \bar{v}_{y_i}/\tau)}{\sum_{k \in \mathcal{Y}_o} \exp(h_i^\top \bar{v}_k/\tau)}. \tag{1}$$

For the relation batch, we extract normalized student and reference features,

$$H_i = \text{Normalize}(f_\theta(x_i)), \qquad R_i = \text{Normalize}(f_\phi(x_i)). \tag{2}$$

Let μ_ϕ be the mean reference feature in the relation batch. We form the centered reference feature matrix and take its leading d_g right singular vectors as the columns of U_ϕ. Both μ_ϕ and U_ϕ are treated as constants during student optimization:

$$\widetilde{H}_i = \text{Normalize}((H_i - \mu_\phi)U_\phi)\,, \qquad \widetilde{R}_i = \text{Normalize}((R_i - \mu_\phi)U_\phi)\,, \tag{3}$$

where $d_g = \min(128, |\mathcal{B}_g| - 1, D)$ is the projection dimension.

For each sample, the reference representation selects its k nearest neighbors,

$$\mathcal{N}_i = \text{TopK}_{j \neq i}\, \widetilde{R}_i^\top \widetilde{R}_j. \tag{4}$$

Over the same neighbor set, student and reference relation distributions are defined as

$$p_{ij}^\psi = \frac{\exp(\widetilde{X}_i^{\psi\top} \widetilde{X}_j^\psi/\tau_g)}{\sum_{r \in \mathcal{N}_i} \exp(\widetilde{X}_i^{\psi\top} \widetilde{X}_r^\psi/\tau_g)}, \qquad \psi \in \{\theta, \phi\}, \tag{5}$$

where $\widetilde{X}^\theta = \widetilde{H}$, $\widetilde{X}^\phi = \widetilde{R}$, and $\tau_g = 0.2$.

We align the student relation distribution with the detached reference:

$$\mathcal{L}_{\text{geo}} = \frac{1}{|\mathcal{B}_g|} \sum_i \text{KL}\left(\text{sg}(p_i^\phi) \,\|\, p_i^\theta\right). \tag{6}$$

The adaptation objective is

$$\mathcal{L}_{\text{adapt}} = \mathcal{L}_{\text{CE,old}} + \lambda_{geo}\mathcal{L}_{\text{geo}}. \tag{7}$$

After each student update, the reference encoder is updated by

$$\phi \leftarrow m\phi + (1-m)\theta. \tag{8}$$

After adaptation, only the student encoder is retained for category discovery. The known-class classifier is discarded, and a new prototype library is constructed from the adapted features. Known prototypes are estimated from labeled samples, while novel prototypes are initialized by clustering and refined using Sinkhorn assignments. The complete initialization procedure is given in Appendix.

3.3 Known-Semantic Anchors

We evaluate each novel candidate relative to a semantic reference induced by the known categories. This reference separates an anchor-aligned contribution from orthogonal residual evidence that can help assess candidate compatibility.

A novel category can be related to multiple known categories rather than to a single nearest prototype. We therefore construct a soft, candidate-specific anchor for each novel prototype c_j:

$$a_j = \mathrm{sg}\left[\mathrm{Normalize}\left(\sum_{k\in\mathcal{Y}_o} \alpha_{jk} c_k\right)\right], \qquad \alpha_{jk} \propto \exp\left(\frac{\mathrm{sg}(c_j)^\top \mathrm{sg}(c_k)}{\tau}\right). \tag{9}$$

The anchor summarizes the known-category semantics most compatible with novel candidate j. Importantly, it is candidate specific. The same sample can therefore have different residual components when evaluated against different novel prototypes. This is distinct from projecting the representation onto a single global "known subspace."

The stop-gradient operation prevents the residual branch from directly modifying the anchor through the current backward pass. However, the anchor is recomputed as prototypes evolve, so it is not permanently fixed throughout training.

3.4 Residual Decomposition and Lifting

For sample h_i and novel prototype c_j, let η_{ij} and β_j denote their projections onto a_j. Let t_{ij} and s_j denote the magnitudes of their corresponding orthogonal residuals, and let q_{ij} denote the cosine similarity between the two residual directions. The cosine score admits the exact decomposition

$$u_{ij} = \eta_{ij}\beta_j + t_{ij}s_j q_{ij}. \tag{10}$$

The first term measures agreement along the known-semantic anchor, whereas the second measures agreement between orthogonal residuals. The residual direction can provide complementary evidence for a candidate, but its contribution is multiplied by both residual magnitudes. When these magnitudes are small, such evidence may have limited influence on an ambiguous category decision.

When the sample and prototype are both strongly aligned with the shared semantic reference, these residual magnitudes can be small. As a result, a potentially informative residual direction may have only a weak effect on the overall score. We refer to this score-level effect as *residual-magnitude attenuation*.

A direct normalization by residual magnitude would remove this attenuation, but would also discard information about residual strength. We instead use a square-root reweighting:

$$\widetilde{r}_{ij} = \sqrt{t_{ij}s_j}\, q_{ij}, \qquad d_{ij} = \widetilde{r}_{ij} - t_{ij}s_j q_{ij}. \tag{11}$$

ARISE adds only the difference d_{ij} rather than adding the lifted residual independently. The original cosine score therefore remains the baseline. The square-root form reduces the multiplicative attenuation while remaining magnitude aware. In particular, the lifted contribution still approaches zero as the residual magnitude vanishes.

We do not assume that every weak residual is useful. Whether the residual direction carries reliable novel-category evidence depends on the sample and candidate. This motivates the selective correction described next.

3.5 SELECTIVE RESIDUAL CORRECTION

Applying residual lifting to every sample and every novel candidate can also amplify unreliable directions. ARISE therefore controls the correction in two complementary ways: candidate-weighted centering and a novelty-and-ambiguity gate.

Let p_i^n denote the original softmax distribution over the novel candidates. We center the lifting difference using

$$\delta_{ij} = d_{ij} - \sum_{r \in \mathcal{Y}_n} \mathrm{sg}(p_{ir}^n) d_{ir}. \tag{12}$$

This removes the candidate-weighted common shift in the correction. Because the same centering term is subtracted from every novel candidate, it does not by itself change pairwise differences among novel candidates. Its primary role is therefore to control the collective shift of novel scores relative to the known group.

Residual correction is most relevant when a sample is likely to belong to the novel group but remains difficult to assign within that group. We therefore estimate two quantities.

First, we compare class-count-normalized log-sum-exp scores over the known and novel groups. Their difference defines a detached novelty tendency ν_i. Second, we measure uncertainty among novel candidates using the normalized entropy of p_i^n, yielding an ambiguity score e_i. The final gate is

$$g_i = \underbrace{\mathrm{sg}\left[\sigma(A_n(h_i) - A_o(h_i))\right]}_{\nu_i} \underbrace{\mathrm{sg}\left[\frac{H(p_i^n)}{\log K_n}\right]}_{e_i}. \tag{13}$$

This gate depends on the model's predictions rather than the true known or novel identity of a sample. It can therefore also affect known-class samples for which novel candidates remain competitive. The entropy term measures uncertainty within the novel group; it is not an entropy measure over all categories.

Combining lifting, centering, and gating gives the corrected logits:

$$\widehat{z}_{ik} = \begin{cases} u_{ik}/\tau, & k \in \mathcal{Y}_o, \\ \left(u_{ik} + \lambda_c g_i \delta_{ik}\right)/\tau, & k \in \mathcal{Y}_n. \end{cases} \tag{14}$$

Only the novel scores receive the explicit residual correction. However, changing novel scores can still alter the final competition between known and novel classes.

3.6 DISCOVERY OBJECTIVES AND INFERENCE

Each labeled or unlabeled image is augmented into two views. The base discovery objective contains three standard terms. First, $\mathcal{L}_{\mathrm{sup}}$ is the mean cross-entropy of the base logits z over both views of labeled samples. Second, $\mathcal{L}_{\mathrm{dist}}$ is symmetric cross-view distillation on unlabeled samples: the sharpened, stop-gradient prediction of either view supervises the base prediction of the other. Third, $\mathcal{L}_{\mathrm{bal}} = \log K - H(\bar{p})$ prevents marginal collapse over the mixed batch. Here $\bar{p}$ is the mean prediction obtained from a balance-only copy of z in which each known logit is multiplied by

$$\xi_i = 1 + \lambda_o \,\mathrm{sg}\left[P_{o,i}\left(1 - \frac{m_i + 1}{2}\right)\right], \quad P_{o,i} = \sum_{k \in \mathcal{Y}_o} \mathrm{softmax}(z_i)_k, \quad m_i = \max_{k \in \mathcal{Y}_o} u_{ik}. \tag{15}$$

This rescaling is confined to $\mathcal{L}_{\mathrm{bal}}$.

The corrected logits $\widehat{z}$ are used for the novel-category information objective. For two augmented views, we average their corrected novel-class probability distributions to obtain ρ_i. Samples are weighted by their estimated novelty tendency, $w_i = \frac{1}{2}(\nu_i^{(1)} + \nu_i^{(2)})$. The resulting objective is

$$\mathcal{L}_{\mathrm{info}} = \frac{\sum_i w_i H(\rho_i)}{\sum_i w_i} - H\left(\frac{\sum_i w_i \rho_i}{\sum_i w_i}\right). \tag{16}$$

The first term favors confident predictions for individual samples, while the second encourages diversity across the aggregate novel distribution. Importantly, the weighting is based on the model-estimated novelty tendency rather than ground-truth novel membership. The complete discovery objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{sup}} + \mathcal{L}_{\mathrm{dist}} + \lambda_{\mathrm{bal}}\mathcal{L}_{\mathrm{bal}} + \lambda_{\mathrm{info}}\mathcal{L}_{\mathrm{info}}. \tag{17}$$

Table 1: Performance comparison across six benchmark datasets. All, Old, and New denote accuracy (%) on all, known, and novel classes, respectively. Average reports the mean accuracy across the six datasets. The best results are **bolded**, and the second-best results are underlined.

Methods	CIFAR-100			ImageNet-100			CUB-200			Stanford Cars			Herbarium19			FGVC-Aircraft			Average		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
GCD	70.8	77.6	57.0	74.1	89.8	66.3	51.3	56.6	48.7	39.0	57.6	29.9	35.4	51.0	27.0	45.0	41.1	46.9	52.6	62.3	46.0
DCCL	75.3	76.8	70.2	80.5	90.5	76.2	63.5	60.8	64.9	43.1	55.7	36.2	–	–	–	–	–	–	–	–	–
PrCAL	81.2	84.2	75.3	83.1	92.7	78.3	62.9	64.4	62.1	50.2	70.1	40.6	37.0	52.0	28.9	52.2	52.2	52.3	61.1	69.3	56.3
SimGCD	80.1	81.2	77.8	83.0	93.1	77.9	60.3	65.6	57.7	53.8	71.9	45.0	44.0	58.0	36.4	54.2	59.1	51.8	62.6	71.5	57.8
SPTNet	81.3	84.3	75.6	85.4	93.2	81.4	65.8	68.8	65.1	59.0	79.2	49.3	43.4	58.7	35.2	59.3	61.8	58.1	65.7	74.3	60.8
CMS	82.3	**85.7**	75.5	84.7	**95.6**	79.2	68.2	**76.5**	64.0	56.9	76.1	47.6	36.4	54.9	26.4	56.0	63.4	52.3	64.1	75.4	57.5
LegoGCD	81.8	81.4	82.5	86.3	94.5	82.1	63.8	71.9	59.8	57.3	75.7	48.4	45.1	57.4	38.4	55.4	68.3	48.8	65.0	74.9	60.0
ProtoGCD	81.9	82.9	80.0	84.0	92.2	79.9	63.2	68.5	60.5	53.8	73.7	44.2	44.5	59.4	36.5	56.8	62.5	53.9	64.0	73.2	59.2
SelEx	82.3	85.3	76.3	83.1	93.6	77.8	73.6	75.3	72.8	58.5	75.6	50.3	39.6	54.9	31.3	57.1	64.7	53.3	65.7	74.9	60.3
DebGCD	83.0	84.6	79.9	85.9	94.3	81.6	66.3	71.8	63.5	65.3	81.6	57.4	44.7	59.4	36.8	61.7	63.9	60.6	67.8	75.9	63.3
CURE	**85.7**	85.3	**86.6**	**87.0**	95.3	**82.9**	75.2	75.5	75.1	68.7	80.8	63.0	48.1	53.1	**45.4**	61.6	**69.4**	57.8	71.1	76.6	68.5
Ours	85.6	85.5	85.9	86.4	94.6	82.2	**76.4**	75.7	**76.8**	**70.1**	**82.8**	**64.0**	**49.6**	**61.2**	43.5	**63.1**	66.8	**61.2**	**71.9**	**77.8**	**68.9**

Table 2: Cumulative component analysis.

ID	G	S	B	R	CUB-200			Stanford Cars		
					All	Old	New	All	Old	New
(a)	No Warmup				57.3	65.0	53.5	40.3	54.3	33.6
(b)	×	×	×	×	69.6	75.6	66.6	65.3	76.8	59.7
(c)	✓	×	×	×	70.8	78.5	67.0	65.0	76.5	59.4
(d)	✓	✓	×	×	74.7	81.5	71.4	66.5	79.7	60.1
(e)	✓	✓	✓	×	75.3	77.1	74.4	68.8	82.2	62.3
(f)	✓	✓	✓	✓	76.4	75.7	76.8	70.1	82.8	64.0

Table 3: Residual correction.

Variant	CUB-200			Stanford Cars		
	All	Old	New	All	Old	New
Common inference	74.7	74.9	74.6	66.2	81.4	58.9
Correction only	75.4	77.5	74.4	66.3	82.5	58.4
Real train/infer	75.4	76.6	74.7	66.3	82.5	58.4
Uncentered	75.0	73.5	75.7	66.4	82.4	58.6
Unweighted InfoMax	74.3	74.4	74.3	65.6	80.3	58.6
Full	76.4	75.7	76.8	70.1	82.8	64.0

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and evaluation protocol. We evaluate ARISE on six benchmarks (CIFAR-100 (Krizhevsky et al., 2009) , ImageNet-100 (Russakovsky et al., 2015), CUB-200 (Wah et al., 2011), Stanford Cars (Krause et al., 2013), Herbarium19 (Tan et al., 2019), and FGVC-Aircraft (Maji et al., 2013)). We follow the standard GCD splits and labeling protocol. Following prior work, we assume that the total number of classes is known. We report clustering accuracy on the unlabeled set using Hungarian matching.

Implementation details. We initialize the encoder with DINO-pretrained weights and train ARISE in two stages. The adaptation stage uses known-class supervision and relational regularization, with class-balanced relation batches containing 16 classes and 4 images per class. We use 8 reference neighbors, a relation temperature of 0.2, an EMA momentum of 0.9. The discovery stage initializes known prototypes from labeled features and novel prototypes through clustering and Sinkhorn refinement. Corrected logits are used for the novel-category information objective and joint inference. we adopt the Vision Transformer (ViT-B/16) (Dosovitskiy et al., 2021) as backbone and optimize the last block. The adaptation and discovery stages run for 200 epochs with learning rates 0.1 and batch sizes 128, respectively. Additional implementation details are provided in the appendix.

4.2 COMPARISON WITH STATE-OF-THE-ART METHODS

Table 1 compares ARISE with representative GCD methods (DCCL (Pu et al., 2023), PrCAL (Zhang et al., 2023), SPTNet (Wang et al., 2024), CMS (Choi et al., 2024), LegoGCD (Cao et al., 2024), ProtoGCD (Ma et al., 2025), Selex (Rastegar et al., 2025), DebGCD (Liu & Han, 2025) and CURE (Bian et al., 2026)) across six benchmarks. ARISE achieves the highest average accuracy among the compared methods on all three metrics. Compared with CURE, the strongest baseline by average All accuracy, these results represent improvements of 0.8, 1.2, and 0.4 percentage points, respectively, based on the reported averages.

Fine-grained category discovery. ARISE achieves the best All accuracy on all four fine-grained benchmarks. It obtains 76.4% on CUB-200, 70.1% on Stanford Cars, 49.6% on Herbarium19, and

Table 4: Warmup and geometric reference.

Variant	CUB-200			Stanford Cars		
	All	Old	New	All	Old	New
No warmup	57.3	65.0	53.5	40.3	54.3	33.6
CE-only warmup	73.0	75.3	71.8	67.8	83.2	60.3
Fixed reference	64.6	71.9	61.0	57.2	76.0	48.1
Full	76.4	75.7	76.8	70.1	82.8	64.0

Table 5: Pseudo-label refinement and balancing.

Variant	CUB-200			Stanford Cars		
	All	Old	New	All	Old	New
No Sinkhorn	71.4	74.4	69.9	66.4	79.1	60.3
Uniform prior	73.1	75.2	72.0	67.1	81.2	60.3
No class scaling	74.9	78.9	72.9	67.5	82.6	60.3
Full	76.4	75.7	76.8	70.1	82.8	64.0

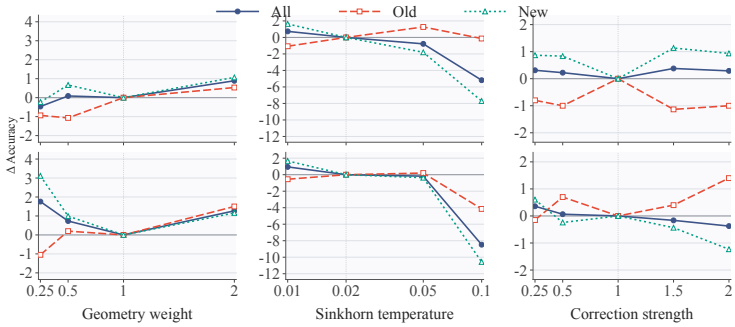


Figure 3: Sensitivity to the hyperparameter. top: CUB-200, bottom: Standford Cars.

63.1% on FGVC-Aircraft, exceeding the strongest baseline on each dataset by 1.2%, 1.4%, 1.5%, and 1.4%, respectively. ARISE also achieves the best New accuracy on CUB-200, Stanford Cars, and FGVC-Aircraft, improving over the strongest competing results by 1.7, 1.0, and 0.6 points. These results establish strong overall performance on fine-grained benchmarks, where shared visual structure makes category assignments challenging. The gains are not uniform across known and novel subsets. On Herbarium19, improved known-class accuracy accompanies a reduction in novel-class accuracy, whereas FGVC-Aircraft shows the opposite trade-off relative to CURE. These differences motivate examining how the corrected scores affect joint category decisions.

The improvements differ across known and novel classes. In Herbarium19, ARISE improves the accuracy over CURE from 53.1% to 61.2% On Old, but reduces the accuracy of New from 45.4% to 43.5%. On FGVC-Aircraft, it increases New accuracy over CURE by 3.4 points while decreasing Old accuracy by 2.6%.

Generic category discovery. ARISE remains competitive on CIFAR-100 and ImageNet-100, achieving the second-best All and New accuracy on both datasets. Its All accuracy reaches 85.6% and 86.4%, respectively, trailing CURE by 0.1 and 0.6 points. Together with the fine-grained results, these comparisons show that ARISE's clearest advantage lies in fine-grained overall recognition, while maintaining strong performance on generic benchmarks.

4.3 Ablation Study

We examine geometry-guided adaptation (G), Sinkhorn refinement (S), old-class MeMax scaling (B), and residual learning and correction (R) on CUB-200 and Stanford Cars.

Cumulative contributions. Combining adaptation, refinement, and balancing improves All accuracy over the internal baseline by 18.0 points on CUB-200 and 28.5 points on Stanford Cars (Table 2). The corresponding New gains of 20.9 and 28.7 points indicate that these components primarily strengthen novel-category discovery. Full-run ARISE achieves 76.4% and 70.1% All accuracy, respectively.

Representation adaptation. Warmup provides the largest initial gains: 12.3 and 25.0 All points ((a)→(b)), highlighting the importance of adapting pretrained features before discovery. Geometric regularization further improves CUB-200 by 1.2 points, but slightly reduces Stanford Cars accuracy by 0.3 points ((b)→(c)). This contrast suggests that preserving local relations is more useful when those relations complement task-specific adaptation. In the full-model design study, CE-only warmup outperforms both no warmup and a fixed geometric reference (Table 4), suggesting that an overly restrictive reference can impede adaptation.

Refinement and balancing. Sinkhorn refinement improves All accuracy by 3.9 and 1.5 points ((c)→(d)), consistent with more coherent assignments supporting discovery. Class scaling subsequently raises New accuracy by 3.0 and 2.2 points ((d)→(e)). On CUB-200, Old accuracy decreases as New accuracy improves, suggesting that scaling redistributes emphasis toward novel classes; on Stanford Cars, both subsets benefit. Together, S and B improve New accuracy over G alone by 7.4 and 2.9 points. A uniform Sinkhorn prior also outperforms removing refinement (Table 5), suggesting that assignment regularization contributes beyond the choice of prior.

Residual correction. Correction-only inference increases CUB-200 All accuracy from 74.7% to 75.4% and Stanford Cars All accuracy from 66.2% to 66.3% (Table 3). These gains accompany improvements of 2.6 and 1.1 points in Old accuracy, with decreases of 0.2 and 0.5 points in New accuracy. Thus, the observed inference gains concern joint category decisions, rather than a uniform improvement in novel-class recognition. On CUB-200, uncentered correction achieves higher New accuracy but lower Old accuracy than correction-only inference, illustrating a different known–novel trade-off. These results motivate evaluating residual correction through both candidate discrimination and competition between category groups.

4.4 Hyperparameter Analysis

We analyze the sensitivity of ARISE to three key hyperparameters: the geometric regularization weight, Sinkhorn temperature, and residual-correction strength. Figure 3 reports the change in accuracy relative to the default configuration for All, Old, and New classes.

Geometric regularization weight λ_{geo}. ARISE is relatively insensitive to the geometric regularization weight. On CUB-200, varying the weight around the default value leads to only minor changes in overall accuracy, while a larger weight can provide a small improvement. On Stanford Cars, reducing the weight to 0.25 yields an approximately $+1.8$ point improvement in All accuracy, and increasing it to 2.0 also remains beneficial. These results indicate that the geometric constraint is effective over a reasonably broad range rather than relying on a narrowly tuned coefficient.

Sinkhorn temperature τ_g. The Sinkhorn temperature exhibits the clearest sensitivity among the three hyperparameters. A low temperature consistently performs best, whereas increasing it to 0.1 causes substantial degradation, especially for novel classes. The drop reaches approximately 5 points in All accuracy on CUB-200 and more than 8 points on Stanford Cars, accompanied by even larger reductions in New accuracy. This behavior suggests that sharp assignment distributions are important for preserving discriminative pseudo-labels during Sinkhorn refinement, while overly smooth assignments obscure fine-grained category structure.

Residual-correction strength λ_c. In contrast, overall performance is stable across a wide range of correction strengths. Varying the coefficient from 0.25 to 2.0 changes All accuracy only marginally on both datasets. Interestingly, Old and New accuracy respond differently to the correction strength: stronger correction can favor one subset while slightly reducing the other, whereas their combined All accuracy remains largely unchanged. This indicates that the correction module is robust to its scaling factor, while the coefficient primarily controls how correction is distributed between known and novel categories rather than determining overall performance.

5 Conclusion

We presented ARISE, a two-stage framework for generalized category discovery that combines structure-regularized adaptation with anchored residual integration. Our analysis identifies residual-magnitude attenuation in cosine classification: directional agreement between orthogonal residuals is weighted by the product of their magnitudes, potentially weakening useful evidence for distinguishing similar novel categories. ARISE addresses this effect through candidate-specific known-semantic anchors and square-root residual reweighting, with selective gating and centering to regulate the correction. Experiments on six benchmarks demonstrate competitive performance, including the highest overall accuracy among the compared methods on all four fine-grained datasets. These results motivate jointly considering representation adaptation and how classification scores incorporate residual evidence.

AI USE STATEMENT

In this work, we used generative AI tools to assist with the refinement of conceptual and methodological descriptions, provide feedback on experimental design and ablation studies, assist with the interpretation of experimental results, and support translation. We also used generative AI tools for drafting and editing parts of the manuscript, improving readability and presentation, suggesting experimental analyses, summarizing and identifying relevant literature, and assisting with manuscript organization.

All AI-assisted research content was subsequently reviewed and verified by the authors. Mathematical formulations and methodological claims were checked independently by the authors, experimental suggestions were validated through actual experiments, and all reported numerical results were obtained from our experimental pipeline rather than generated by AI. Literature references suggested with AI assistance were manually checked against the original sources. The authors made the final decisions on the research methodology, experimental design, analysis, and presentation.

We did not use generative AI tools to generate synthetic datasets, clean or reformat datasets, or perform qualitative or thematic data analysis. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Yuwei Bian, Shidong Wang, and Haofeng Zhang. CURE: Consistency-under-unified semantic regularization for generalized category discovery. In *Forty-third International Conference on Machine Learning*, 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 *CVPR*, pp. 16880–16889, 2024.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pp. 9650–9660, October 2021.

Sua Choi, Dahyun Kang, and Minsu Cho. Contrastive mean-shift learning for generalized category discovery. In *CVPR*, pp. 23094–23104, June 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

Enrico Fini, Enver Sangineto, Stéphane Lathuilière, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery. In *ICCV*, 2021.

Jizhou Han, Chenhao Ding, Yuhang He, Qiang Wang, Shaokun Wang, SongLin Dong, and Yihong Gong. Learning like humans: Analogical concept learning for generalized category discovery. In *CVPR*, pp. 28734–28743, June 2026.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, June 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, June 2020.

Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better? In *CVPR*, June 2019.

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

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

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, volume 25, 2012.

Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-Tuning Can Distort Pretrained Features and Underperform Out-of-Distribution. In *ICLR*, 2022.

Xingjian Li, Haoyi Xiong, Hanchao Wang, Yuxuan Rao, Liping Liu, and Jun Huan. Delta: Deep learning transfer using feature map with attention for convolutional networks. In *ICLR*, 2019.

Xuhong LI, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. In *ICML*, volume 80, pp. 2825–2834, 10–15 Jul 2018.

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

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

Yuanpei Liu, Zhenqi He, and Kai Han. Hyperbolic category discovery. In *CVPR*, pp. 9891–9900, 2025b.

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*, 47(7):6022–6038, 2025.

Subhransu Maji, Juho Kannala, Esa Rahtu, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. Technical report, University of Oxford, 2013.

Nan Pu, Zhun Zhong, and Nicu Sebe. Dynamic conceptional contrastive learning for generalized category discovery. In *CVPR*, pp. 7579–7588, 2023.

Sarah Rastegar, Mohammadreza Salehi, Yuki M. Asano, Hazel Doughty, and Cees G. M. Snoek. Selex: Self-expertise in fine-grained generalized category discovery. In *ECCV*, pp. 440–458, 2025.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015.

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, Jiewei Zheng, Kunze Huang, Chaoqi Chen, Yue Huang, and Cheng Chen. Coge-gcd: Reframing generalized category discovery with compositional generalization. In *ICML*, 2026.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, volume 30, 2017.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *CVPR*, pp. 7492–7501, June 2022.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Hongjun Wang, Sagar Vaze, and Kai Han. Sptnet: An efficient alternative framework for generalized category discovery with spatial prompt tuning. In *ICLR*, volume 2024, pp. 53803–53838, 2024.

Xin Wen, Bingchen Zhao, and Xiaojuan Qi. Parametric classification for generalized category discovery: A baseline study. In *ICCV*, pp. 16590–16600, 2023.

Yulin Xu, Chunqi Guo, Yuanzhen Shuai, and Jianyuan Ni. Relational retrieval: Leveraging known-novel interactions for generalized category discovery. In *Proceedings of the 2026 International Conference on Multimedia Retrieval*, pp. 410–414, 2026.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger (eds.), *NeurIPS*, volume 27, 2014.

Sheng Zhang, Salman Khan, Zhiqiang Shen, Muzammal Naseer, Guangyi Chen, and Fahad Shahbaz Khan. Promptcal: Contrastive affinity learning via auxiliary prompts for generalized novel category discovery. In *CVPR*, pp. 3479–3488, June 2023.

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 *CVPR*, 2026.

A Additional Method Details

This appendix complements the method description in the main paper with implementation details, exact definitions, and derivations omitted for space. We refer to the equations in the main text whenever the corresponding definition has already been introduced, and focus here on details that are useful for reproducing ARISE.

A.1 Implementation Details of Geometry-Regularized Adaptation

Relation-batch sampling. As described in the main text, each adaptation step uses a standard classification batch $\mathcal{B}_c$ and a class-balanced relation batch $\mathcal{B}_g$. For $\mathcal{B}_g$, we sample P known categories and M instances from each selected category, with $P = 16$ and $M = 4$ by default. If fewer than P known categories are available, P is reduced accordingly. When a selected category contains fewer than M samples, sampling with replacement is used.

Class labels in $\mathcal{B}_g$ determine only the sampling strategy. They are not used to construct positive or negative pairs for the relational objective. Consequently, the geometry regularizer in Eq. equation 6 is determined entirely by the neighborhood structure of the reference representation.

Reference coordinates and local relations. For every relation-batch sample, the same augmented image tensor is passed to the student encoder f_θ and the EMA reference encoder f_ϕ. The reference features, their batch mean μ_ϕ, and the reference-derived basis U_ϕ in Eq. equation 3 are detached during the student update.

The projection dimension is

$$d_g = \min(128, |\mathcal{B}_g| - 1, D),$$

where D denotes the encoder feature dimension. The local relation distributions in Eq. equation 5 are constructed using the $k = 8$ nearest neighbors selected by the reference representation. We use $\tau_g = 0.2$ for the relation distributions.

The adaptation loss in Eq. equation 7 uses $\lambda_{geo} = 1$. After every student update, the reference parameters are updated with EMA momentum

$$m = 0.9.$$

Although both encoders are initialized from the same DINO-pretrained model, the reference encoder is not a fixed DINO teacher. It evolves with the student through EMA and therefore serves as a slowly changing relational reference rather than as a constraint that exactly preserves the original DINO geometry.

Adaptation classifier. The classifier used during this stage contains only the K_o known categories. It is used solely for the supervised adaptation objective and is discarded after adaptation. Its parameters are not transferred to the prototype library used during category discovery.

A.2 Prototype Initialization

Prototype initialization is performed once after geometry-regularized adaptation and before discovery training.

We first sample up to $200K_o$ examples from the mixed training data and extract normalized features using a deterministic evaluation transformation. Ward hierarchical clustering is then applied to obtain initial cluster centers. Labeled samples are used to align clusters with the known categories. After alignment, each known center is replaced by the normalized mean feature of the labeled samples from the corresponding category.

The remaining centers are treated as novel-category candidates and refined using entropy-regularized Sinkhorn assignments. Let N denote the initialization subset size and let n_k^l denote the number of labeled samples from known category k. To account for labeled samples that have already contributed explicitly to known-category initialization, we define a residual class mass

$$m_k = \begin{cases} \max\big(N/K - n_k^l,\, 0.25\big), & k \in \mathcal{Y}_o, \\ N/K, & k \in \mathcal{Y}_n, \end{cases} \qquad \pi_k = \frac{m_k}{\sum_r m_r}. \tag{18}$$

For the N_u unlabeled samples in the initialization subset, let $S \in \mathbb{R}^{N_u \times K}$ denote their similarity matrix to the current centers. The Sinkhorn assignment approximately solves

$$\max_{Q \geq 0} \langle Q, S \rangle + \varepsilon H(Q), \qquad Q\mathbf{1} = \mathbf{1}, \qquad Q^\top \mathbf{1} = N_u \pi. \tag{19}$$

During this refinement, the known-category centers remain fixed. Only the novel centers are updated:

$$c_j \leftarrow \text{Normalize}\left(\frac{\sum_i Q_{ij} h_i}{\sum_i Q_{ij}}\right), \qquad j \in \mathcal{Y}_n. \tag{20}$$

The resulting known and novel centers initialize the trainable prototype parameters used in discovery training. The Sinkhorn assignments are used only for initialization; they are not maintained as persistent pseudo-labels during subsequent optimization.

A.3 Residual Decomposition: Derivation and Numerical Details

The main text gives the decomposition in Eq. equation 10. Here we provide its explicit derivation.

For normalized sample feature h_i, novel prototype c_j, and its candidate-specific normalized anchor a_j, define the anchor projections

$$\eta_{ij} = h_i^\top a_j, \qquad \beta_j = c_j^\top a_j. \tag{21}$$

The corresponding orthogonal residual components are

$$h_{ij}^\perp = h_i - \eta_{ij} a_j, \qquad c_j^\perp = c_j - \beta_j a_j. \tag{22}$$

Their magnitudes and directional agreement are

$$t_{ij} = \|h_{ij}^\perp\|_2, \qquad s_j = \|c_j^\perp\|_2, \qquad q_{ij} = \frac{(h_{ij}^\perp)^\top c_j^\perp}{t_{ij} s_j}. \tag{23}$$

Because

$$a_j^\top h_{ij}^\perp = 0, \qquad a_j^\top c_j^\perp = 0,$$

the cosine score can be written as

$$\begin{aligned} u_{ij} &= h_i^\top c_j \\ &= \left(\eta_{ij} a_j + h_{ij}^\perp\right)^\top \left(\beta_j a_j + c_j^\perp\right) \\ &= \eta_{ij}\beta_j + (h_{ij}^\perp)^\top c_j^\perp \\ &= \eta_{ij}\beta_j + t_{ij} s_j q_{ij}, \end{aligned} \tag{24}$$

which gives Eq. equation 10 in the main text.

This derivation also makes explicit that the decomposition is candidate dependent: the anchor a_j changes with novel candidate j, and therefore the orthogonal component of the same sample may differ across candidates.

The residual lifting in Eq. equation 11 changes only the magnitude weight applied to the residual-direction agreement. In particular, the square-root weighting remains magnitude aware: the lifted residual contribution still converges to zero as $t_{ij} s_j \to 0$.

For numerical stability, the implementation applies small positive lower bounds to residual norms and their products when evaluating normalized residual directions and the square-root weighting. These clamps prevent numerical instability in nearly degenerate residuals but do not alter the idealized decomposition above.

A.4 Centering and Novelty-Gating Details

This section provides two properties of the candidate-centering operation in Eq. equation 12 and the explicit group scores used by the novelty gate.

Candidate-weighted centering. Before centering, the distribution over novel candidates is computed from the base cosine scores:

$$p_{ij}^{n} = \frac{\exp(u_{ij}/\tau)}{\sum_{r\in\mathcal{Y}_n}\exp(u_{ir}/\tau)}. \tag{25}$$

Define the candidate-weighted correction mean

$$b_i = \sum_{j\in\mathcal{Y}_n} \mathrm{sg}(p_{ij}^{n}) d_{ij}. \tag{26}$$

Then Eq. equation 12 can be written as $\delta_{ij} = d_{ij} - b_i$. Because $\sum_j p_{ij}^{n} = 1$, the centered corrections satisfy

$$\sum_{j\in\mathcal{Y}_n} p_{ij}^{n}\delta_{ij} = 0. \tag{27}$$

Moreover, since the same b_i is subtracted from every novel candidate,

$$\delta_{ij} - \delta_{ik} = d_{ij} - d_{ik}. \tag{28}$$

Thus, centering does not by itself alter pairwise correction differences between novel candidates. Its direct effect is to remove their candidate-weighted common shift, thereby controlling the relative displacement of the novel-score group with respect to the known-score group.

Novelty tendency. The novelty component of the gate in Eq. equation 13 compares class-count-normalized aggregate scores from the known and novel groups. Specifically,

$$A_o(h_i) = \log\sum_{k\in\mathcal{Y}_o}\exp(u_{ik}/\tau) - \log K_o, \tag{29}$$

and

$$A_n(h_i) = \log\sum_{j\in\mathcal{Y}_n}\exp(u_{ij}/\tau) - \log K_n. \tag{30}$$

The novelty tendency is therefore

$$\nu_i = \mathrm{sg}\left[\sigma\left(A_n(h_i) - A_o(h_i)\right)\right]. \tag{31}$$

Subtracting $\log K_o$ and $\log K_n$ reduces the direct dependence of the aggregate scores on the number of categories in each group. However, ν_i should be interpreted as a model-derived novelty tendency, not as a calibrated posterior probability that sample i belongs to an unseen category.

Together with the normalized novel-category entropy defined in Eq. equation 13, this quantity activates residual correction primarily when novel candidates are competitive and their within-group assignment remains ambiguous. The correction strength in Eq. equation 14 is $\lambda_c = 1$ in the default configuration.

A.5 Additional Details of the Discovery Objective

The main paper summarizes the complete discovery objective in Eq. equation 17. Here we specify the view-level definitions and implementation conventions used by its individual terms.

Each training image is independently augmented into two views $v \in \{1, 2\}$. Both views are processed by the same trainable encoder and prototype library.

Supervised classification. For a labeled mini-batch $\mathcal{B}_l$, supervision is applied to both augmented views using the base logits z:

$$\mathcal{L}_{\mathrm{sup}} = \frac{1}{2|\mathcal{B}_l|}\sum_{i\in\mathcal{B}_l}\sum_{v=1}^{2} -\log\mathrm{softmax}\left(z_i^{(v)}\right)_{y_i}. \tag{32}$$

Residual-corrected logits are therefore not used by this supervised term.

Cross-view self-distillation. For each augmented view, we construct a sharpened stop-gradient prediction

$$q_i^{(v)} = \mathrm{sg}\left[\mathrm{softmax}\left(z_i^{(v)}/T_t\right)\right], \tag{33}$$

where T_t denotes the target temperature. The prediction from either view supervises the other:

$$\begin{aligned}\mathcal{L}_{\mathrm{dist}} = \frac{1}{2|\mathcal{B}_u|}\sum_{i\in\mathcal{B}_u}\Big[&\mathrm{CE}\left(q_i^{(1)}, \mathrm{softmax}(z_i^{(2)})\right)\\ &+\mathrm{CE}\left(q_i^{(2)}, \mathrm{softmax}(z_i^{(1)})\right)\Big].\end{aligned} \tag{34}$$

Both views are generated by the same trainable encoder. There is no independent EMA teacher during discovery training; the EMA reference encoder is used only in the preceding geometry-regularized adaptation stage.

Balance-only old-class scaling. The old-class scaling factor ξ_i is defined in Eq. equation 15. It is applied only to the copy of the logits used to compute the marginal balance objective:

$$z_{ik}^{\mathrm{bal}} = \begin{cases}\xi_i z_{ik}, & k \in \mathcal{Y}_o,\\ z_{ik}, & k \in \mathcal{Y}_n.\end{cases} \tag{35}$$

For a mixed batch $\mathcal{B}$ containing the labeled and unlabeled training samples, the marginal prediction is

$$\bar{p} = \frac{1}{2|\mathcal{B}|}\sum_{i\in\mathcal{B}}\sum_{v=1}^{2}\mathrm{softmax}\left(z_i^{\mathrm{bal},(v)}\right). \tag{36}$$

The scaling in Eq. equation 35 is confined to $\mathcal{L}_{\mathrm{bal}}$. It is not used by the supervised loss, cross-view distillation, residual correction, or prediction outside this balance objective.

Novelty-weighted information maximization. The information-maximization term in Eq. equation 16 operates on the corrected novel-category logits from Eq. equation 14. For each sample, the novel-category distributions from the two views are averaged before computing the information objective.

The corresponding sample weight is

$$w_i = \frac{1}{2}\left(\nu_i^{(1)} + \nu_i^{(2)}\right).$$

Importantly, this weight uses only the novelty tendency ν_i. It does not use the complete correction gate $g_i = \nu_i e_i$. Thus, novel-category ambiguity controls where residual correction is applied, whereas the information-maximization objective is weighted according to the model's estimated tendency for a sample to belong to the novel group.

The four discovery objectives are combined exactly as specified in Eq. equation 17.

A.6 Inference and Score Paths

For clarity, the objectives use the following score paths:

Component	Scores	Samples
Supervised CE	original z	labeled
Cross-view distillation	original z	unlabeled
Marginal entropy	old-scaled original scores	mixed batch
Novel InfoMax	corrected novel $\widehat{z}$	mixed batch
Inference	original z	evaluation samples

Inference predicts

$$\widehat{y}_i = \arg\max_{k\in\mathcal{Y}_o\cup\mathcal{Y}_n} z_{ik}. \tag{37}$$

Although the residual branch leaves the numerical values of the known logits unchanged for fixed features and prototypes, changing the novel scores can alter known-versus-novel competition. The method therefore does not guarantee unchanged known-class predictions.

Table 6: Stability of ARISE across three random seeds. Results are reported as mean ± standard deviation (%). All, Old, and New denote accuracy on all, known, and novel classes, respectively.

Dataset	All	Old	New
CIFAR-100	85.3 ± 0.21	84.90 ± 0.42	86.47 ± 0.17
ImageNet-100	86.33± 0.17	94.77 ± 0.24	82.07 ± 0.19
CUB-200	76.47 ± 0.25	76.50 ± 0.57	76.46 ± 0.41
Stanford Cars	70.03 ± 0.09	82.53 ± 0.37	64.03 ± 0.05
Herbarium19	49.57 ± 0.21	61.07 ± 0.50	43.43 ± 0.57
FGVC-Aircraft	63.00 ± 0.22	66.33 ± 0.46	61.33 ± 0.50

A.7 Sequential Optimization

The two stages of ARISE are optimized separately. The geometry loss $\mathcal{L}_{\text{geo}}$ is used only during known-class adaptation and is not evaluated during discovery training. The adapted encoder initializes the subsequent discovery stage, where the prototype library is reconstructed from the adapted features.

In the current implementation, the main discovery script loads an offline adapted checkpoint rather than running adaptation inline. For reproducibility, the provenance and configuration used to produce this checkpoint should therefore be reported together with the discovery-stage configuration.

B Additional Experimental Analysis

This appendix provides additional experiments that complement the main results and clarify the effectiveness of the major components in ARISE. We focus on three questions: (1) whether representation adaptation provides gains beyond additional optimization, (2) how prototype initialization and old-class balancing affect discovery, and (3) how the reference design influences geometry-aware adaptation.

B.1 Stability across Random Seeds

We further evaluate the stability of ARISE by repeating each experiment with three different random seeds. Table 6 reports the mean accuracy and standard deviation over the three runs for All, Old, and New classes. The results are presented as mean ± std (%).

B.2 Comparison with Plug-and-Play Extensions

A notable trend in recent GCD research is to develop generic plug-and-play components that can be incorporated into existing discovery frameworks. This design is particularly common among recent works. For example, ATCG (Han et al., 2026) augments existing GCD pipelines with analogically generated textual concepts; HypCD (Liu et al., 2025b) replaces the conventional spherical representation and classifier with their hyperbolic counterparts; EAGC (Zheng et al., 2026) introduces a gradient-level coordinator without redesigning the underlying discovery objective; and CoGe-GCD (Tang et al., 2026) inserts a compositional inductive-bias module between the backbone and projection head while retaining the original head and loss functions.

Since these methods are explicitly designed to improve existing GCD frameworks, we use their SimGCD-based variants as a common reference for this supplementary comparison. This setting allows us to examine how different recent mechanisms improve a representative parametric GCD baseline, rather than conflating their gains with differences between entirely different base frameworks.

Importantly, ARISE should not be interpreted as another plug-in applied directly on top of SimGCD. We use the parametric formulation of SimGCD as a starting point to construct our own framework, but substantially redesign the learning pipeline with geometry-guided representation adaptation, Sinkhorn-based assignment refinement, known-class bias calibration, and residual-aware learning and correction. Therefore, the comparison below is intended to position ARISE against recent

Table 7: Comparison with recent SimGCD-based extensions on fine-grained benchmarks. For plug-and-play methods, we report the SimGCD baseline used in the corresponding paper and its absolute improvement in parentheses. ARISE is not a direct SimGCD plug-in; SimGCD is used only as the reference formulation from which our framework is constructed.

Method	Backbone	CUB-200	SCars	Aircraft	Avg. Δ
$\text{SimGCD}_{\text{ATCG}}$	CLIP	69.6	69.4	53.5	–
+ ATCG (Han et al., 2026)	CLIP	74.7 (+5.1)	78.3 (+8.9)	58.6 (+5.1)	+6.4
SimGCD	DINO	60.3	53.8	54.2	–
+ HypCD (Liu et al., 2025b)	DINO	64.8 (+4.5)	62.8 (+9.0)	58.7 (+4.5)	+6.0
+ EAGC (Zheng et al., 2026)	DINO	66.5 (+6.2)	62.9 (+9.1)	57.7 (+3.5)	+6.3
$\text{SimGCD}_{\text{CoGe}}$	DINO	61.6	51.8	54.5	–
+ CoGe-GCD (Tang et al., 2026)	DINO	63.9 (+2.3)	58.6 (+6.8)	57.9 (+3.4)	+4.1
ARISE (ours)	DINO	**76.4**	**70.1**	**63.1**	–

SimGCD-based extensions under a related reference framework, rather than to claim that all methods share an identical architecture.

Analysis. Table 7 compares ARISE with recent methods that extend representative GCD frameworks through modular mechanisms. The paired SimGCD results show that these plug-and-play designs are effective: ATCG, HypCD, EAGC, and CoGe-GCD improve their corresponding SimGCD baselines by 6.4, 6.0, 6.3, and 4.1 points on average, respectively. Their gains arise from complementary directions, including external semantic concepts, hyperbolic representation geometry, gradient coordination, and compositional inductive bias.

A consistent pattern is that the largest improvements are often observed on Stanford Cars. ATCG, HypCD, EAGC, and CoGe-GCD improve their respective SimGCD baselines by 8.9, 9.0, 9.1, and 6.8 points, respectively. This observation is consistent with the motivation that fine-grained discovery is particularly sensitive to representation ambiguity, since visually similar categories share substantial semantic and structural information.

Under the DINO setting, ARISE achieves 76.4%, 70.1%, and 63.1% All accuracy on CUB-200, Stanford Cars, and FGVC-Aircraft, respectively. Compared with the reported DINO-based extensions, ARISE obtains substantially higher absolute accuracy on all three fine-grained benchmarks. In particular, the improvement is pronounced on CUB-200, where ARISE reaches 76.4%, compared with 64.8% for HypCD, 66.5% for EAGC, and 63.9% for CoGe-GCD.

This comparison also highlights a distinction in methodology. Recent plug-and-play approaches typically target one particular bottleneck of an existing framework: HypCD modifies representation geometry, EAGC regulates gradient interference, and CoGe-GCD injects compositional structure. In contrast, ARISE is constructed as an integrated discovery framework in which representation adaptation, assignment refinement, known-class bias calibration, and residual modeling are jointly designed to address ambiguity throughout the learning and inference pipeline. Although its formulation is developed from the parametric GCD paradigm exemplified by SimGCD, ARISE does not simply attach an additional module to a frozen SimGCD pipeline.

The absolute values across all rows should nevertheless be interpreted with appropriate care. ATCG adopts a CLIP-based SimGCD variant, whereas ARISE, HypCD, EAGC, and CoGe-GCD use DINO-based representations. Moreover, CoGe-GCD reports its own SimGCD reproduction. We therefore use the paired baseline improvements to characterize the effectiveness of individual plug-and-play mechanisms, while using the DINO-based results primarily to contextualize the overall performance of ARISE.

B.3 Effect of Prototype Initialization and Class Balancing

We next isolate the contributions of Sinkhorn initialization and old-class balancing. These components address complementary aspects of GCD: Sinkhorn improves the initialization of latent novel categories, whereas old-class scaling adjusts the competition between known and novel prototypes.

Table 8: Ablation of prototype initialization and balancing objectives.

Configuration	CUB			SCars		
	All	Old	New	All	Old	New
Full	76.4	75.7	76.8	70.1	82.8	64.0
No Sinkhorn	69.8	72.4	68.5	66.6	79.9	60.3
No old-class scaling	74.2	79.8	71.4	67.6	80.9	61.2
No marginal entropy	73.8	72.9	74.2	66.9	79.9	60.6

Table 9: Reference-design variants within the same downstream discovery configuration.

Adaptation	CUB			SCars		
	All	Old	New	All	Old	New
Fixed DINO reference	65.4	71.5	62.4	57.2	77.3	47.4
EMA feature matching	73.4	73.9	73.2	67.6	81.2	61.0
EMA relations w/o PCA	74.2	73.9	74.3	67.3	81.2	60.6
Full	76.4	75.7	76.8	70.1	82.8	64.0

Removing Sinkhorn consistently degrades discovery performance. All accuracy decreases by **4.0** points on CUB and **0.7** points on SCars, while New accuracy decreases by **5.1** and **1.1** points, respectively. The larger degradation on novel categories confirms that improved prototype initialization is particularly important for category discovery.

Old-class scaling mainly controls the balance between known- and novel-category predictions. Removing it increases CUB Old accuracy from 74.0% to 79.8%, but reduces New accuracy from 73.6% to 71.4%. A similar, although smaller, tradeoff appears on SCars. Thus, the balancing term does not simply maximize Old accuracy; instead, it prevents known categories from dominating the joint prediction space and maintains stronger performance on novel categories.

Removing the marginal-entropy component has only a modest effect on aggregate accuracy, but its effect differs across the two datasets. We therefore treat it as a complementary regularizer rather than the main source of the performance improvement.

B.4 Analysis of Geometry-Regularized Adaptation

Finally, we examine the reference design used during representation adaptation. The central comparison is between a fixed pretrained reference and an adaptive EMA reference.

Using a fixed pretrained reference causes a substantial degradation: compared with the full EMA-based formulation, All accuracy decreases by **8.4** points on CUB and **10.2** points on SCars, while New accuracy decreases by **11.3** and **13.9** points. This result indicates that rigidly preserving the pretrained representation is unsuitable for the downstream discovery objective.

In contrast, all EMA-based variants recover substantially stronger performance. The differences among EMA feature matching, relational alignment without PCA, and the full formulation are relatively small at the final checkpoint. Accordingly, the strongest conclusion supported by these experiments is the importance of using an *adaptive* reference rather than a fixed pretrained geometry, rather than claiming that every individual geometric design choice independently improves accuracy.

B.5 Summary of Additional Results

The additional experiments support three main observations. First, known-class representation adaptation provides the dominant improvement and cannot be reproduced by simply increasing the number of discovery updates. Second, Sinkhorn-based prototype initialization consistently improves novel-category recognition, with particularly large gains on CUB. Third, maintaining an adaptive EMA reference is substantially more effective than constraining the encoder to a fixed pretrained representation.

Table 10: Correction gates and anchor construction.

Variant	CUB-200			SCars		
	All	Old	New	All	Old	New
Correction gates						
No novelty gate	74.7	73.8	75.1	66.2	81.9	58.7
No ambiguity gate	74.8	73.9	75.3	65.7	82.1	57.8
No gates	74.7	74.9	74.6	66.1	81.6	58.5
Anchor construction						
Uniform anchor	75.2	73.8	75.9	66.2	81.8	58.6
Nearest anchor	74.8	73.5	75.4	66.0	81.8	58.3

Table 11: Geometry-batch construction.

Variant	CUB-200			SCars		
	All	Old	New	All	Old	New
Random batches	74.7	75.0	74.5	66.9	82.4	59.4
Full	76.4	75.7	76.8	70.1	82.8	64.0

The balancing components further regulate the Old–New tradeoff, preventing improvements on known categories from dominating novel-category discovery.

B.6 Gates, anchors, and batches.

Gate-removal variants span 74.7–74.8% All accuracy on CUB-200 and 65.7–66.2% on SCars (Table 10). Uniform anchors exceed nearest anchors by 0.5 and 0.3 New points, respectively. Random geometry batches yield 74.7% and 66.9% All accuracy (Table 11).

Underlying objectives. Removing global MeMax yields 61.7% and 49.9% All accuracy; removing view distillation yields 70.3% and 58.9% (Table 12).

C Visualization

C.1 How Residual Correction Reshapes Novel-Class Logit Distributions

We examine the distributional effect of anchored residual correction immediately after warm-up, with the backbone and prototypes held fixed, as shown in Fig. 4. The three supplied 10-epoch warm-up checkpoints do not contain classifier prototypes; these are reconstructed using the current Ward–Sinkhorn initialization, without using unlabeled ground-truth labels. This is a separate diagnostic from the discovery-stage checkpoint experiments. We include all 2,997, 4,143, and 3,335 true-novel examples on CUB, Stanford Cars, and FGVC-Aircraft, respectively. No additional training, temperature search, or outcome-based sample selection is performed.

For each example, let z^B, z^F denote the actual Base and Full novel logits and $p^v = \mathrm{softmax}(z^v)$ over all novel candidates. We measure

$$H_i^v = -\frac{\sum_j p_{ij}^v \log p_{ij}^v}{\log K_n}, \qquad G_i^v = z_{i,(1)}^v - z_{i,(2)}^v, \qquad v \in \{B, F\}. \tag{38}$$

Candidate identities in the rank-conditioned logit analysis are fixed by Base; the two largest logits defining G_i^v are computed separately for each variant.

Observed redistribution. The Base-leading novel candidate receives a mean logit change of $+0.061$, $+0.096$, and $+0.064$, whereas the Base runner-up receives -0.057, -0.024, and -0.022,

Table 12: Underlying training objectives.

Variant	CUB-200			SCars		
	All	Old	New	All	Old	New
No global MeMax	61.7	68.5	58.3	49.9	72.3	39.0
No distillation	70.3	76.6	67.1	58.9	80.0	48.7
Full	76.4	75.7	76.8	70.1	82.8	64.0

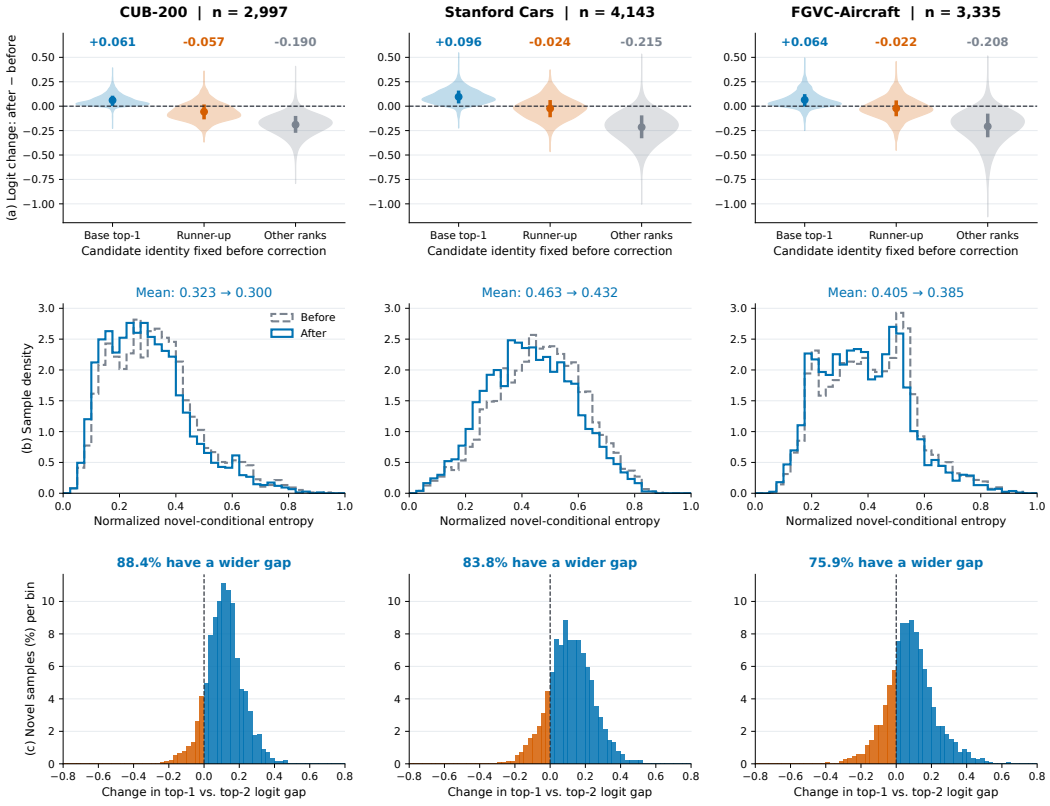


Figure 4: **Residual correction redistributes novel-class logits.** Columns correspond to datasets; all true-novel examples are included. (a) Actual logit-change distributions grouped by the Base top-1, runner-up, and remaining candidates. Dots are means and thick lines are interquartile ranges; violin widths are normalized separately by group. (b) Before/after distributions of normalized novel-conditional entropy. (c) Paired changes in the top-two novel logit gap, with positive changes in blue and negative changes in orange. The axes and histogram bins are shared within each row, and no histogram tails are discarded. The gap is between leading predictions, not necessarily the true class and an incorrect rival.

respectively. Normalized entropy decreases for 98.2%, 98.0%, and 93.5% of novel examples, and the top-two logit gap increases for 88.4%, 83.8%, and 75.9%. These observations show a widespread concentration of novel-candidate scores, even though only a small fraction of predictions change.

C.1.1 Direct Pre-Discovery Verification of Residual-Magnitude Attenuation

The algebraic factorization of a residual inner product does not by itself show that magnitude attenuation suppresses useful evidence. We therefore perform a direct diagnostic immediately after the 10-epoch warm-up, before discovery training can adapt the representation or prototypes to

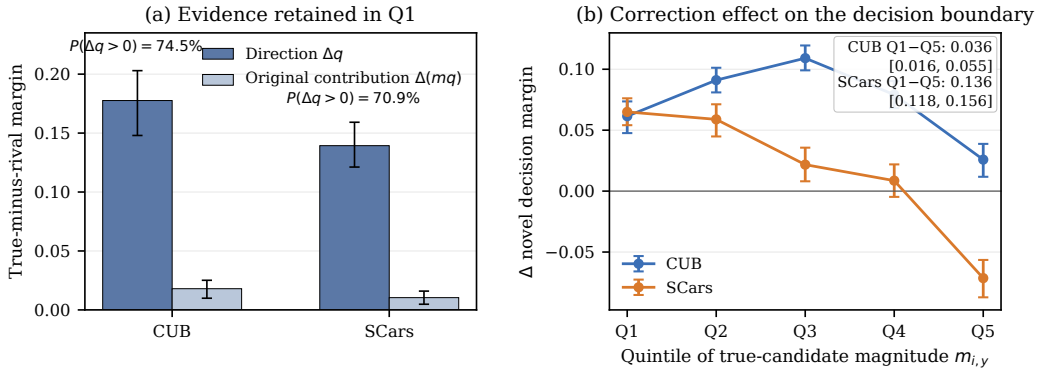


Figure 5: **Low residual magnitude suppresses directional evidence that remains discriminative.** All intervals are 95% bootstrap intervals over true-novel examples. **Left:** In the lowest true-candidate magnitude quintile (Q1), $\Delta q = q_{i,y} - q_{i,r}$ is positive on average and favors the true candidate for 74.5% of CUB examples and 70.9% of SCars examples. Multiplication by the candidate-specific residual magnitudes reduces the corresponding contribution $\Delta(mq) = m_{i,y}q_{i,y} - m_{i,r}q_{i,r}$. **Right:** Change in the true-candidate novel decision margin after correction, with the strongest incorrect candidate reselected for each scoring rule. The Q1-minus-Q5 contrasts are positive on both datasets, showing that the measured boundary improvement is stronger at low residual magnitude.

Table 13: Direct residual-magnitude diagnostic in Q1. Δq is the true-minus-rival directional margin, $\Delta(mq)$ is its contribution under the original residual inner product, and ΔM_{dec} is the corrected minus Base novel decision margin after reselecting the strongest incorrect candidate. The final column directly bootstraps the Q1-minus-Q5 difference in decision-margin gain. $P(\Delta q > 0)$ is in percent; bracketed values are 95% bootstrap intervals.

Dataset	N	$m_{i,y}$	$P(\Delta q > 0)$	Δq	$\Delta(mq)$	ΔM_{dec}	Q1−Q5 ΔM_{dec}
CUB	600	0.220	74.5	0.178 [0.148, 0.203]	0.018 [0.010, 0.025]	0.061 [0.048, 0.074]	0.036 [0.016, 0.055]
SCars	829	0.250	70.9	0.139 [0.121, 0.159]	0.010 [0.005, 0.016]	0.065 [0.054, 0.076]	0.136 [0.118, 0.156]

corrected scores. We freeze the CUB and SCars features, reconstruct the official labeled-known and Ward–Sinkhorn prototype initialization, and change only the production correction strength from zero to one. Unlabeled ground-truth labels are hidden during initialization and used only for post-hoc correspondence and evaluation.

For novel candidate j, write the residual contribution as $m_{ij}q_{ij}$, where $m_{ij} = t_{ij}s_j$ is the product of the sample and prototype residual norms and q_{ij} is their directional cosine. We group all true-novel examples into quintiles of the true-candidate magnitude $m_{i,y}$. A novel-only Hungarian correspondence is fitted once from Base predictions. For directional and residual margins, the rival is the strongest incorrect novel candidate under Base and remains fixed. For the decision margin, the strongest incorrect candidate is reselected separately under Base and corrected scores; hence this diagnostic measures the actual novel top-1 boundary rather than only one favorable candidate pair.

Figure 5 and Table 13 establish both parts of the diagnostic. First, Q1 retains a positive directional margin even though its contribution to the original cosine score is much smaller. Because the true and rival candidates have their own magnitudes, this pairwise comparison is not interpreted as a single common shrinkage factor. Second, correction increases the decision margin against the currently strongest incorrect novel candidate. The Q1-minus-Q5 gain is 0.0356 with interval $[0.0156, 0.0545]$ on CUB and 0.1364 with interval $[0.1176, 0.1560]$ on SCars. Thus the intervention has a targeted and measurable effect on low-magnitude decision boundaries. These intervals condition on the fixed warm-up checkpoints, initialized prototypes, correspondence, and quintile boundaries; they do not estimate training-seed variation.