SCoRP: Soft Consensus and Residual KL Projection for Feasible Clustering

Anonymous Authors

Abstract

Frozen foundation representations often make simple clustering methods competitive, but converting their outputs into stable, feasible hard partitions remains difficult. Soft assignments preserve uncertainty but may lose requested clusters after rowwise decoding; early hard pseudo-labels can amplify errors; and strict equipartition can distort imbalanced data. We introduce *Soft Consensus and Residual Projection* (SCoRP), a label-free soft-to-hard partition operator for fixed feature representations. It aligns multiple seed partitions into a soft consensus, propagates that consensus with a stochastic graph resolvent, estimates an adaptive cluster-mass target, and applies an anchored residual generalized Kullback–Leibler (GKL) projection before a decoder that restores feasibility with the minimum number of changes to the unconstrained rowwise-$\arg\max$ partition. We establish joint graph–posterior perturbation stability, uniqueness and strict convexity of the reduced projection problem, an exact generalized-KL projection identity, exact soft mass control, minimum-Hamming hard feasibility, set-valued label-permutation equivariance with a conditional pointwise guarantee, and a sufficient neighborhood-preservation condition for optional bounded *Confidence-Refined Adaptive Feature Transformation* (CRAFT). An 18-dataset audit spans image, text, tabular, and sensor data and retains favorable and unfavorable outcomes. Representative clustering-accuracy (ACC) changes against the strongest standard same-feature baseline are $+3.91$ points on CIFAR-10, $+3.15$ on CIFAR-100, $+1.64$ on EuroSAT, and $+8.61$ on COIL-20. The mixed results support SCoRP as a controlled soft-to-hard partition operator when the supplied representation is informative, not as a universal remedy for inadequate feature geometry.

1 Introduction

Frozen self-supervised and multimodal encoders increasingly make clustering a *post-representation* problem. DINOv2, Contrastive Language–Image Pre-training (CLIP), and related features can support strong partitions without encoder fine-tuning (Oquab et al., 2024; Gadetsky et al., 2024). Yet a useful feature space does not by itself determine how uncertain assignments, neighborhood geometry, cluster-size information, and hard occupancy should be reconciled. This final partitioning stage remains consequential: small changes in graph construction, initialization, or decoding can produce different hard partitions from the same frozen features.

The difficulty is partly a mismatch between soft and hard requirements. A soft assignment matrix may give every cluster positive column mass even though a cluster is never the largest entry in any row, so rowwise $\arg\max$ can remove a cluster that is present in the soft solution. Enforcing hard pseudo-labels earlier avoids that failure but can preserve incorrect decisions and amplify them through later updates. Global size constraints introduce a separate trade-off: strict equipartition prevents collapse but can distort genuinely imbalanced groups, while relaxed balancing addresses only part of this tension (Caron et al., 2020; Zhang et al., 2024). PRO-DSC gives structural anti-collapse conditions in deep subspace clustering (Meng et al., 2025), while probabilistic-cut methods optimize graph-partition objectives (Ghriss and Monteleoni, 2025; Ghriss, 2026). Our focus is complementary: a controlled partition operator on top of fixed features, not another encoder or cut surrogate.

We address these issues after representation extraction. SCoRP first aligns a diverse bank of seed partitions and preserves their agreement and disagreement in a soft consensus. A stochastic graph resolvent then propagates this evidence through local neighborhoods. Instead of imposing uniform cluster sizes, an adaptive target interpolates between the consensus mass estimate and equipartition. High-margin samples may be anchored, while only the remaining mass is reconciled through residual generalized-KL projection. A discrete repair is deferred to the end and changes the minimum possible number of unconstrained $\arg\max$ labels needed to satisfy the requested hard support.

Design rationale. The separation is deliberate. Consensus represents uncertainty across plausible seeds; propagation incorporates local geometry; residual projection controls global cluster mass; anchors preserve selected high-confidence

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.

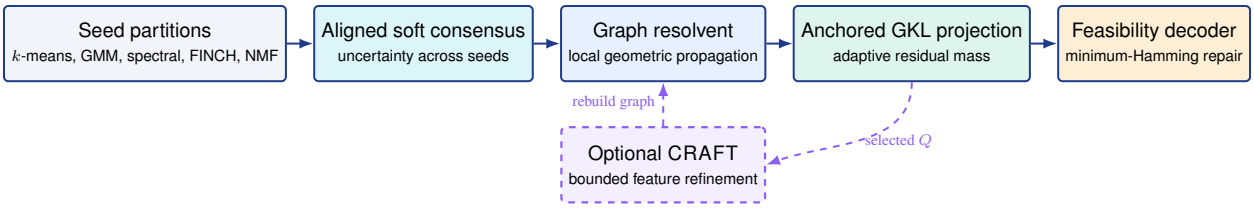

Figure 1: **SCoRP pipeline.** GMM, FINCH, and NMF denote Gaussian mixture model, First Integer Neighbor Clustering Hierarchy, and nonnegative matrix factorization, respectively. Colors distinguish functional roles: gray denotes seed generation, cyan consensus, blue graph propagation, green residual GKL mass projection, orange hard feasibility decoding, and purple the optional CRAFT refinement. Solid arrows form the mandatory path. After the base candidate is selected, the dashed branch trains CRAFT from its soft posterior, rebuilds the graph on the adapted features, and reruns the downstream propagation–projection–decode stages; the seed bank and aligned consensus are not regenerated. The refined branch is retained only when the predefined label-free stability criteria improve.

assignments; and the hard repair acts only at the final discrete interface. The soft projection is minimal in generalized-KL geometry, whereas the decoder is minimal in Hamming distance. Neither guarantee asserts that the frozen representation already matches an external semantic taxonomy. Optional CRAFT provides a bounded residual feature update retained only when the predefined label-free stability diagnostics improve (Figure 1).

Contributions.

1. We formulate the post-representation stage as a label-free soft-to-hard operator: SCoRP combines aligned consensus, graph-resolvent propagation, adaptive residual generalized-KL projection with confidence anchors, and minimum-Hamming feasibility decoding.
2. We establish guarantees for the transitions between these stages: adaptive mass interpolation contracts deviation from uniform; graph propagation is stable to joint graph/posterior perturbations; the anchored residual projection has a unique positive solution and a strictly convex $(K-1)$-dimensional reduced dual; and the combined posterior attains the prescribed soft cluster masses.
3. We show that the hard decoder changes the minimum possible number of unconstrained $\arg\max$ assignments subject to the required counts. The downstream soft stages are label-permutation equivariant, while the set of minimum-Hamming hard repairs is equivariant and the point-valued decoder is equivariant under label-symmetric unique choices. For optional bounded CRAFT, we also give a sufficient cosine-geometry and kNN-preservation condition.
4. We evaluate an 18-dataset audit with development, frozen-feature evaluation, and post-development evaluation roles, retain favorable and unfavorable outcomes, and include a matched-feature TURTLE comparison on identical DINOv2-S/14 matrices. Native TURTLE and PRO-DSC runs that use different CLIP representations remain separate contextual comparisons rather than same-feature claims.

2 Related Work

Clustering, graph partitioning, and consensus. Clustering is intrinsically underdetermined because different objectives encode different notions of a useful grouping (MacQueen, 1967; Jain, 2010). Graph-based methods replace Euclidean prototypes by neighborhood structure; normalized cuts and spectral clustering connect partitions to graph Laplacian eigenvectors (Shi and Malik, 2000; Ng et al., 2001; von Luxburg, 2007), with consistency results clarifying when normalized spectral constructions converge (von Luxburg et al., 2008). A complementary line combines multiple imperfect partitions through cluster ensembles, evidence accumulation, and probabilistic consensus (Strehl and Ghosh, 2002; Fred and Jain, 2005; Topchy et al., 2005; Boongoen and Iam-On, 2018). Graph- and sparsity-regularized factorization also improves robustness by combining neighborhood structure with constrained representations (Esmaeili Salehani et al., 2018). SCoRP instead operates after representation extraction: seed disagreement is retained as a soft posterior and then reconciled with graph geometry and explicit cluster-mass constraints.

Graph diffusion, random walks, and stability. Random-walk operators provide a classical bridge between local neighborhoods and multiscale geometry. Diffusion maps use powers of a Markov operator to expose data geometry (Coifman and Lafon, 2006); personalized-PageRank propagation gives a closely related resolvent form in modern graph learning (Gasteiger et al., 2019). Recent multi-view clustering work has also used adaptive multi-scale diffusion and random-walk-based contrastive structure to couple local and global graph information (Chen et al., 2026). A separate line studies clustering stability under graph perturbations, including spectral-gap/distance-to-ambiguity analyses (Andreotti et al., 2021) and average sensitivity of spectral clustering to edge perturbations (Peng and Yoshida, 2020). These works motivate diffusion and stability as geometric mechanisms but do not provide the particular posterior-level soft-to-hard feasibility interface considered here. The resolvent itself is classical; the SCoRP-specific result is the joint graph/initialization perturbation bound in Theorem 1 and its posterior-margin certificate in Corollary 1, which turn a graph/posterior perturbation radius into a sufficient condition for hard-label invariance.

Transport, imbalance, and mass control. The residual projection block sits on established information-geometric foundations. Csiszár characterized I-projections onto convex constraint sets (Csiszár, 1975); Sinkhorn scaling and its convergence theory provide the matrix-scaling perspective (Sinkhorn and Knopp, 1967; Franklin and Lorenz, 1989); and entropic transport can be computed through KL/Bregman projections (Cuturi, 2013; Benamou et al., 2015; Peyré and Cuturi, 2019). In representation learning, SwAV uses balanced transport to avoid collapse (Caron et al., 2020), while BPA transforms features using a self-transport construction (Shalam and Korman, 2024). Size-constrained clustering has also been studied directly, including balanced k-means (Malinen and Fränti, 2014). Recent transport-based clustering work includes progressive partial transport for deep imbalance (Zhang et al., 2024), partial transport within multi-view contrastive learning (Xue et al., 2025), and entropic-transport objectives for model-based clustering (Mena, 2026). These methods optimize pseudo-labeling, representation-learning, or model-based objectives. SCoRP uses a different interface: the target mass is estimated from a seed ensemble, hard anchors first consume part of the marginal, and only the residual free mass is projected. Here "transport" in the method name refers to this residual-marginal generalized-KL projection; the method does not introduce a ground-cost matrix. This yields both a reduced strictly convex dual and the KL-minimal-intervention identity of Corollary 2.

Deep clustering and frozen foundation representations. Deep embedded clustering jointly learns features and assignments (Xie et al., 2016); DeepCluster alternates pseudo-label construction and representation updates (Caron et al., 2018); and SCAN separates representation learning from semantic clustering (Van Gansbeke et al., 2020). DINOv2 demonstrates strong transferable self-supervised visual features (Oquab et al., 2024), motivating clustering layers that do not fine-tune the encoder. TURTLE directly searches for clusterings compatible with frozen foundation-model spaces (Gadetsky et al., 2024). TURTLE is especially relevant to our setting because it also operates on frozen representations. We therefore run a matched-feature version on the identical saved feature matrices; results from its released native representation are reported separately and are not used for same-feature claims.

Subspace, graph-cut, and anti-collapse methods. Sparse subspace clustering provides theory for self-expressive unions of subspaces (Elhamifar and Vidal, 2013). The Principled Framework for Deep Subspace Clustering (PRO-DSC) develops a deep self-expressive framework with structural anti-collapse guarantees (Meng et al., 2025); I-Con gives an information-geometric view relating clustering, spectral objectives, contrastive learning, and supervision (Alshammari et al., 2025). PRCut and subsequent probabilistic-cut work optimize differentiable relaxations of RatioCut and broader graph-cut families (Ghriss and Monteleoni, 2025; Ghriss, 2026), while MiniClustering targets long-tailed deep clustering (Li et al., 2026). SCoRP is complementary: it neither learns a self-expressive representation nor replaces the clustering problem by a cut objective. It composes a consensus prior, a stochastic geometry operator, a convex mass projection, and a minimal discrete repair.

Relation to established components. Markov diffusion, KL/I-projection, matrix scaling, cluster ensembles, and size constraints are established tools. The contribution here is their soft–hard composition and the guarantees required at the interfaces: adaptive mass contraction, graph-to-label perturbation certification, anchored residual generalized-KL projection with a reduced dual and exact KL minimality, minimum-Hamming feasibility repair, and a bounded refinement with a sufficient neighborhood-preservation condition. Appendix A separates these specialized conclusions from their classical foundations.

3 Problem Setting

Let $X = [x_1, \dots, x_n]^\top \in \mathbb{R}^{n \times d}$ be unlabeled features and let $1 \le K \le n$ be the requested number of clusters. Ground-truth labels are unavailable to fitting, hyperparameter selection, and final candidate selection. After ℓ_2 normalization, we construct a sparse symmetric affinity $A \ge 0$ from cosine kNN neighborhoods with local scaling. With $D = \mathrm{diag}(A\mathbf{1})$ and no isolated vertices,

$$P = D^{-1}A, \tag{1}$$

Thus $P \ge 0$ and $P\mathbf{1} = \mathbf{1}$, so each row of P is a probability distribution over graph neighbors and defines a random-walk transition distribution. A soft partition is $Q \in \mathbb{R}_+^{n \times K}$ with $Q\mathbf{1}_K = \mathbf{1}_n$. The desired output is a hard partition $y \in [K]^n$ in which all K requested clusters are represented. Throughout, $\|M\|_\infty = \max_i \sum_j |M_{ij}|$ denotes the induced matrix infinity norm. The central problem is to transform a noisy soft consensus into such a feasible hard partition while changing high-confidence local evidence as little as possible and without using labels to select the transformation.

4 Method: Soft Consensus and Residual Projection

4.1 Aligned consensus and adaptive target mass

We execute a diverse bank of noncollapsed unsupervised partitions, including k-means (MacQueen, 1967), diagonal Gaussian mixtures, spectral-kNN (Ng et al., 2001), FINCH (Sarfraz et al., 2019), and NMF+k-means (Lee and Seung, 1999) when feasible. Because cluster indices are arbitrary, each candidate partition is aligned to a reference seed by maximum-weight bipartite (Hungarian) matching (Kuhn, 1955); the best label-free seeds then form

$$Q^{(0)} = \sum_{s \in \mathcal{S}} w_s H_s, \qquad w_s \ge 0, \quad \sum_s w_s = 1, \tag{2}$$

where H_s is one-hot after alignment. Weights are a fixed softmax of internal seed-quality scores.

Let $\pi_0 = \frac{1}{n}(Q^{(0)})^\top \mathbf{1}$ and $u = \mathbf{1}_K/K$. We write $\pi_{\rho,c} = [\pi_\rho]_c$ for the cth cluster mass. Instead of strict equipartition,

$$\pi_\rho = (1-\rho)\pi_0 + \rho u, \qquad \rho \in [0,1]. \tag{3}$$

The center $\rho_\star$ of a small deterministic grid is estimated from normalized entropy of π_0 and Jensen–Shannon disagreement between seed-mass histograms.

Proposition 1 (Adaptive balancing contracts divergence to uniform) *For any norm,*

$$\|\pi_\rho - u\| = (1-\rho)\,\|\pi_0 - u\|\,, \qquad \pi_{\rho,c} \ge \rho/K.$$

Moreover,

$$D_{\mathrm{KL}}(\pi_\rho \| u) \le (1-\rho) D_{\mathrm{KL}}(\pi_0 \| u), \tag{4}$$

and therefore Shannon entropy satisfies $H(\pi_\rho) \ge H(\pi_0)$.

The first identity gives the exact interpolation geometry, while Eq. (4) shows that the same parameter also controls information-theoretic deviation from equipartition.

4.2 Graph-resolvent propagation and perturbation stability

The use of a row-stochastic random walk follows classical diffusion/spectral geometry (Coifman and Lafon, 2006; von Luxburg, 2007); the role of the theorem below is not to re-prove diffusion theory, but to specialize it to posterior propagation and quantify when graph perturbations cannot change hard labels. For $\alpha \in [0,1)$ define

$$G_\alpha(P) = (1-\alpha)(I - \alpha P)^{-1} = (1-\alpha)\sum_{\ell=0}^{\infty} \alpha^\ell P^\ell. \tag{5}$$

We compute the resolvent by the fixed-point iteration

$$R_{m+1} = (1-\alpha)Q + \alpha P R_m, \qquad R_0 = Q. \tag{6}$$

Theorem 1 (Stochasticity, contraction, convergence, and graph perturbation stability) *Let $P, \widetilde{P}$ be row-stochastic, let $Q, \widetilde{Q}$ be soft partitions, and $0 \le \alpha < 1$. Then $G_\alpha(P)$ is row-stochastic and Eq.* (6) *remains in the row simplex. The map $T_P(Z) = (1-\alpha)Q + \alpha PZ$ is an α-contraction in $\|\cdot\|_\infty$ with unique fixed point $R^\star = G_\alpha(P)Q$, and*

$$\|R_m - R^\star\|_\infty \le \alpha^m \|R_0 - R^\star\|_\infty .$$

Furthermore,

$$\left\|G_\alpha(P)Q - G_\alpha(\widetilde{P})\widetilde{Q}\right\|_\infty \le \left\|Q - \widetilde{Q}\right\|_\infty + \frac{\alpha}{1-\alpha}\left\|P - \widetilde{P}\right\|_\infty . \tag{7}$$

Corollary 1 (Margin-certified hard stability) *Let $R = G_\alpha(P)Q$ and $\widetilde{R} = G_\alpha(\widetilde{P})\widetilde{Q}$. If every row of R has top-two gap larger than 2δ and the right-hand side of Eq.* (7) *is at most δ, then $\arg\max_c R_{ic} = \arg\max_c \widetilde{R}_{ic}$ for all i.*

Corollary 1 certifies the exact resolvent posterior. For the computed finite iterate, the same argument applies after adding its fixed-point approximation error to the perturbation radius. Thus large computed margins motivate high-confidence anchoring, while formal certification additionally requires control of the iteration residual.

4.3 Confidence anchors and residual generalized-KL projection

Let $q_{i,(1)} \ge q_{i,(2)}$ denote the largest two entries of the current posterior row i, $\delta_i = q_{i,(1)} - q_{i,(2)}$, and $n_c(Q) = |\{i : \arg\max_j q_{ij} = c\}|$. For anchor fraction $f_A \in [0,1)$ and feasibility cap $\beta \in [0,1)$, the largest-margin samples in cluster c are selected with

$$a_c = \min\{\lfloor f_A n_c(Q)\rfloor,\ \lfloor \beta n \pi_{\rho,c}\rfloor\}, \qquad f_A, \beta \in [0,1). \tag{8}$$

The reported implementation searches $f_A \in \{0, 0.15\}$ and fixes $\beta = \frac{1}{2}$, which conservatively leaves at least half of each target mass available to the residual projection before integer effects. Let A_c contain these selected anchors and let $F = [n] \setminus \bigcup_c A_c$ be the free set. The remaining $m = |F| = n - \sum_c a_c$ rows carry residual masses

$$c_c = n\pi_{\rho,c} - a_c, \qquad \sum_c c_c = m.$$

For propagated free-row scores R_F, temperature $T > 0$, and numerical floor $\varepsilon_0 > 0$, define

$$\mathsf{K}_{ic} = (R_{F,ic} + \varepsilon_0)^{1/T} > 0.$$

Define the generalized KL divergence

$$D_{\mathrm{GKL}}(U\|\mathsf{K}) = \sum_{i,c}\left[U_{ic}\log\frac{U_{ic}}{\mathsf{K}_{ic}} - U_{ic} + \mathsf{K}_{ic}\right].$$

We use the standard convention $0\log(0/a) = 0$ for $a > 0$. The residual generalized-KL projection is the finite-dimensional I-projection of K onto the intersection of row- and column-marginal constraints (Csiszár, 1975; Benamou et al., 2015). Its SCoRP-specific feature is that the column marginal is the *residual* mass left after confidence anchors:

$$U^\star = \arg\min_{U \ge 0} D_{\mathrm{GKL}}(U\|\mathsf{K}) \quad \text{s.t.} \quad U\mathbf{1} = \mathbf{1}_m, \quad U^\top\mathbf{1} = c. \tag{9}$$

Theorem 2 (Unique residual projection and strictly convex reduced dual) *Assume $\pi_\rho > 0$ componentwise and Eq.* (8) *holds with $\beta < 1$. Then $c_c > 0$ for every c, the feasible polytope contains a strictly positive point, and Eq.* (9) *has a unique strictly positive minimizer. Moreover, ϕ has a unique minimizer $v^\star$ on $\mathbf{1}^\perp$; every unconstrained minimizer is $v^\star + \tau\mathbf{1}$, $\tau \in \mathbb{R}$, and*

$$U^\star_{ic} = \frac{\mathsf{K}_{ic}e^{v_c^\star}}{\sum_{j=1}^K \mathsf{K}_{ij}e^{v_j^\star}}. \tag{10}$$

Here

$$\phi(v) = \sum_{i=1}^m \log\left(\sum_{c=1}^K \mathsf{K}_{ic}e^{v_c}\right) - c^\top v, \tag{11}$$

with $\nabla\phi(v) = U(v)^\top\mathbf{1} - c$, where $U(v)$ is the row-normalized form in Eq. (10)*. The Hessian is positive definite on $\mathbf{1}^\perp$, so the reduced $(K-1)$-dimensional dual is strictly convex.*

The reduced dual permits optimization over only the cluster-level variables with L-BFGS, followed by direct verification of row and column residuals. It also makes the projection block analytically distinct from a generic heuristic reweighting step.

Corollary 2 (KL-minimal soft intervention) *Let $\mathcal{C} = \{U \geq 0 : U\mathbf{1} = \mathbf{1}_m,\ U^\top \mathbf{1} = c\}$ be the residual-marginal constraint polytope and use generalized KL divergence. For every $V \in \mathcal{C}$,*

$$D_{\mathrm{GKL}}(V\|\mathsf{K}) = D_{\mathrm{GKL}}(V\|U^\star) + D_{\mathrm{GKL}}(U^\star\|\mathsf{K}). \tag{12}$$

Consequently, $U^\star$ is the unique feasible posterior that changes the propagated positive kernel as little as possible in KL geometry.

The term *Pythagorean* refers to the exact divergence decomposition in Eq. (12): the divergence from any feasible V to K splits into a projection term through $U^\star$ plus the residual divergence from $U^\star$ to K, analogous to the distance decomposition of an orthogonal Euclidean projection. This is the anchored-residual specialization of classical I-projection geometry (Csiszár, 1975). It is the soft counterpart of Theorem 3: SCoRP first makes a minimum generalized-KL intervention and then a minimum-count hard repair.

Corollary 3 (Exact soft mass and non-collapse) *After every projection round, the combined anchored/free posterior satisfies*

$$Q^\top \mathbf{1} = n\pi_\rho.$$

Hence for $\rho > 0$, every cluster receives soft mass at least $n\rho/K$.

4.4 Minimum-Hamming feasibility decoding

Soft non-collapse does not imply hard coverage. Let

$$b = \min\Big\{\max(1, \lfloor \eta n \rfloor), \Big\lfloor \frac{n}{K} \Big\rfloor\Big\}.$$

Starting from $y^{(0)}$ with $y_i^{(0)} = \arg\max_c q_{ic}$, let $n_c(y) = |\{i : y_i = c\}|$. Call a cluster *deficient* if $n_c(y) < b$, and call a sample i an *admissible donor* when $n_{y_i}(y) > b$. For an admissible move $i \to c$, define the posterior-cost increase $\zeta_{i\to c} = -\log q_{ic} + \log q_{iy_i}$. At each step, choose a currently deficient cluster $c^\star$. The minimum-Hamming guarantee does not require $c^\star$ to be the uniquely or maximally deficient cluster; indeed, equal deficits are common when several clusters are empty. The reference implementation processes deficient clusters in a fixed order. Among the admissible donors for the selected $c^\star$, choose $i^\star \in \arg\min_i \zeta_{i\to c^\star}$, resolving donor ties by the fixed sample order, and update $y_{i^\star} \leftarrow c^\star$. Repeat until every cluster has count at least b.

Theorem 3 (Minimum-Hamming feasible repair) *Assume $Kb \leq n$, let $n_c^{(0)}$ be the initial hard counts, and define*

$$\Delta_0 = \sum_{c=1}^{K} (b - n_c^{(0)})_+.$$

The decoder terminates after exactly Δ_0 reassignments and returns $n_c \geq b$ for every cluster. No feasible hard partition can differ from the initial $\arg\max$ *partition on fewer than Δ_0 samples. Thus the decoder is a minimum-Hamming exact-feasibility repair. For whichever deficient cluster is selected at a given step, the decoder minimizes the instantaneous posterior-cost increase $\zeta_{i\to c}$ over its admissible donor moves; no "maximally deficient" condition is required for the Hamming-optimality claim.*

This is stronger than merely guaranteeing termination: the decoder changes the smallest possible number of initial hard decisions, although it does not claim globally minimum posterior cost among all repairs with that Hamming distance.

Corollary 4 (End-to-end partition feasibility) *Under the assumptions of Theorems 1 and 2, every SCoRP round remains in the row simplex and has column mass exactly $n\pi_\rho$. If $Kb \leq n$, the returned hard partition contains all K requested clusters; if its unconstrained* $\arg\max$ *partition is already feasible, the decoder is the identity.*

4.5 Permutation equivariance

Theorem 4 (Label-permutation equivariance) *Let Π be a $K \times K$ permutation matrix. If the aligned seed posterior and cluster-mass target are transformed as $Q^{(0)} \mapsto Q^{(0)}\Pi$ and $\pi_\rho \mapsto \Pi^\top \pi_\rho$, then graph propagation, anchor selection, and residual projection commute with the same permutation. For the hard-repair stage, the* set *of admissible minimum-Hamming repairs is permutation-equivariant. The point-valued decoder itself commutes with Π only when each deficient-cluster choice is uniquely determined by a permutation-invariant criterion and each donor argmin is unique (or ties are resolved by a label-symmetric rule). If several deficient clusters remain admissible and the implementation resolves that choice by cluster index, a different representative of the same equivariant repair set may be returned after relabeling; pointwise equivariance should therefore not be claimed in that case.*

This property is fundamental for clustering: the method depends on the partition, not on arbitrary integer names assigned to clusters.

4.6 Blind model selection and optional bounded CRAFT

Candidates use label-free diagnostics only. Let $r_j \in [0,1]$ denote percentile-ranked internal diagnostics and let nonnegative selector weights sum to one; screening and final selection are weighted sums, so the weights are design choices rather than theoretical constants. For the reported production runs, screening over agreement, silhouette, separation, balance, and margin used $(.25, .25, .20, .15, .15)$, while final ranking over stability, agreement, silhouette, balance, margin, and separation used $(.35, .20, .15, .10, .10, .10)$. The top six candidates are rerun at graph scales $\{0.75, 1, 1.25\}$, with stability intentionally dominant in the final score because graph-perturbation robustness is the primary selection criterion. The production search uses $\alpha \in \{0.95, 0.985\}$, $T \in \{0.10, 0.20, 0.35\}$, $f_A \in \{0, 0.15\}$, and a deterministic adaptive-ρ grid; $\beta = 1/2$ and $\eta = 10^{-3}$ are fixed. Failed stability reruns are ineligible and score ties are broken by candidate ID. Appendix C gives the component formulas, normalizations, exact fixed weights, adaptive-ρ rule, and tie rules. ACC/NMI/ARI are isolated until predictions are frozen.

CRAFT is an optional module introduced here, not an external clustering algorithm. It learns a bounded residual adapter from the selected SCoRP posterior,

$$z_i = \mathrm{norm}(x_i + r_\theta(x_i)), \qquad \|r_\theta(x_i)\|_2 \le \varepsilon < 1.$$

The final method compares the selected SCoRP output with one refinement branch. CRAFT is trained from the selected soft posterior; adapted features rebuild the graph, after which that posterior is propagated and reprojected with the selected (α, T, ρ, f_A) settings before feasibility decoding. The seed bank and consensus are not regenerated. Branch comparison is a nonnegative convex combination of predefined label-free diagnostics; the reported runs use $(.40, .25, .15, .10, .10)$ for worst-view stability, seed agreement, scaled worst-view silhouette, bounded worst-view separation, and target-mass fit. The refined branch must strictly improve this score and pass fixed stability/silhouette gates; otherwise the base branch is retained. CRAFT followed by k-means remains an ablation only.

Theorem 5 (CRAFT geometry preservation) *If $\|x_i\| = 1$, $\|r_i\| \le \varepsilon < 1$, and $z_i = \mathrm{norm}(x_i + r_i)$, then*

$$\|z_i - x_i\| \le 2\varepsilon, \qquad |z_i^\top z_j - x_i^\top x_j| \le 4\varepsilon.$$

If every point's cosine-similarity gap between its kth and $(k+1)$th neighbors exceeds 8ε, its kNN set is unchanged.

Complexity. Condition on the sparse graph and seed partitions being available. With $O(nk)$ graph edges, each propagation iteration costs $O(nkK)$, so M iterations cost $O(MnkK)$. Computing propagated $\arg\max$ labels and margins costs $O(nK)$, and sorting margins within clusters costs at most $O(n \log n)$. For m free samples, each reduced-dual objective/gradient evaluation costs $O(mK)$; L L-BFGS evaluations therefore cost $O(LmK)$. Initial hard decoding costs $O(nK)$. A direct implementation that scans all samples for each of the Δ_0 repairs costs $O(\Delta_0 n)$, although indexed candidate lists can reduce this term. Thus one candidate costs $O(MnkK + LmK + nK + n \log n + \Delta_0 n)$, excluding graph construction, seed generation, and optional CRAFT training. Building the S seed contingency tables costs $O(Sn)$, their Hungarian alignments cost $O(SK^3)$, and the mandatory propagation/projection stages use $O(nk + nK)$ memory without forming a dense $n \times n$ affinity matrix.

Algorithm 1 SCoRP with optional bounded CRAFT

Require: Unlabeled features X, requested K, fixed seed.

Build sparse cosine graph and diverse noncollapsed seed partitions.
Align seed labels; form soft consensus $Q^{(0)}$ and adaptive mass target π_ρ.
for each label-free candidate (α, T, ρ, f_A) **do**
    Diffuse the current posterior by Eq. (6).
    Clamp only high-margin anchors subject to Eq. (8).
    Solve the residual generalized-KL projection in Eq. (9).
    Decode with the minimum-Hamming feasibility repair.
    Score with label-free consensus/geometry/stability diagnostics.
end for
Select by cross-graph stability-weighted score.
Optionally train bounded CRAFT from the selected posterior; rebuild the graph and rerun propagation–projection–decode with the selected settings; retain the refined branch only if the predefined label-free stability criteria improve.

return Soft posterior and feasible hard partition.

5 Experiments

5.1 Protocol and evaluation philosophy

The benchmark is split by role rather than pooled indiscriminately. Nine light datasets informed method development; the subsequent five-dataset vision suite used frozen DINOv2-S/14 feature matrices and the carried-forward protocol; CIFAR-10, CIFAR-20 (CIFAR-100 coarse labels), CIFAR-100, and LEDGAR were evaluated later as *post-development* definitions. We use this conservative term rather than relying on a stronger holdout claim. CIFAR-20 and CIFAR-100 share the same 10,000 DINOv2 feature vectors and differ only in label granularity. Within every matched comparison all methods receive the identical saved feature matrix, K comes from benchmark metadata, and labels are isolated until predictions are frozen. Appendix C records the versioned protocol chronology; Appendix B gives sizes and preprocessing.

The suite intentionally spans faces, object views, handwritten digits, fashion, natural images, fine-grained pets and flowers, remote-sensing scenes, news text, legal clauses, tabular chemistry, and wearable-sensor activity. Public sources include COIL object views (Nene et al., 1996), Extended Yale B illumination variation (Georghiades et al., 2001), Fashion-MNIST (Xiao et al., 2017), UCI HAR (Anguita et al., 2013), STL-10 (Coates et al., 2011), Oxford-IIIT Pets (Parkhi et al., 2012), Oxford Flowers (Nilsback and Zisserman, 2008), EuroSAT (Helber et al., 2019), CIFAR (Krizhevsky, 2009), and LEDGAR (Tuggener et al., 2020).

Metrics are clustering accuracy (ACC) after optimal label matching, normalized mutual information (NMI), and adjusted Rand index (ARI). Standard baselines include k-means, normalized k-means, MiniBatch k-means, diagonal GMM, Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH), agglomerative clustering when feasible, sparse spectral clustering, FINCH, and NMF+k-means. Matched structural comparators include Graph-SymNMF, local self-expression plus spectral clustering, and graph-continuation spectral clustering. Probabilistic graph-cut (PGCuts) variants are audited separately because a probabilistic cut can return fewer occupied clusters than the benchmark-provided K. Modern TURTLE and PRO-DSC comparisons are protocol-labeled rather than merged with matched-feature rows when their native representation differs.

5.2 Results

Table 1 summarizes representative matched results; Appendix D reports all 18 datasets, including every unfavorable row.

CIFAR-10 gains +3.91/+4.05/+8.64 ACC/NMI/ARI points, CIFAR-100 gains +3.15 ACC and +5.67 ARI, and COIL-20 gains +8.61 ACC and +13.91 ARI. CIFAR-20 is near-tied in ACC; Pets, Flowers-102, Wine, and LEDGAR decline on all three metrics; and UCI HAR trades higher ACC for lower NMI/ARI.

The selector improves ACC on 11/18 definitions and at least one metric on 12/18, but these development/evaluation roles are not exchangeable and the counts are not estimates of future-dataset success. CRAFT is selected on COIL-20,

Table 1: **Selected matched-feature results (%).** Final scores with Δ pp relative to the metric-wise strongest standard same-feature baseline. Appendix D reports all 18 rows; Appendix C specifies the ledger audit that records the maximizing baseline name separately for ACC, NMI, and ARI.

Dataset	ACC (Δ pp)	NMI (Δ pp)	ARI (Δ pp)
COIL-20	82.57 (+8.61)	87.24 (+4.56)	78.68 (+13.91)
20 Newsgroups	44.97 (+1.46)	37.78 (-5.68)	25.38 (+2.81)
EuroSAT	68.05 (+1.64)	68.15 (-0.10)	57.75 (+4.35)
CIFAR-10	93.74 (+3.91)	88.21 (+4.05)	87.01 (+8.64)
CIFAR-20	62.61 (-0.43)	66.00 (-0.48)	48.35 (+0.22)
CIFAR-100	67.74 (+3.15)	76.82 (-0.85)	53.08 (+5.67)
LEDGAR	42.84 (-3.36)	64.23 (-1.25)	34.00 (-6.95)

20 Newsgroups, Oxford Pets, and STL-10; otherwise the base SCoRP branch is retained. CIFAR-20/100, which share features but differ in label granularity, further expose taxonomy sensitivity.

Modern matched comparator. Table 6 controls the representation by giving SCoRP and TURTLE the identical saved DINOv2-S/14 feature matrices. In the fixed-seed runs, SCoRP exceeds TURTLE by +15.81/+9.88/+16.78 ACC/NMI/ARI points on CIFAR-10 and +30.86/+15.81/+24.88 on CIFAR-100. TURTLE occupies 98/100 clusters on CIFAR-100 whereas SCoRP returns 100/100; this is consistent with the hard-feasibility motivation but does not attribute the full metric gap to the decoder. These values are not mean-performance estimates; the table and protocol details are retained in Appendix I to preserve the 9-page main-text limit.

6 Discussion and Limitations

The evidence supports a structural, not universal, claim: SCoRP controls stability, feasibility, and the size of the intervention conditional on the supplied unlabeled geometry, but it cannot guarantee recovery of an external semantic taxonomy. If the representation or seed consensus is misaligned with that taxonomy, graph propagation can stabilize the mismatch and the residual GKL projection can correct cluster masses without improving ACC/NMI/ARI. LEDGAR is a clear example because the source corpus is multi-label (Tuggener et al., 2020), whereas our audit asks for one hard 100-way partition of fixed sentence embeddings. Pets/Flowers show a related fine-grained geometry–taxonomy mismatch, while UCI HAR illustrates that improving one external metric need not improve the others. Appendix E gives the detailed interpretation.

The matched-feature TURTLE comparison confirms that the downstream partition stage can matter even when the representation is held fixed, but the current comparison uses one fixed seed and should not be read as a mean-performance claim. Other limitations are known K, heuristic ρ-grid centering, a sufficient rather than necessary margin certificate, positive residual-mass/kernel requirements, and a finite benchmark audit. The guarantees therefore constrain how SCoRP changes a supplied representation; they do not assert semantic identifiability.

7 Conclusion

SCoRP separates uncertainty preservation, graph propagation, cluster-mass reconciliation, and hard feasibility instead of forcing them into a single objective. Its guarantees are tied to these interfaces: perturbation-to-label stability, a unique residual GKL projection, minimum-Hamming repair, set-valued permutation equivariance with a conditional pointwise guarantee, and optional bounded CRAFT. Across the 18-dataset audit, gains and failures are both retained, while the matched-feature TURTLE comparison isolates the downstream partition stage from the representation. The evidence supports SCoRP as a controlled post-representation operator when the supplied geometry is informative, rather than as a replacement for representation learning.

Reproducibility Statement

Appendices C, B, D, F, and I document the exact label-free selector, chronology, datasets/preprocessing, complete audit, proofs, and comparator protocols. The accompanying code preserves the exact production implementation used for the reported SCoRP/CRAFT experiments and includes the selector configuration, saved-output checks, and reproducibility/audit routines. Legacy internal identifiers are retained only where needed to reproduce archived runs exactly. Ground-truth labels are read only for post-freeze evaluation metrics.

Ethics Statement

The study uses public benchmark datasets and does not involve intervention, private records, or deployment decisions. Clusters can inherit biases or artifacts from their input representation; their identities should not be interpreted as socially or scientifically meaningful categories without domain validation.

AI Use Statement

Generative-AI tools were used for literature discovery, language editing, code assistance/debugging, methodological and experimental-design review, and mathematical/proof checking. The authors independently reviewed the cited sources, verified the mathematical arguments, executed the reported experiments, checked the resulting tables and figures, and made the final methodological and writing decisions. The authors take responsibility for all claims, citations, code, results, and text in the submission.

References

J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1967.

Anil K. Jain. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE TPAMI*, 22(8):888–905, 2000.

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *NeurIPS*, 2001.

Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

Ulrike von Luxburg, Mikhail Belkin, and Olivier Bousquet. Consistency of spectral clustering. *The Annals of Statistics*, 36(2):555–586, 2008.

Fan R. K. Chung. *Spectral Graph Theory*. American Mathematical Society, 1997.

Alexander Strehl and Joydeep Ghosh. Cluster ensembles—A knowledge reuse framework for combining multiple partitions. *JMLR*, 3:583–617, 2002.

Ana L. N. Fred and Anil K. Jain. Combining multiple clusterings using evidence accumulation. *IEEE TPAMI*, 27(6):835–850, 2005.

Alexander Topchy, Anil K. Jain, and William Punch. Clustering ensembles: Models of consensus and weak partitions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1866–1881, 2005.

Tossapon Boongoen and Natthakan Iam-On. Cluster ensembles: A survey of approaches with recent extensions and applications. *Computer Science Review*, 28:1–25, 2018.

Yaser Esmaeili Salehani, Ehsan Arabnejad, and Mohamed Cheriet. Graph and sparse-based robust nonnegative block value decomposition for clustering. *IEEE Journal of Selected Topics in Signal Processing*, 12(6):1561–1574, 2018.

Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

Johannes Gasteiger, Aleksandar Bojchevski, and Stephan Günnemann. Predict then propagate: Graph neural networks meet personalized PageRank. In *ICLR*, 2019.

Feiyang Chen, Ruiqiang Guo, and Zhibin Gu. Multi-scale diffusion-guided graph learning with power-smoothing random walk contrast for multi-view clustering. In *ICLR*, 2026.

Pan Peng and Yuichi Yoshida. Average sensitivity of spectral clustering. In *KDD*, 2020.

Eleonora Andreotti, Dominik Edelmann, Nicola Guglielmi, and Christian Lubich. Measuring the stability of spectral clustering. *Linear Algebra and its Applications*, 610:673–697, 2021.

Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975.

Joel Franklin and Jens Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and its Applications*, 114–115:717–735, 1989.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, 2013.

Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019.

Mikko I. Malinen and Pasi Fränti. Balanced K-means for clustering. In *Structural, Syntactic, and Statistical Pattern Recognition*, LNCS 8621, pages 32–41. Springer, 2014.

Gonzalo Mena. On model-based clustering with entropic optimal transport. *arXiv:2605.03240*, 2026.

Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *ICML*, 2016.

Ehsan Elhamifar and René Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE TPAMI*, 35(11):2765–2781, 2013.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, 2018.

Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. SCAN: Learning to classify images without labels. In *ECCV*, 2020.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020.

Maxime Oquab et al. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024.

Artyom Gadetsky, Yulun Jiang, and Maria Brbic. Let go of your labels with unsupervised transfer. In *ICML*, 2024.

Chuyu Zhang, Hui Ren, and Xuming He. P^2OT: Progressive partial optimal transport for deep imbalanced clustering. In *ICLR*, 2024.

Xuqian Xue, Yiming Lei, Qi Cai, Hongming Shan, and Junping Zhang. PROTOCOL: Partial optimal transport-enhanced contrastive learning for imbalanced multi-view clustering. In *ICML*, 2025.

Daniel Shalam and Simon Korman. The balanced-pairwise-affinities feature transform. In *ICML*, 2024.

Xianghan Meng, Zhiyuan Huang, Wei He, Xianbiao Qi, Rong Xiao, and Chun-Guang Li. Exploring a principled framework for deep subspace clustering. In *ICLR*, 2025.

Shaden Alshammari, John R. Hershey, Axel Feldmann, William T. Freeman, and Mark Hamilton. I-Con: A unifying framework for representation learning. In *ICLR*, 2025.

Ayoub Ghriss and Claire Monteleoni. Deep clustering via probabilistic ratio-cut optimization. In *AISTATS*, 2025.

Ayoub Ghriss. Beyond spectral clustering: Probabilistic cuts for differentiable graph partitioning. In *AISTATS*, 2026.

Zhixin Li, Yuheng Jia, Guanliang Chen, Hui Liu, and Junhui Hou. Mini-cluster guided long-tailed deep clustering. In *ICLR*, 2026.

Harold W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955.

Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

M. Saquib Sarfraz, Vivek Sharma, and Rainer Stiefelhagen. Efficient parameter-free clustering using first neighbor relations. In *CVPR*, 2019.

Sameer A. Nene, Shree K. Nayar, and Hiroshi Murase. Columbia object image library (COIL-20/COIL-100). Technical Report CUCS-005-96, Columbia University, 1996.

Athinodoros S. Georghiades, Peter N. Belhumeur, and David J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE TPAMI*, 23(6):643–660, 2001.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. arXiv:1708.07747, 2017.

Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge L. Reyes-Ortiz. A public domain dataset for human activity recognition using smartphones. In *ESANN*, 2013.

Adam Coates, Andrew Y. Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS*, 2011.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, 2012.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE JSTARS*, 12(7):2217–2226, 2019.

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

Don Tuggener, Pius von Däniken, Thomas Peetz, and Mark Cieliebak. LEDGAR: A large-scale multi-label corpus for text classification of legal provisions in contracts. In *LREC*, 2020.

A Theoretical foundations and what SCoRP adds

Block	Established foundation	SCoRP-specific conclusion
Graph propagation	Markov diffusion, spectral/random-walk geometry, PPR (Coifman and Lafon, 2006; von Luxburg, 2007; Gasteiger et al., 2019)	Joint graph–posterior perturbation bound and a top-two-margin certificate for hard-label invariance.
Residual GKL projection	I-projection, matrix scaling, entropic/Bregman projections (Csiszár, 1975; Franklin and Lorenz, 1989; Benamou et al., 2015)	Anchor-residual feasibility, a strictly convex $(K-1)$-dimensional reduced dual, and an exact generalized-KL projection identity.
Mass control	Balanced/size-constrained clustering (Malinen and Fränti, 2014)	Continuous seed-to-uniform mass path with explicit per-cluster floor and KL/entropy contraction.
Hard feasibility	Constrained cluster support	Feasibility repair with the minimum Hamming distance to the unconstrained $\arg\max$ partition while satisfying the required count floor.
Bounded refinement	Perturbation continuity of normalized features	Explicit cosine-similarity drift and a sufficient top-k neighborhood-preservation condition for CRAFT.

Table 2: Established foundations and results specific to the SCoRP composition. The right column lists the guarantees proved for the interfaces used in this paper; it does not present the underlying classical components as new.

B Dataset details

Dataset	n	K	d	Feature protocol
ORL	400	40	66	pixels → PCA-whiten → ℓ_2 norm
Extended Yale B	2,432	38	160	pixels → PCA-whiten → ℓ_2 norm
COIL-20	1,440	20	100	32×32 gray → PCA → norm
Fashion-MNIST	10,000	10	128	test pixels → PCA-whiten → norm
MNIST	10,000	10	128	test pixels → PCA-whiten → norm
Digits	1,797	10	64	normalized raw 8×8 pixels
Wine	178	3	13	standardized tabular features → norm
20 Newsgroups	7,532	20	128	test text → TF–IDF → SVD → norm
UCI HAR	10,299	6	128	train+test 561-D → standardize/PCA/norm
COIL-100	7,200	100	384	DINOv2-S/14
STL-10	13,000	10	384	DINOv2-S/14
Oxford Pets	7,349	37	384	DINOv2-S/14
Flowers-102	8,189	102	384	DINOv2-S/14
EuroSAT	27,000	10	384	DINOv2-S/14
CIFAR-10	10,000	10	384	test split, DINOv2-S/14
CIFAR-20	10,000	20	384	CIFAR-100 test split, DINOv2-S/14, coarse labels
CIFAR-100	10,000	100	384	test split, DINOv2-S/14
LEDGAR	10,000	100	384	test legal clauses, all-MiniLM-L6-v2

Table 3: Complete dataset and representation audit. The first nine rows are development datasets; COIL-100 through EuroSAT form the frozen-feature evaluation group; and CIFAR-10 through LEDGAR are post-development evaluation definitions. CIFAR-20 and CIFAR-100 intentionally share the same image feature matrix and differ only in label granularity.

C Exact label-free selection protocol

Candidate diagnostics and normalization. The seed bank first removes collapsed seeds and scores the remaining hard partitions with the same label-free geometry family used for screening. For a seed h, mean agreement with the other valid seeds, cosine silhouette, normalized hard-count entropy, and centroid separation are percentile ranked; the hard-seed posterior-margin field is constant and therefore does not affect ordering. The top three seeds are retained, aligned by Hungarian matching to the highest-scoring seed, and assigned weights proportional to $\exp(5(s_s - \max_r s_r))$ before normalization, exactly as in the production notebook.

For a candidate posterior Q and its repaired hard partition $\hat{y}$, define

$$C = \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}} \mathrm{ARI}(\hat{y}, h_s), \qquad B = -\frac{1}{\log K}\sum_{c=1}^{K} p_c \log p_c,$$

where h_s are the independent seed partitions and $p_c = n_c(\hat{y})/n$. The silhouette S uses cosine distance on the fixed

feature matrix and $\hat{y}$; centroid separation is $D = 1 - \max_{c \neq d} \bar{\mu}_c^\top \bar{\mu}_d$ for normalized cluster centroids of $\hat{y}$; and posterior margin is $M = n^{-1} \sum_i (q_{i,(1)} - q_{i,(2)})$, computed from the soft Q. Thus C, S, B, D are hard-partition diagnostics, M is a soft-posterior diagnostic, and cross-graph stability below is computed from repaired hard partitions. Larger is better for every component. There is no additional target-fit term in the base candidate selector: B is its mass-regularity component; target-mass fit appears only in the final CRAFT branch selector. Within the candidate pool, the implementation converts each quantity to an average-tie percentile rank. Screening uses

$$S_{\mathrm{scr}} = .25r_C + .25r_S + .20r_D + .15r_B + .15r_M.$$

The top six candidates are rerun on graph scales $0.75, 1, 1.25$ times the base k. Stability A is the mean pairwise ARI of the resulting hard partitions; after percentile ranking among valid shortlisted candidates,

$$S_{\mathrm{sel}} = .35r_A + .20r_C + .15r_S + .10r_B + .10r_M + .10r_D.$$

Candidates with failed stability reruns are ineligible. The final sort is descending in S_{sel} and ascending in deterministic candidate ID, which is the tie rule.

Search grid and adaptive mass target. The production search uses $\alpha \in \{0.95, 0.985\}$, $T \in \{0.10, 0.20, 0.35\}$, $f_A \in \{0, 0.15\}$, three propagation–projection rounds, and 30 fixed-point diffusion steps per round. The graph-scale set is $\{0.75, 1, 1.25\}$; the decoder uses fixed $\eta = 10^{-3}$, and the anchor feasibility cap uses fixed $\beta = 1/2$. These two quantities are therefore *not* tuned grids. With seed-consensus mass π_0, normalized entropy $H_m = H(\pi_0)/\log K$, and D_m the mean pairwise Jensen–Shannon divergence between sorted seed-mass vectors, the code sets

$$\rho_\star = \begin{cases} 0, & H_m > 0.985 \text{ and } D_m < 0.002, \\ \mathrm{clip}_{[0,0.95]}\big(0.05 + 4(1 - H_m) + 3\sqrt{D_m}\big), & \text{otherwise}, \end{cases}$$

and searches the clipped unique values $\rho_\star - 0.20, \rho_\star, \rho_\star + 0.20$, adding 0 when $\rho_\star < 0.15$ and 0.90 when $\rho_\star > 0.65$.

CRAFT branch selector. The base and CRAFT→re-SCoRP branches are evaluated by the same hard-partition diagnostics. Let $A_{\min}$ be worst-view graph-diffusion stability over original/adapted features, $S_{\min}$ and $D_{\min}$ the corresponding worst-view silhouette and separation, C seed agreement on the original features, and $F = 1 - \mathrm{JS}(p(\hat{y}), \pi_\rho)/\log 2$ the hard-mass target fit. The common score is

$$S_{\mathrm{br}} = .40A_{\min} + .25C + .15\ \mathrm{clip}_{[0,1]}((S_{\min} + 1)/2) + .10(1 - e^{-\max(D_{\min}, 0)}) + .10F.$$

CRAFT is accepted only if its score strictly improves, $A_{\min} \geq 0.90$, its stability drop relative to the base is at most 0.08, and its silhouette drop is at most 0.01. Otherwise the base branch is returned. CRAFT-only k-means is never eligible.

Chronology and terminology. The archived notebooks identify the nine light datasets as the development suite, record the medium DINOv2 protocol as frozen from the light-stage method, and carry the same SCoRP search/selector settings into the later CIFAR/LEDGAR benchmark notebook. Because "holdout" can imply stronger temporal isolation than is necessary for the claims here, the manuscript conservatively calls the latter four *post-development evaluations*. Labels remain separated from fitting and blind candidate/branch selection throughout.

D Complete result audit

	ACC		NMI		ARI	
Dataset	Final	Δ	Final	Δ	Final	Δ
20 Newsgroups	44.97	+1.46	37.78	-5.68	25.38	+2.81
CIFAR-10	93.74	+3.91	88.21	+4.05	87.01	+8.64
CIFAR-100	67.74	+3.15	76.82	-0.85	53.08	+5.67
CIFAR-20	62.61	-0.43	66.00	-0.48	48.35	+0.22
COIL-100	80.29	+0.50	95.54	+1.04	80.68	+0.93
COIL-20	82.57	+8.61	87.24	+4.56	78.68	+13.91
Digits	88.81	-0.33	87.39	-0.69	82.51	-0.33
EuroSAT	68.05	+1.64	68.15	-0.10	57.75	+4.35
Extended Yale B	89.51	+2.01	90.21	+0.99	80.83	+7.97
Fashion-MNIST	60.44	+3.61	62.30	-2.02	47.03	+1.96
Flowers-102	94.04	-2.77	98.11	-0.89	92.87	-2.61
LEDGAR	42.84	-3.36	64.23	-1.25	34.00	-6.95
MNIST	68.70	+0.52	64.47	+1.57	54.06	+1.26
ORL	79.00	+2.00	88.68	-0.01	71.44	+3.71
Oxford Pets	82.79	-3.21	90.56	-0.44	78.72	-2.60
STL-10	99.08	-0.13	97.61	-0.35	97.97	-0.29
UCI HAR	60.98	+0.58	51.63	-9.17	39.50	-7.43
Wine	95.51	-1.12	84.84	-1.77	86.66	-3.09

Table 4: Complete matched-feature audit (%). Δ is the final score minus the best score among the predefined standard same-feature baselines for that dataset and metric. The maximizing baseline may differ across ACC, NMI, and ARI. Positive and negative values denote gains and losses in percentage points.

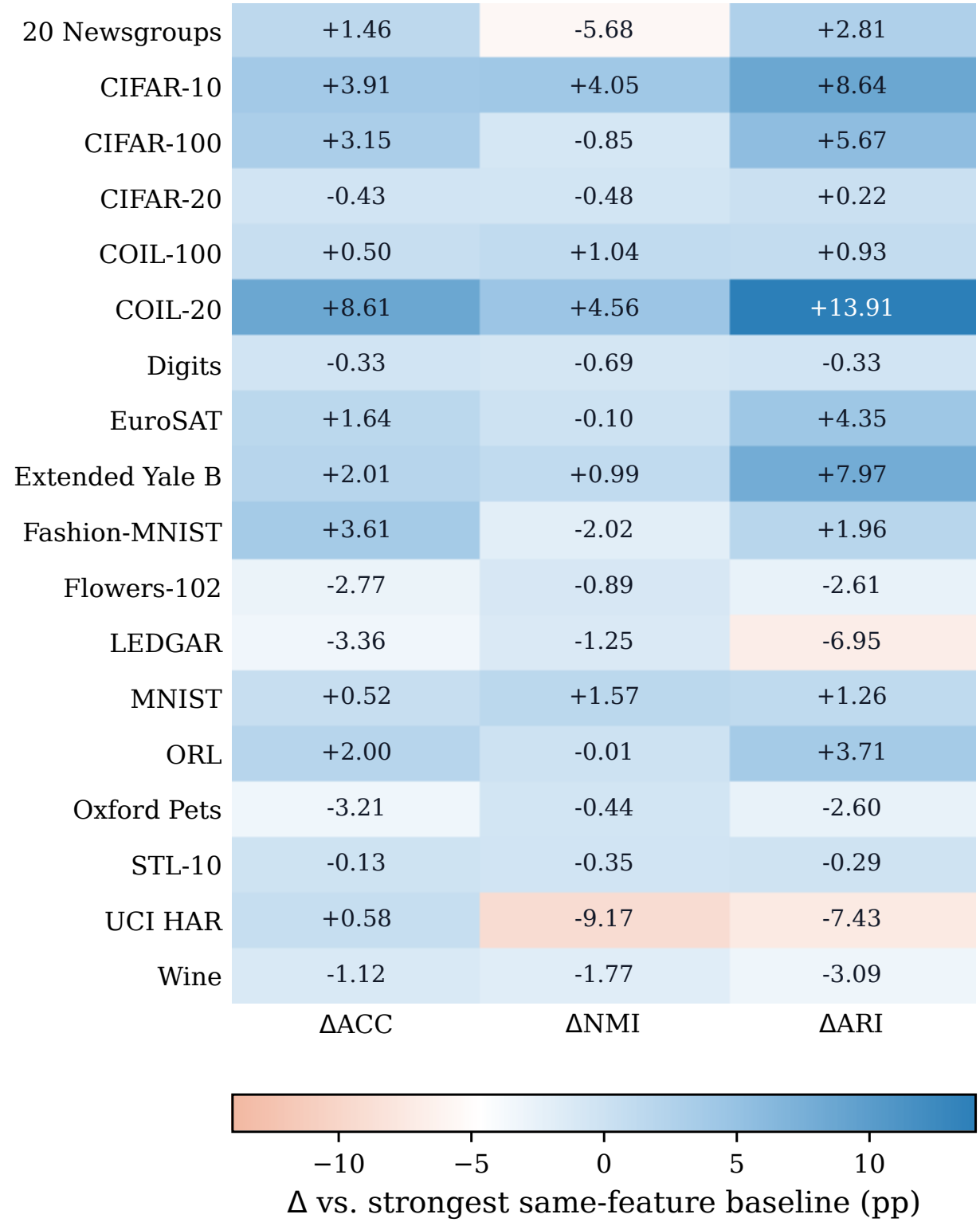

Figure 2: **Complete signed metric audit.** Cells show Final minus the strongest standard same-feature baseline in percentage points. The common zero-centered scale exposes gains, losses, and metric trade-offs without filtering datasets by outcome.

E Failure-case interpretation

The unfavorable rows identify a limitation of the representation-to-partition interface rather than a contradiction of the guarantees. The theory is deliberately conditional: it controls how much the posterior can move under graph perturbations, how little the residual projection changes a positive propagated kernel while meeting the requested masses, and how few hard labels must change to restore cluster support. None of these statements identifies which external semantic taxonomy the representation ought to encode. Thus, when a fixed representation places semantically different benchmark classes in the same neighborhoods—or separates examples that share a benchmark label—SCoRP can faithfully preserve and stabilize the wrong geometry from the viewpoint of that benchmark.

A one-step margin calculation makes this limitation explicit. For post-hoc interpretation only, let y denote an external reference class and c a competitor, and define $m_i = Q_{iy} - Q_{ic}$. The first diffusion update gives

$$m_i^+ = (1-\alpha)m_i + \alpha g_i, \qquad g_i = (PQ)_{iy} - (PQ)_{ic}.$$

If the local-neighbor contrast g_i is smaller than the original margin—and especially if $g_i < 0$—propagation reduces, and can reverse, the reference-class margin. Theorem 1 certifies stability of this operation to perturbations; it does not assert that g_i is aligned with unavailable semantic labels. Likewise, the residual projection block is the minimum generalized-KL correction needed to satisfy the chosen masses, not a supervised semantic correction. This separates the guarantees from a claim of universal benchmark improvement.

LEDGAR is the clearest example. The original corpus is a multi-label legal-clause resource (Tuggener et al., 2020), while the present audit evaluates one hard 100-cluster partition of fixed all-MiniLM-L6-v2 clause embeddings. Legal provisions can share vocabulary, drafting templates, and overlapping functions, so embedding neighborhoods need not correspond to a mutually exclusive 100-way taxonomy. Graph diffusion then propagates evidence along those neighborhoods; adaptive residual projection can repair global mass/support but cannot manufacture missing semantic separation. A similar issue can arise for fine-grained Oxford Pets and Flowers-102, where breed/species distinctions can be subtler than the dominant frozen-feature geometry. UCI HAR is different but equally informative: the positive ACC delta together with negative NMI/ARI deltas shows that a mass/stability correction can improve label-matched accuracy while changing the global partition structure in a way penalized by other metrics. These cases motivate an abstention or reliability diagnostic that detects weak agreement between seed consensus and graph geometry before applying corrective residual projection.

F Proofs of the strengthened theoretical results

F.1 Proof of Proposition 1

The norm identity follows from

$$\pi_\rho - u = (1-\rho)(\pi_0 - u).$$

The componentwise lower bound follows from $\pi_{\rho,c} = (1-\rho)\pi_{0,c} + \rho/K \geq \rho/K$. Convexity of KL in its first argument gives

$$D_{\mathrm{KL}}((1-\rho)\pi_0 + \rho u \| u) \leq (1-\rho) D_{\mathrm{KL}}(\pi_0 \| u) + \rho D_{\mathrm{KL}}(u \| u),$$

which proves Eq. (4). Since $D_{\mathrm{KL}}(p\|u) = \log K - H(p)$, entropy is nondecreasing. □

F.2 Proof of Theorem 1

Because P is row stochastic, $P^\ell \geq 0$, $P^\ell \mathbf{1} = \mathbf{1}$, and $\left\|P^\ell\right\|_\infty = 1$. The Neumann series therefore yields a nonnegative row-stochastic $G_\alpha(P)$ and $\|G_\alpha(P)\|_\infty = 1$. The fixed-point error satisfies

$$R_{m+1} - R^\star = \alpha P(R_m - R^\star),$$

giving geometric convergence.

For perturbation stability, add and subtract $G_\alpha(P)\widetilde{Q}$:

$$\left\|G_\alpha(P)Q - G_\alpha(\widetilde{P})\widetilde{Q}\right\|_\infty \leq \left\|Q - \widetilde{Q}\right\|_\infty + \left\|G_\alpha(P) - G_\alpha(\widetilde{P})\right\|_\infty \left\|\widetilde{Q}\right\|_\infty.$$

A soft partition has induced infinity norm one. The resolvent identity gives

$$G_\alpha(P) - G_\alpha(\widetilde{P}) = (1-\alpha)(I-\alpha P)^{-1}\alpha(P-\widetilde{P})(I-\alpha\widetilde{P})^{-1}.$$

Since $\left\|(I-\alpha P)^{-1}\right\|_\infty \le (1-\alpha)^{-1}$ for any row-stochastic P, the second term is bounded by $\frac{\alpha}{1-\alpha}\left\|P-\widetilde{P}\right\|_\infty$. □

F.3 Proof of Corollary 1

If every entry changes by at most δ, the winner in a row can decrease by at most δ while any competitor can increase by at most δ. A top-two gap greater than 2δ therefore cannot change sign. The entrywise maximum perturbation is bounded by the induced infinity norm. □

F.4 Proof of Theorem 2

The anchor cap yields

$$c_c = n\pi_{\rho,c} - a_c \ge (1-\beta)n\pi_{\rho,c} > 0, \qquad \sum_c c_c = m.$$

Hence a feasible point exists, e.g. every row equal to c/m. Because $\mathsf{K}_{ic} > 0$ and all row/column marginals are positive, the feasible polytope contains a strictly positive point. Generalized KL is strictly convex in U, so the constrained minimizer is unique. Since the directional derivative of $u \mapsto u\log(u/\mathsf{K}_{ic})$ tends to $-\infty$ as $u \downarrow 0$, an optimum cannot contain a zero entry when a positive feasible perturbation is available; hence the unique minimizer is strictly positive.

With row and column multipliers, stationarity gives $U_{ic} = a_i\mathsf{K}_{ic}b_c$ with positive factors. Writing $b_c = e^{v_c}$ and eliminating a_i using row normalization yields Eq. (10). Substitution gives Eq. (11), whose gradient is the vector of current column sums minus c. Its Hessian is

$$\nabla^2\phi(v) = \sum_i \left[\mathrm{Diag}(u_i) - u_i u_i^\top\right],$$

where u_i is the ith row of Eq. (10). Each u_i is strictly positive, so for any nonconstant vector z, $z^\top[\mathrm{Diag}(u_i) - u_i u_i^\top]z = \mathrm{Var}_{u_i}(z_c) > 0$. The only null direction is $\mathbf{1}$, corresponding to the gauge $v \mapsto v + a\mathbf{1}$. Thus the restriction to $\mathbf{1}^\perp$ is strictly convex. □

F.5 Proof of Corollary 2

Write generalized KL as $D_{\mathrm{GKL}}(A\|B) = \sum_{ic}[A_{ic}\log(A_{ic}/B_{ic}) - A_{ic} + B_{ic}]$. The KKT conditions for Theorem 2 give $\log(U^\star_{ic}/\mathsf{K}_{ic}) = r_i + s_c$ for row/column multipliers after absorbing an additive constant. For any feasible V, both $V - U^\star$ row sums and column sums vanish. Therefore

$$\sum_{ic}(V_{ic} - U^\star_{ic})\log\frac{U^\star_{ic}}{\mathsf{K}_{ic}} = \sum_i r_i \sum_c (V_{ic} - U^\star_{ic}) + \sum_c s_c \sum_i (V_{ic} - U^\star_{ic}) = 0.$$

Expanding $D_{\mathrm{GKL}}(V\|\mathsf{K}) - D_{\mathrm{GKL}}(V\|U^\star) - D_{\mathrm{GKL}}(U^\star\|\mathsf{K})$ leaves exactly this zero inner product, proving Eq. (12). Uniqueness follows from Theorem 2. □

F.6 Proof of Corollary 4

Theorem 1 preserves row stochasticity. Theorem 2 gives the exact target column mass after anchors are restored. Theorem 3 guarantees all hard counts are at least $b \ge 1$ and uses zero moves when the initial deficit is zero. □

F.7 Proof of Theorem 3

Let $\Delta = \sum_c (b - n_c)_+$. If some cluster is deficient and $Kb \le n$, another cluster must have count $> b$; otherwise the total count would be strictly less than n. A move from such a donor to a deficient cluster preserves donor feasibility and decreases Δ by exactly one. Thus the algorithm makes exactly Δ_0 moves.

Any feasible final assignment must supply at least $b - n_c^{(0)}$ new points to each initially deficient cluster c. Summed over those clusters, at least Δ_0 samples must change their initial labels. Hence no feasible repair can have smaller Hamming distance. The instantaneous posterior-cost statement follows directly from the implementation's donor minimization. □

F.8 Proof of Theorem 4

Right multiplication by Π relabels posterior columns, while graph propagation acts on rows and therefore commutes with Π. Moreover, $\pi_0' = n^{-1}(Q^{(0)}\Pi)^\top \mathbf{1}_n = \Pi^\top \pi_0$ and $\Pi^\top u = u$, hence $\pi_\rho' = \Pi^\top \pi_\rho$. Posterior margins are invariant to column permutations, and corresponding anchor sets are permuted consistently. The residual target therefore transforms as $c' = \Pi^\top c$, and the free-row kernel as $\mathsf{K}' = \mathsf{K}\Pi$. The matrix $U^\star\Pi$ has the required row marginal and column marginal $\Pi^\top c$. Since generalized KL divergence is invariant under a common column permutation, uniqueness in Theorem 2 gives $(U^\star)' = U^\star\Pi$. Decoder deficits and move costs are likewise permuted. Thus every admissible repair sequence for the original labeling maps to an admissible sequence for the permuted labeling, so the *set* of minimum-Hamming repairs is equivariant. If the deficient cluster at every step is uniquely selected by a permutation-invariant criterion and the donor minimizer is unique (or tied choices use a label-symmetric rule), the correspondence is point-valued and every decoder update commutes with Π. A fixed cluster-index choice among several deficient clusters does not satisfy this premise, so in that case only the set-valued statement is guaranteed. □

F.9 Proof of Theorem 5

Let $u = x + r$ and $z = u/\|u\|$. Then

$$\|z - x\| \le \|u/\|u\| - u\| + \|r\| = |1 - \|u\|\,| + \|r\| \le 2\|r\| \le 2\varepsilon.$$

For unit vectors,

$$|z_i^\top z_j - x_i^\top x_j| \le \|z_i - x_i\| + \|z_j - x_j\| \le 4\varepsilon.$$

A neighbor similarity can decrease by at most 4ε and a non-neighbor similarity can increase by at most 4ε; a boundary gap larger than 8ε cannot swap ordering. □

G Development-suite details

The light-suite matched ACC table is retained for transparency but should not carry the main generalization claim because these datasets informed method development.

Dataset	Best standard	Base ACC	Final ACC	Δ
ORL	Spectral-kNN	77.00	79.00	+2.00
Extended Yale B	Spectral-kNN	87.50	89.51	+2.01
COIL-20	FINCH	73.96	82.57	+8.61
Fashion-MNIST	FINCH	56.83	60.44	+3.61
MNIST	FINCH	68.18	68.70	+0.52
Digits	Spectral-kNN	89.14	88.81	-0.33
Wine	diagonal GMM	96.63	95.51	-1.12
20 Newsgroups	KMeans	43.51	44.97	+1.46
UCI HAR	BIRCH	60.40	60.98	+0.58

Table 5: Development-suite ACC summary. Displayed deltas equal the displayed final score minus the displayed baseline score.

H Theory illustrations

These schematics visualize three non-Euclidean/combinatorial guarantees and their scope; they are explanatory aids, not additional benchmark evidence.

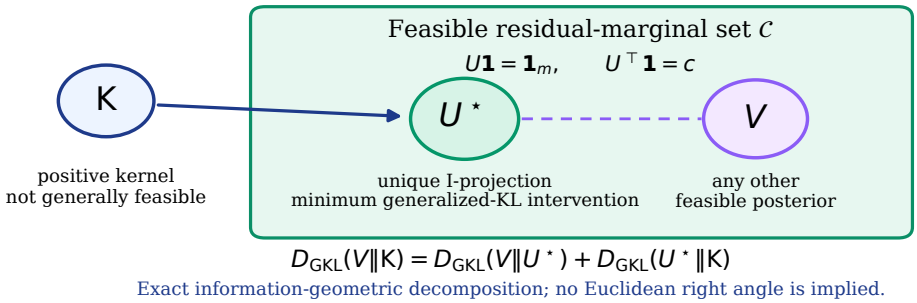

Figure 3: **Generalized-KL projection geometry.** The positive kernel matrix K generally lies outside the affine residual-marginal set $\mathcal{C}$. The solution $U^\star$ is its unique generalized-KL I-projection onto $\mathcal{C}$; every feasible V satisfies the exact decomposition in Corollary 2. The dashed segment is schematic only: no Euclidean orthogonality is claimed.

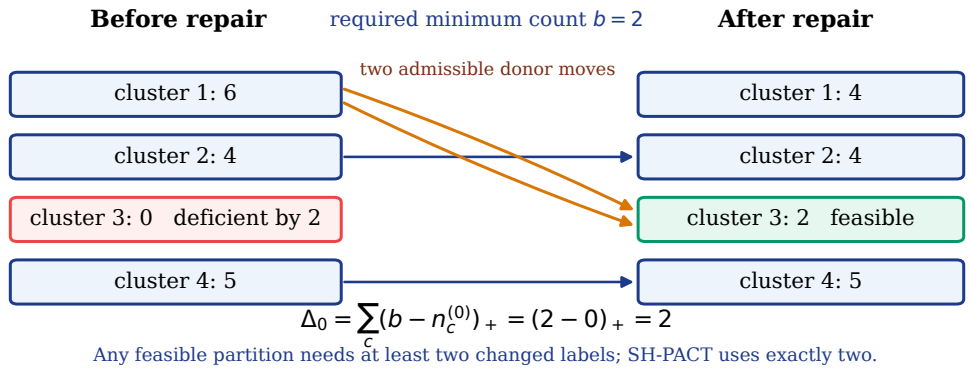

Figure 4: **Minimum-Hamming feasibility repair.** With required minimum count $b = 2$, the empty third cluster has deficit $\Delta_0 = (2-0)_+ = 2$. Every feasible repair must therefore change at least two initial labels, and SCoRP changes exactly two. Posterior cost only chooses the donor samples among currently admissible moves; the theorem does not claim globally minimum posterior cost among all minimum-Hamming repairs.

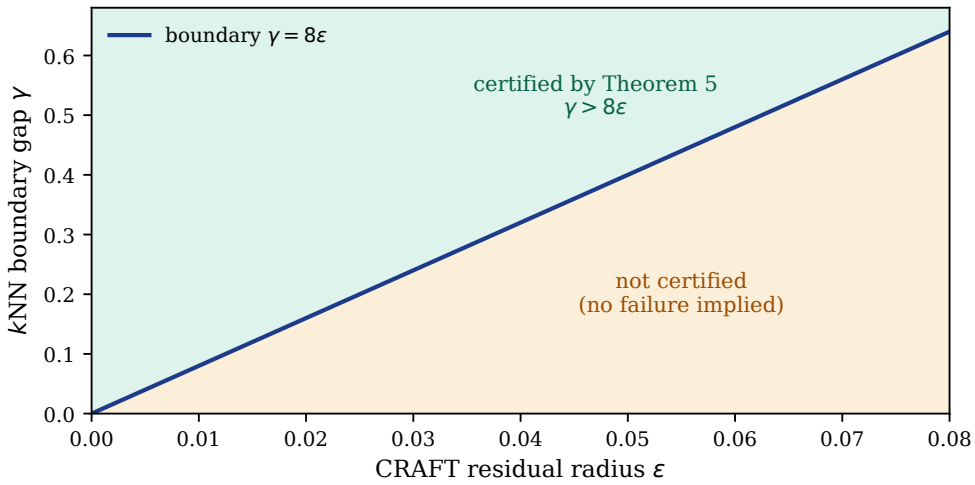

Figure 5: **CRAFT neighborhood certificate.** For a given point, let γ be the cosine-similarity gap between its kth and $(k+1)$th neighbors. Theorem 5 certifies that point's kNN set whenever $\gamma > 8\varepsilon$; if the condition holds pointwise for all samples, the whole kNN graph support is unchanged. On or below the boundary the theorem is silent and does not assert a neighborhood change.

Table 6: **Matched-feature comparison with TURTLE (%).** Both methods use the same 10,000-sample DINOv2-S/14 feature matrices. Values are from the predeclared fixed-seed runs, not mean±standard-deviation estimates. K_{occ} is the number of occupied hard clusters.

Dataset	Method	ACC	NMI	ARI	K_{occ}/K
CIFAR-10	SCoRP	93.74	88.21	87.01	10/10
	TURTLE	77.93	78.33	70.23	10/10
CIFAR-100	SCoRP	67.74	76.82	53.08	100/100
	TURTLE	36.88	61.01	28.20	98/100

I Modern comparator audit

Matched-feature TURTLE. The final matched run uses the official TURTLE optimization equations on the same saved DINOv2-S/14 feature matrices as SCoRP, with the predeclared CPU configuration $N_{\mathrm{out}} = 6000$ outer iterations, $N_{\mathrm{in}} = 10$ inner steps, full batch size 10,000, and seed 42. Here N_{out} and N_{in} are notation used only to avoid collision with SCoRP's symbols; they correspond to the official TURTLE command-line flags $T = 6000$ and $M = 10$. No labels are used in optimization or model selection. The resulting CIFAR-10 ACC/NMI/ARI is 77.93/78.33/70.23 with 10 occupied clusters; CIFAR-100 is 36.88/61.01/28.20 with 98 occupied clusters. The corresponding SCoRP values on the same matrices are 93.74/88.21/87.01 and 67.74/76.82/53.08, with full 10/10 and 100/100 hard support. These are fixed-seed values, not mean±standard-deviation estimates; the accompanying audit notebook independently repeats SCoRP on CIFAR-10/100 over five seeds, while the TURTLE result remains explicitly fixed-seed unless the matched TURTLE rerun is repeated.

Dataset	Method	Representation/protocol	ACC	NMI	ARI
CIFAR-10	TURTLE native pipeline	CLIP ViT-B/32, native context	85.88	80.20	76.60
CIFAR-100	TURTLE native pipeline	CLIP ViT-B/32, native context	45.73	60.96	32.36
CIFAR-10	PRO-DSC native pipeline	CLIP ViT-L/14, native context	97.08	92.75	–
CIFAR-20	PRO-DSC native pipeline	CLIP ViT-L/14, native context	69.91	71.68	–
CIFAR-100	PRO-DSC native pipeline	CLIP ViT-L/14, native context	79.05	83.07	–

Table 7: Native-pipeline comparator context (%). These rows use the representations and evaluation pipelines associated with the released methods and are not used for matched-feature superiority claims. Values correspond to the released native evaluation protocols used in this audit. A dash indicates that ARI was not produced by the corresponding protocol.

The native-context rows serve a different purpose from Table 6. They show the performance of the released methods in their own representation pipelines. In particular, the PRO-DSC CLIP ViT-L/14 values are numerically higher than the DINOv2-S SCoRP values on CIFAR-10/20/100, but the representation and learning protocol differ, so this is not evidence for or against the partition operator in a matched setting. Conversely, the matched TURTLE experiment controls the representation and therefore provides the cleaner test of the post-representation partition stage. The CIFAR-100 occupancy result is consistent with the motivation for explicit hard-feasibility repair, but does not by itself establish that the decoder is responsible for the full metric gap.