GDP: RELATION-PRESERVING CONSTRAINED CLUSTERING FROM GRADED PAIRWISE SUPERVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Pairwise constrained clustering typically uses binary supervision, while graded relations may reflect weak evidence, mixed membership, or coarser-resolution agreement. The same intermediate grade can imply different cluster structures, while relational orderings may appear on different numerical scales. Preserving graded relations and forming explicit cluster assignments are distinct but coordinated tasks. We introduce *Graded DualPair* (GDP), separating these roles across Local and Global representations, coordinating their pairwise relations while using relative grade ordering for direct Global supervision. GDP-Hier extends this principle across prescribed hierarchy resolutions. Our theory analyzes the relation–partition design and establishes sufficient conditions for multi-resolution hierarchy recovery. Across diverse benchmarks, GDP achieves consistently strong and frequently leading clustering performance and remains robust to targeted grade recalibration, while GDP-Hier yields strong multi-resolution clustering with cross-level coherence.

1 INTRODUCTION

Clustering uncovers structure in unlabeled data, but unsupervised partitions need not align with domain knowledge (Luo et al., 2018; Manduchi et al., 2021; Zhang & Chen, 2025). Pairwise constrained clustering (CC) injects sparse judgments of whether instances should be grouped together or apart (Davidson & Basu, 2007; González-Almagro et al., 2025); classical methods modify metrics, objectives, or penalties (Wagstaff & Cardie, 2000; Wagstaff et al., 2001; Lu & Leen, 2004; Bilenko et al., 2004; Basu et al., 2004; Kulis et al., 2005), while deep constrained clustering (DCC) learns from such supervision in complex high-dimensional data. Modern DCC either learns constraint-aware representations for subsequent clustering (Ohi et al., 2020; Fogel et al., 2019; Zhang & Chen, 2025), or directly shapes cluster assignments or co-assignment predictions (Hsu et al., 2018; Ren et al., 2019; Zhang et al., 2021; Śmieja et al., 2020; Tian et al., 2021; Goschenhofer et al., 2023; Nguyen et al., 2023). For binary supervision, the implication for clustering is direct: positive and negative relations correspond naturally to same- and different-cluster decisions.

Beyond binary supervision, existing formulations assign intermediate pairwise values specific interpretations: probability or confidence (Law et al., 2004; 2005), fuzzy relation strengths or penalties (Diaz-Valenzuela et al., 2016; Wang et al., 2022; Mei et al., 2019), or relation certainty and degrees of belief (Manduchi et al., 2021; Wang & Davidson, 2010). Yet the same numerical value need not have the same implication for clustering. Figure 1 illustrates the same grade of 0.5 for an occluded lion despite definite species membership, a lion–tiger hybrid (liger), or a lion–tiger pair in the same family (Felidae) but different species. Separately, the same relational ordering may be expressed on different

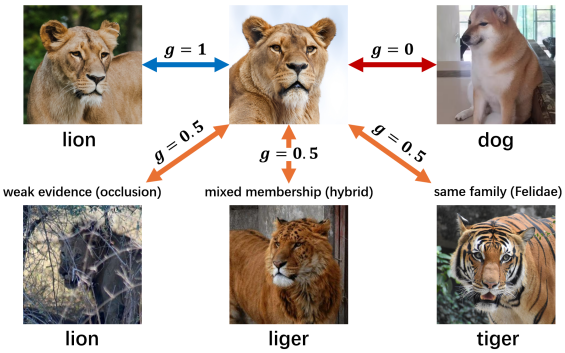

Figure 1: Illustration of graded pairwise supervision. Edge labels denote the relation grade g. The three $g = 0.5$ relations arise from weak evidence, mixed membership, and coarse hierarchical agreement, and need not imply the same flat cluster relation.

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.

numerical scales. Together, these observations motivate separating relation preservation from partition formation: directly fitting an intermediate grade to co-assignment imposes a particular numerical interpretation, whereas preserving it only in a learned representation leaves the final cluster assignments implicit. Our goal is therefore to preserve graded relations while forming explicit cluster assignments, without treating the raw grade as a universally calibrated flat co-membership target.

To address this tension, we introduce *Graded DualPair* (GDP), which assigns graded relation learning and prototype-based partition formation to Local and Global representations, respectively. GDP coordinates the representations through their pairwise relations, retaining richer relational information while relative grade ordering complements the learned relation values in guiding partition formation. GDP-Hier extends the same principle to hierarchy-induced grades, jointly learning levelwise cluster assignments from a shared graded relation. Our theoretical analysis supports the relation–partition design and its multi-resolution extension.

Our contributions are: (i) a structural formulation of graded pairwise constrained clustering separating relation preservation from partition formation without assuming universal flat co-membership semantics; (ii) GDP and GDP-Hier, a coordinated Local–Global framework for relation-preserving clustering from graded pairwise supervision at one or more resolutions; (iii) theoretical analysis of relation preservation, partition formation, and their coordination, with sufficient conditions for multi-resolution hierarchy recovery; (iv) extensive experiments demonstrating consistently strong and frequently leading clustering performance, robustness to targeted grade recalibration, effective multi-resolution hierarchy recovery with cross-level coherence, and ablations supporting the design.

2 Related work

Deep constrained clustering. Pairwise supervision in DCC has been used both to learn constraint-aware representations for subsequent clustering (Ohi et al., 2020; Fogel et al., 2019; Zhang & Chen, 2025) and to directly shape cluster assignments or co-assignment predictions (Hsu et al., 2018; Ren et al., 2019; Zhang et al., 2021; Śmieja et al., 2020; Tian et al., 2021; Goschenhofer et al., 2023; Nguyen et al., 2023). For binary positive/negative supervision, both routes naturally align with the same together/apart partition decision. With graded supervision, however, the two routes emphasize different parts of the problem: representation learning can preserve graded relations without determining the final partition, whereas direct co-assignment fitting assigns intermediate values a specific partition-level interpretation. GDP unifies these roles through coordinated Local and Global representations, preserving graded relations while forming explicit cluster assignments.

Graded pairwise supervision. Non-binary pairwise methods assign intermediate values specific semantics: probabilistic constraints encode relation uncertainty or confidence (Law et al., 2004; 2005); fuzzy formulations encode graded strengths or penalties, with SC-DEC relaxing should-link supervision through fuzzy memberships (Diaz-Valenzuela et al., 2016; Wang et al., 2022; Mei et al., 2019; Mei & Lv, 2018; Antoine et al., 2014; Li et al., 2018; AlZuhair et al., 2023); DC-GMM encodes must-/cannot-link type by sign and certainty by magnitude, while other work uses real-valued degrees of belief (Manduchi et al., 2021; Wang & Davidson, 2010). ECI-PP instead separates a canonical probabilistic relation from expert-conditioned judgment and stochastic corruption (Zhang & Chen, 2026). Across these formulations, numerical values have meanings defined by the supervision or observation model. Task-level grades can carry different meanings: the same value can reflect weak evidence for definite membership or genuine overlap, while the same ordering may appear on different numerical scales. GDP addresses this relation-to-partition gap by preserving graded relations while forming explicit assignments without treating raw grades as calibrated flat co-membership targets.

Hierarchical constrained clustering. Existing hierarchy-aware clustering has incorporated binary pairwise constraints (Davidson & Ravi, 2009), relative comparison information (Zheng & Li, 2011; Ghoshdastidar et al., 2019), and structural constraints or priors (Chatziafratis et al., 2018; Zheng et al., 2026). Real-valued pairwise supervision derived from class hierarchies has also been explored in constrained spectral clustering (Wang & Davidson, 2010), but in a transductive single-partition setting. Meanwhile, recent unsupervised work develops efficient hierarchical k-means for multi-resolution clustering (Zhao et al., 2026). Together, these lines leave open how graded pairwise supervision can support joint learning across semantic resolutions; GDP-Hier decomposes the graded relation across resolutions to learn levelwise cluster assignments in a single jointly trained model.

3 PROBLEM FORMULATION

Graded pairwise constrained clustering. Given a dataset $\mathcal{X} = \{x_j\}_{j=1}^{|\mathcal{X}|}$, a task-specified cluster count K, and directly specified graded constraints $\mathcal{C} = \{(a_i, b_i, g_i)\}_{i=1}^{|\mathcal{C}|}$, where (a_i, b_i) indexes (x_{a_i}, x_{b_i}) and $g_i \in [0, 1]$ is its *relation grade*, we seek a K-way partition $\mathcal{S} = \{S_k\}_{k=1}^{K}$ of $\mathcal{X}$.

The endpoints $g_i = 1$ and $g_i = 0$ recover binary positive and negative pairwise constraints, respectively. Between these endpoints, larger grades indicate stronger grouping relations, but an intermediate grade need not be a calibrated probability of same-cluster membership; its precise meaning can depend on how the graded supervision is defined. Thus, preserving graded pairwise relations and forming explicit cluster assignments are related but distinct requirements.

Hierarchy-induced graded relations. For hierarchy-induced supervision, we consider a prescribed nested hierarchy with L semantic resolutions ordered from coarse to fine and cluster counts $K^{(1)} < \cdots < K^{(L)}$. At resolution ℓ, the target partition is $\mathcal{S}^{(\ell)} = \{S_k^{(\ell)}\}_{k=1}^{K^{(\ell)}}$, where every cluster at level $\ell + 1$ has one parent at level ℓ. Prescribed hierarchy grade breakpoints $0 = \beta_0 < \beta_1 < \cdots < \beta_L = 1$ define the grading scheme. Let $\ell_i^\star \in \{0, \ldots, L\}$ be the deepest level at which x_{a_i} and x_{b_i} belong to the same cluster, with $\ell_i^\star = 0$ if they differ at the coarsest level; the corresponding relation grade is

$$g_i = \beta_{\ell_i^\star}. \tag{1}$$

Thus, an intermediate hierarchy grade identifies the deepest resolution at which the pair remains co-clustered. Accordingly, the objective is to recover accurate partitions at each resolution while preserving the cross-level nesting structure that makes them a hierarchy.

4 GRADED DUALPAIR

Graded DualPair (GDP) separates relation preservation from partition formation using sequential Local and Global representations: the Local representation captures graded pairwise relations, while the Global representation organizes samples into a prototype-based cluster partition. The representations are coordinated during learning, and GDP-Hier extends the same principle to hierarchy-induced relations through resolution-specific Global heads. Figure 2 provides an overview.

4.1 LOCAL GRADED RELATION LEARNING

We use a Local encoder $f_{\phi_{\text{loc}}}$, yielding $z_j^{\text{loc}} = f_{\phi_{\text{loc}}}(x_j) \in \mathbb{R}^{D_{\text{loc}}}$ and $\mathcal{Z}^{\text{loc}} = \{z_j^{\text{loc}}\}_{j=1}^{|\mathcal{X}|}$. To retain graded pairwise structure as an explicit relational signal for Local–Global coordination, we use the ProbPair angular readout (Zhang & Chen, 2026) and fit its output directly to the supplied grades:

$$r_i^{\text{loc}} := \sigma\big((\cos(z_{a_i}^{\text{loc}}, z_{b_i}^{\text{loc}}) - m)/T\big) \in (0, 1), \tag{2}$$

$$\mathcal{L}_{\text{PP}} = -\frac{1}{|\mathcal{C}|} \sum_{i=1}^{|\mathcal{C}|} \left[g_i \log r_i^{\text{loc}} + (1 - g_i) \log(1 - r_i^{\text{loc}}) \right]. \tag{3}$$

Here σ is the sigmoid function, and learnable m and $T > 0$ control the operating point and sharpness of the angular mapping. Following Zhang & Chen (2025; 2026), reconstruction over all samples complements pairwise supervision by preserving intrinsic data structure in the Local representation. With decoder d_ψ, we use $\hat{x}_j = d_\psi\big(z_j^{\text{loc}}/\|z_j^{\text{loc}}\|_2\big)$ and $\mathcal{L}_{\text{rec}} = |\mathcal{X}|^{-1} \sum_j \|x_j - \hat{x}_j\|_2^2$.

4.2 GLOBAL PARTITION FORMATION

Prototype assignment. A Global encoder $f_{\phi_{\text{glob}}}$ maps Local features to a Global feature space for prototype-based clustering, yielding $z_j^{\text{glob}} = f_{\phi_{\text{glob}}}(z_j^{\text{loc}}) \in \mathbb{R}^{D_{\text{glob}}}$ and $\mathcal{Z}^{\text{glob}} = \{z_j^{\text{glob}}\}_{j=1}^{|\mathcal{X}|}$. We fix K unit prototypes $\{w_k\}_{k=1}^{K}$ at regular-simplex vertices, with $w_k^\top w_{k'} = -1/(K-1)$ for $k \neq k'$. Each sample receives a soft assignment $q_j \in \Delta^{K-1}$, giving the Global co-assignment relation

$$q_{jk} = \text{softmax}_k\big(\rho(z_j^{\text{glob}}, w_k)\big), \qquad r_i^{\text{glob}} = q_{a_i}^\top q_{b_i}. \tag{4}$$

Here $\rho(\cdot, \cdot)$ denotes the norm-scaled angular assignment score, which preserves the fixed simplex geometry while allowing assignment concentration to vary with the Global feature norm; its construction is given in Appendix B.1. The resulting hard partition is given by $S_k = \{x_j : \arg\max_{k'} q_{jk'} = k\}$.

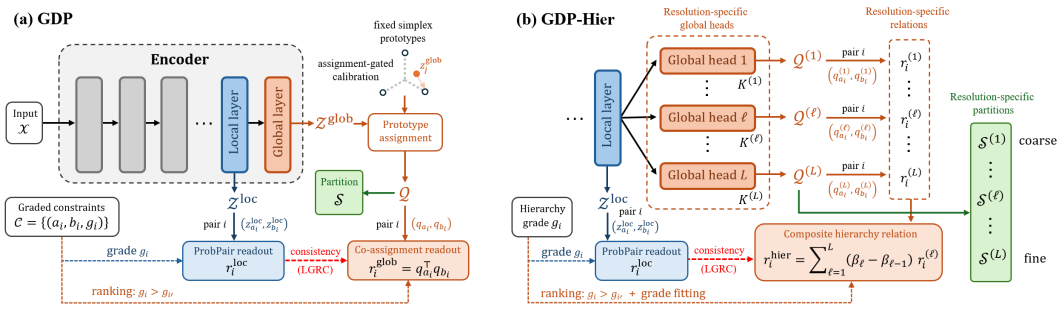

Figure 2: Overview of GDP and GDP-Hier. In GDP, the **Local representation** learns graded relations r_i^{loc} from supplied grades, while the **Global representation** forms soft assignments $\mathcal{Q} = \{q_j\}_j$ over fixed simplex prototypes and co-assignment relations r_i^{glob} in Eq. (4). The Global representation uses order-based supervision and assignment-gated calibration; the **Local-to-Global link** aligns Local and Global relations. Dominant assignments yield **partition** $\mathcal{S}$. GDP-Hier combines resolution-specific Global relations into r_i^{hier} in Eq. (10), yielding coarse-to-fine **partitions** $\{\mathcal{S}^{(\ell)}\}_{\ell=1}^{L}$.

Order-based supervision. Since an intermediate grade need not be a calibrated flat co-membership target, enforcing $r_i^{\mathrm{glob}} = g_i$ would impose a stronger numerical interpretation than partition formation requires. Direct Global supervision instead uses relative grade ordering. For $\mathcal{P} = \{(i, i') : g_i > g_{i'}\}$, the ranking objective is

$$\mathcal{L}_{\mathrm{rank}} = \frac{1}{|\mathcal{P}|} \sum_{(i,i') \in \mathcal{P}} \left[r_{i'}^{\mathrm{glob}} - r_i^{\mathrm{glob}} \right]_+, \qquad [x]_+ := \max(0, x). \tag{5}$$

A comparison is therefore penalized only when the predicted Global relations reverse the supplied grade order. $\mathcal{L}_{\mathrm{rank}}$ therefore enforces only the required ordering, without assuming that grade differences correspond to fixed numerical gaps between Global relation scores.

Assignment-adaptive geometric calibration. Order-based supervision constrains pairwise relations but does not determine how Global representations consolidate around the prototypes fixed at simplex vertices. GDP therefore makes the corresponding vertex calibration assignment-adaptive. We project z_j^{glob} onto the simplex subspace and divide by a learnable radius to obtain u_j (Appendix B.1). Let $\mathrm{sg}[\cdot]$ denote stop-gradient; we define the assignment-dependent gate and vertex loss as

$$c(q_j) = \frac{K\|q_j\|_2^2 - 1}{K - 1}, \qquad \mathcal{L}_{\mathrm{vertex}} = \frac{1}{|\mathcal{X}|} \sum_j c(\mathrm{sg}[q_j]) \sum_k \mathrm{sg}[q_{jk}] \|u_j - w_k\|_2^2. \tag{6}$$

The score $c(q_j)$ ranges from 0 for a uniform to 1 for a one-hot assignment and acts as an assignment-dependent gate: concentrated assignments receive stronger prototype calibration, while diffuse ones retain greater freedom to reorganize. Detaching assignments in $\mathcal{L}_{\mathrm{vertex}}$ calibrates the Global geometry without adding an assignment target. Finally, we maximize marginal assignment entropy H_{marg} to discourage degenerate prototype usage.

4.3 Local–Global Coordination

Because order-based supervision constrains Global relation order but not values, the Local relation provides a complementary value-based reference. For pair (a, b), let r_{ab}^{loc} denote the ProbPair relation in Eq. (2) and $r_{ab}^{\mathrm{glob}} = q_a^\top q_b$ its Global co-assignment. Over $\mathcal{B}_{\mathrm{LG}}$, containing constrained and additional sampled pairs, we define the Local–Global relation consistency (LGRC) objective as

$$\mathcal{L}_{\mathrm{LGRC}} = \frac{1}{|\mathcal{B}_{\mathrm{LG}}|} \sum_{(a,b) \in \mathcal{B}_{\mathrm{LG}}} \left(r_{ab}^{\mathrm{glob}} - \mathrm{sg}[r_{ab}^{\mathrm{loc}}] \right)^2, \tag{7}$$

Here $\mathrm{sg}[\cdot]$ denotes stop-gradient. Within LGRC, the detached Local relation is the reference and the Global relation is optimized toward it. Ranking and LGRC are complementary: the former penalizes grade-order reversals, while the latter transfers the richer Local relation values.

With all components defined, the Local and Global objectives are

$$\mathcal{L}_{\text{loc}} = \mathcal{L}_{\text{PP}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}}, \tag{8}$$

$$\mathcal{L}_{\text{glob}} = \mathcal{L}_{\text{rank}} - \alpha H_{\text{marg}} + \gamma\mathcal{L}_{\text{vertex}} + \eta\mathcal{L}_{\text{LGRC}}, \tag{9}$$

where λ_{rec}, α, γ, and η control reconstruction, marginal entropy, vertex calibration, and Local–Global consistency, respectively. During training, this directed Local-to-Global matching is complemented by periodic joint updates that allow the emerging Global partition to refine the Local representation. This two-way coordination allows relational learning and partition formation to inform each other without conflating their roles. The resulting training procedure is summarized in Appendix B.1.

4.4 GDP-HIER FOR HIERARCHY-INDUCED RELATIONS

Resolution-specific assignments. For the hierarchy-induced grades in Eq. (1), partition formation must span semantic resolutions rather than a single Global co-assignment. GDP-Hier retains the shared Local representation and uses L resolution-specific Global heads, with head ℓ containing $K^{(\ell)}$ fixed simplex prototypes. Applying Eq. (4) at each level yields $q_j^{(\ell)} \in \Delta^{K^{(\ell)}-1}$ and an explicit partition $\mathcal{S}^{(\ell)}$ from dominant assignments. The levelwise relations explain the hierarchy grade through the breakpoint-weighted composite:

$$r_i^{(\ell)} = q_{a_i}^{(\ell)\top} q_{b_i}^{(\ell)}, \qquad r_i^{\text{hier}} = \sum_{\ell=1}^{L} (\beta_\ell - \beta_{\ell-1})\, r_i^{(\ell)}. \tag{10}$$

Composite hierarchy relation. The composite r_i^{hier} expresses the hierarchy grade through breakpoint-weighted levelwise agreement across resolutions. Unlike generic task-level grades, hierarchy grades are prescribed numerically by the breakpoints in Eq. (1), making direct fitting of the composite semantically well-defined: $\mathcal{L}_{\text{fit}} = |\mathcal{C}|^{-1} \sum_{i=1}^{|\mathcal{C}|} (r_i^{\text{hier}} - g_i)^2$.

GDP-Hier applies the ranking objective to r_i^{hier}, which also serves as LGRC's Global-side relation. Assignment-adaptive vertex calibration and marginal entropy act on each resolution-specific head. With $\overline{H}_{\text{marg}}$ and $\overline{\mathcal{L}}_{\text{vertex}}$ denoting head averages, the hierarchy Global objective is

$$\mathcal{L}_{\text{glob}}^{\text{hier}} = \mathcal{L}_{\text{rank}}^{\text{hier}} + \mathcal{L}_{\text{fit}} - \alpha\overline{H}_{\text{marg}} + \gamma\overline{\mathcal{L}}_{\text{vertex}} + \eta\mathcal{L}_{\text{LGRC}}^{\text{hier}}. \tag{11}$$

Here $\mathcal{L}_{\text{rank}}^{\text{hier}}$ and $\mathcal{L}_{\text{LGRC}}^{\text{hier}}$ are the objectives in Eqs. (5) and (7) evaluated on the composite relation; the shared Local representation retains Eq. (8). The hierarchy heads are thus coupled through shared graded supervision and the composite relation, coordinating partition learning across resolutions.

At inference, dominant Global assignments yield the partition $\mathcal{S}$ for GDP and the resolution-specific partitions $\{\mathcal{S}^{(\ell)}\}_{\ell=1}^{L}$ for GDP-Hier; training procedures are summarized in Appendix B.1.

5 THEORETICAL ANALYSIS

We analyze the separation and coordination of graded relations and partition formation, then establish hierarchy factorization and recovery across resolutions.

Relation–partition separation and coordination. We first quantify how decisive prototype assignments constrain Global pairwise relations.

Lemma 5.1 (Relation bounds under decisive assignments)**.** *After aligning distinct latent clusters with distinct prototypes, suppose that every sample assigns at least $1-\epsilon$ mass to its aligned prototype, where $0 \le \epsilon < 1/2$. Then $r_{ab}^{\text{glob}} \ge (1-\epsilon)^2$ for a same-cluster pair, whereas $r_{ab}^{\text{glob}} \le 2\epsilon$ for a different-cluster pair, quantifying the compression induced by decisive assignments.*

Appendix A.1 proves the bounds. Decisive Global assignments therefore push same- and different-cluster relations toward 1 and 0, respectively, motivating the Local representation to preserve graded variation. Zero-margin Ranking is likewise partition-compatible: with latent hard co-membership $h_i \in \{0,1\}$, $g_i > g_{i'} \Rightarrow h_i \ge h_{i'}$ gives zero loss even when distinct grades share the same h_i. The formal result and proof are given in Appendix A.2.

Ranking constrains relation order, while LGRC complements this with Local reference values. For a relation-space characterization, let $\mathbf{r} = (r_i)_{i=1}^{|\mathcal{C}|}$ collect constrained-pair relations with Local counterpart $\mathbf{r}^{\mathrm{loc}}$. With $\mathcal{P} = \{(i, i') : g_i > g_{i'}\}$ and $\mathcal{R}_g = \{\mathbf{r} \in [0,1]^{|\mathcal{C}|} : r_i \geq r_{i'}\ \forall (i, i') \in \mathcal{P}\}$, we obtain:

Theorem 5.2 (Local relation transfer under order constraints)**.** *Let $\bar{\mathbf{r}}$ be the Euclidean projection of $\mathbf{r}^{\mathrm{loc}}$ onto $\mathcal{R}_g$, and let $\mathbf{r}_\lambda$ minimize $\frac{1}{2}\|\mathbf{r} - \mathbf{r}^{\mathrm{loc}}\|_2^2 + \lambda \mathcal{L}_{\mathrm{rank}}(\mathbf{r})$ over $[0,1]^{|\mathcal{C}|}$. Then $\mathbf{r}_\lambda \to \bar{\mathbf{r}}$ as $\lambda \to \infty$; moreover, $\bar{\mathbf{r}} = \mathbf{r}^{\mathrm{loc}}$ whenever $\mathbf{r}^{\mathrm{loc}} \in \mathcal{R}_g$, so feasible Local relations remain unchanged.*

Appendix A.3 proves the result. In relation space, increasing Ranking pressure drives relations toward the grade order, while LGRC favors the values closest to the Local reference; the theorem does not imply convergence of the full GDP optimization. Global partition formation further depends on how representations consolidate around the prototypes. For the calibration gate in Eq. (6), the weighted prototype average $s(q) = \sum_k q_k w_k$ satisfies $c(q) = \|s(q)\|_2^2$ under regular-simplex geometry; with q detached, differentiating $\mathcal{L}_{\mathrm{vertex}}$ with respect to u yields a gradient proportional to $c(q)(u - s(q))$. Thus, $s(q)$ sets the calibration target and $c(q)$ its strength; see Appendix A.4 for the derivation.

Hierarchy factorization and recovery. A single decisive finest-resolution head cannot in general represent a positive intermediate hierarchy grade. Indeed, for a pair separated at the finest level, Lemma 5.1 gives $r_{ab}^{(L)} \leq 2\epsilon$; fitting an intermediate grade β_ℓ within error $\delta < \beta_\ell$ therefore requires $\epsilon \geq (\beta_\ell - \delta)/2$. This motivates distributing hierarchy relations across their semantic resolutions.

Proposition 5.3 (Multi-resolution factorization)**.** *For the nested hierarchy, let g_{ab} be the grade of pair (a, b), $\Delta_\ell^\beta := \beta_\ell - \beta_{\ell-1}$, and $h_{ab}^{(\ell)}$ its level-ℓ co-membership indicator. Then $g_{ab} = \sum_{\ell=1}^{L} \Delta_\ell^\beta h_{ab}^{(\ell)}$, and $\{\Delta_\ell^\beta\}_{\ell=1}^{L}$ are the unique additive weights reproducing all hierarchy breakpoints.*

Appendix A.5 proves the factorization. Replacing $h_{ab}^{(\ell)}$ by the soft levelwise relations $r_{ab}^{(\ell)}$ yields the GDP-Hier composite relation in Eq. (10) and the following sufficient recovery result.

Theorem 5.4 (General-L hierarchy recovery)**.** *At every level ℓ, suppose each assignment places at least $1 - \epsilon_\ell$ mass on its argmax ($0 \leq \epsilon_\ell < 1/2$), the argmax partition has $K^{(\ell)}$ nonempty clusters, and constrained within-cluster pairs connect every true cluster. If $|r_{ab}^{\mathrm{hier}} - g_{ab}| \leq \delta$ on these pairs and $\delta + 2\sum_{\ell=1}^{L} \Delta_\ell^\beta \epsilon_\ell < \Delta_{\min}^\beta$, where $\Delta_{\min}^\beta := \min_\ell \Delta_\ell^\beta$, then all levelwise argmax partitions recover the corresponding true hierarchy partitions up to label permutation.*

Appendix A.6 proves the result under the stated assumptions, linking hierarchy recovery to accurate composite relations, decisive levelwise assignments, and hierarchy-breakpoint separation.

6 Experiments

6.1 Experimental settings

We structure our evaluation around four questions: (i) how well GDP performs when pairwise grades carry different clustering implications; (ii) whether GDP remains robust across numerical grade scales; (iii) whether GDP-Hier recovers accurate and coherent hierarchies across semantic resolutions; and (iv) whether the proposed mechanisms behave as intended.

Datasets. We evaluate single-resolution clustering on seven benchmarks (CIFAR100-20, CIFAR10, FMNIST, ImageNet10, MNIST, Reuters, and STL10) and multi-resolution hierarchy recovery on CIFAR100-HIER, RCV1-HIER, Aircraft (Maji et al., 2013), and tieredImageNet-H (Bertinetto et al., 2020) at two scales (T200 and T608). Together, they span image and text, varying cluster counts and class balance, and hierarchies with up to four semantic resolutions and 608 finest-level clusters. T608 contains 608 ImageNet classes organized through WordNet; T200 is its 200-class subset.

Compared methods. We compare with VanillaDCC (Hsu et al., 2019), VolMaxDCC (Nguyen et al., 2023), CIDEC (Zhang et al., 2021), SpherePair (Zhang & Chen, 2025), ProbPair, Weighted ProbPair (WPP), and ECI-PP (Zhang & Chen, 2026). These methods cover representative end-to-end DCC, deep constraint embedding, and probabilistic pairwise-relation learning approaches.

Pairwise supervision. We use four forms of pairwise supervision with distinct clustering interpretations. (i) *Binary* records class agreement: 1 for same-class and 0 for different-class pairs. (ii)

Table 1: Test ACC/NMI/ARI (%; five-run means) at 5k constraints. Best/second-best are **bold**/underlined per dataset/source/metric; WPP reduces to ProbPair under Binary and is omitted.

Method	CIFAR100-20	CIFAR10	FMNIST	ImageNet10	MNIST	Reuters	STL10
Binary							
VanillaDCC	48.9 / 51.7 / 35.0	88.1 / 84.4 / 81.3	76.7 / 67.3 / 61.3	88.7 / 82.6 / 80.3	82.1 / 74.8 / 71.5	95.0 / 81.2 / 87.9	82.2 / 74.8 / 70.7
VolMaxDCC	44.6 / 47.8 / 29.5	89.1 / 84.5 / 82.1	68.6 / 62.9 / 54.4	77.5 / 71.6 / 66.7	81.1 / 73.1 / 71.0	93.4 / 76.7 / 84.3	78.9 / 71.4 / 62.6
CIDEC	57.6 / 60.5 / 43.6	94.8 / 89.0 / 89.1	67.6 / 63.6 / 53.8	95.5 / 89.9 / 90.4	96.6 / 91.8 / 92.7	82.3 / 69.5 / 77.6	88.9 / 79.6 / 77.8
SpherePair	69.4 / 64.5 / 53.4	94.7 / 88.8 / 89.0	79.5 / 71.2 / 65.2	95.8 / 90.2 / 90.9	96.4 / 90.9 / 92.2	94.5 / 79.9 / 86.6	89.3 / 80.3 / 78.7
ProbPair	66.6 / 64.0 / 50.7	94.7 / 88.8 / 88.9	76.9 / 68.7 / 61.3	95.8 / 90.2 / 90.9	94.8 / 87.8 / 88.9	94.9 / 81.0 / 87.7	88.7 / 79.6 / 77.5
WPP	–	–	–	–	–	–	–
ECI-PP	63.9 / 63.2 / 49.3	94.6 / 88.9 / 88.9	70.2 / 66.0 / 56.5	**96.3 / 91.5 / 92.1**	93.5 / 85.9 / 86.5	94.8 / 81.3 / 88.0	88.9 / 79.9 / 78.0
GDP (Ours)	**71.0 / 66.2 / 54.7**	**94.9 / 89.3 / 89.4**	**81.8 / 73.4 / 68.1**	96.1 / 91.0 / 91.7	**97.1 / 92.5 / 93.8**	**95.2 / 82.1 / 88.5**	**89.9 / 81.2 / 79.7**
Difficulty							
VanillaDCC	17.4 / 15.3 / 4.1	74.3 / 70.2 / 63.3	47.7 / 43.2 / 27.1	40.7 / 34.1 / 23.6	57.3 / 51.6 / 41.3	88.3 / 68.4 / 76.2	52.8 / 52.2 / 32.6
VolMaxDCC	45.4 / 49.0 / 30.6	87.1 / 83.4 / 80.9	70.2 / 62.1 / 54.2	74.3 / 68.3 / 62.4	75.6 / 69.6 / 64.4	91.9 / 75.1 / 82.5	83.5 / 71.2 / 68.7
CIDEC	25.9 / 29.7 / 2.5	93.2 / 85.7 / 85.9	45.7 / 42.7 / 19.8	81.1 / 67.7 / 61.4	84.8 / 78.0 / 66.5	87.9 / 66.3 / 75.3	76.9 / 61.6 / 56.8
SpherePair	68.0 / 62.9 / 50.8	94.7 / 88.6 / 88.8	77.4 / 67.7 / 62.0	93.7 / 86.2 / 86.6	94.3 / 87.3 / 88.0	**93.9 / 78.8 / 85.7**	86.7 / 76.5 / 74.5
ProbPair	60.8 / 59.5 / 42.3	94.5 / 88.6 / 88.5	69.4 / 65.5 / 55.5	94.2 / 87.7 / 87.5	92.6 / 83.7 / 84.6	86.7 / 70.3 / 73.3	87.1 / 77.7 / 74.6
WPP	63.2 / 61.5 / 45.7	94.4 / 88.5 / 88.3	72.6 / 65.7 / 56.8	94.3 / 87.8 / 87.8	92.6 / 83.9 / 84.6	86.9 / 70.6 / 73.6	83.8 / 75.9 / 70.9
ECI-PP	64.3 / 62.6 / 47.9	94.3 / 88.4 / 88.2	72.9 / 67.3 / 58.5	95.7 / 90.3 / 90.7	92.3 / 83.7 / 84.0	91.0 / 75.7 / 81.1	87.7 / 78.8 / 75.4
GDP (Ours)	**69.9 / 64.5 / 52.9**	**94.9 / 89.1 / 89.3**	**81.1 / 72.8 / 67.4**	**96.2 / 91.3 / 91.9**	**96.3 / 90.6 / 92.0**	93.5 / 78.7 / 85.6	**89.4 / 80.5 / 78.8**
Mix							
VanillaDCC	48.3 / 49.4 / 34.1	90.7 / 85.6 / 83.5	72.8 / 64.4 / 57.7	84.9 / 80.1 / 76.6	85.0 / 73.8 / 72.5	92.2 / 75.7 / 83.6	84.2 / 75.3 / 72.2
VolMaxDCC	42.0 / 46.6 / 27.3	86.2 / 81.9 / 77.5	67.0 / 61.4 / 52.6	68.4 / 62.9 / 55.4	68.3 / 61.2 / 54.6	93.1 / 76.8 / 84.3	71.9 / 63.7 / 52.0
CIDEC	59.4 / 59.8 / 44.9	92.9 / 86.3 / 85.4	70.6 / 64.8 / 56.4	95.5 / 89.8 / 90.3	94.0 / 87.3 / 87.1	90.1 / 71.0 / 80.2	88.1 / 78.6 / 76.4
SpherePair	63.9 / 61.0 / 47.3	92.9 / 86.8 / 85.9	73.4 / 65.4 / 57.7	93.6 / 86.1 / 86.4	89.7 / 82.5 / 80.9	94.3 / 80.0 / 86.8	85.3 / 75.9 / 72.7
ProbPair	**68.5 / 63.7 / 51.9**	94.6 / 88.8 / 88.8	70.0 / 64.4 / 55.2	96.2 / 90.9 / 91.7	92.5 / 83.5 / 84.4	**94.7 / 80.6 / 87.1**	88.9 / 79.7 / 77.9
WPP	65.5 / 62.5 / 49.6	94.6 / 88.8 / 88.8	71.8 / 65.4 / 56.7	96.0 / 90.5 / 91.3	93.2 / 84.8 / 85.8	**94.7** / 80.4 / 87.0	88.7 / 79.3 / 77.6
ECI-PP	62.6 / 61.6 / 47.8	94.0 / 88.3 / 87.8	69.3 / 65.6 / 55.6	96.0 / 90.7 / 91.3	91.6 / 82.8 / 82.7	94.1 / 79.6 / 86.4	87.7 / 78.4 / 75.9
GDP (Ours)	64.0 / 62.3 / 47.7	**95.1 / 89.4 / 89.7**	**79.3 / 71.3 / 65.1**	**96.5 / 91.7 / 92.4**	**95.8 / 89.6 / 91.0**	94.5 / 80.2 / 87.0	**89.8 / 81.0 / 79.6**

Difficulty corresponds to the weak-evidence case in Figure 1: class membership remains definite; using easiness scores from a frozen auxiliary classifier, lower-easiness pairs receive grades nearer the default boundary $b = 0.5$, while higher-easiness pairs approach Binary endpoints. (iii) *Mix* represents genuine partial membership. We augment each training split with an equal number of cross-class mixtures $\boldsymbol{p} = \lambda \boldsymbol{e}_u + (1-\lambda)\boldsymbol{e}_v$ from one-hot vectors $\boldsymbol{e}_u, \boldsymbol{e}_v$, with $\lambda \sim \mathcal{U}(0,1)$, set $g_{ab} = \boldsymbol{p}_a^\top \boldsymbol{p}_b$, and sample constraints jointly from original and mixed samples. (iv) *Hierarchy* encodes the deepest shared semantic resolution, with internal grade breakpoints evenly spaced between 0 and 1 by default.

Protocol. We report ACC/NMI/ARI over five runs; embedding methods use K-means on learned embeddings, while end-to-end methods use native assignments. Main single-resolution results use 5k constraints for Binary, Difficulty, and Mix; Mix is evaluated only on original test samples because mixtures lack unambiguous labels. Grade-scale robustness varies the Difficulty boundary and Mix calibration under a sufficiently large 10k constraint budget. Hierarchy recovery uses 5k constraints and reports levelwise quality and hierarchy violation rate (HVR): the fraction of samples whose predicted parent disagrees with the majority parent of their predicted child cluster (lower is better). Component ablations and mechanism analyses test whether the proposed designs behave as intended.

Implementation. For fair comparison, autoencoder-based methods share a 500–500–2000 fully connected backbone where applicable, while architecture-specific baselines retain native networks. Global and autoencoder latent dimensions are set to K, but 10 on Reuters. Fixed real-world image features are learned via self-supervision on complete pre-split collections and shared across methods; autoencoder pretraining provides model initialization on original training samples. Component ablations use random initialization. GDP uses **one shared setting without dataset-specific tuning**: $\alpha = 0.2$, $\gamma = 1$, $\eta = 2$, and $\lambda_{\text{rec}} = 0.02$, with Global updates and feedback every five epochs.

Full experimental details are provided in Appendix C to support reproducibility.

6.2 Experimental results

Main clustering effectiveness. Table 1 compares GDP with seven baselines under Binary, Difficulty, and Mix supervision at 5k constraints. Overall, GDP is strongest, ranking first in 51 of 63 dataset–supervision–metric comparisons and within the top two in 59/63. The advantage holds across all three forms; on FMNIST, its ACC exceeds the next-best method by 2.3–5.9 points. Without autoencoder pretraining, GDP's advantage is even clearer, indicating limited dependence on autoencoder initialization. Appendix D.1 reports both initialization regimes at 1k/5k/10k constraint budgets.

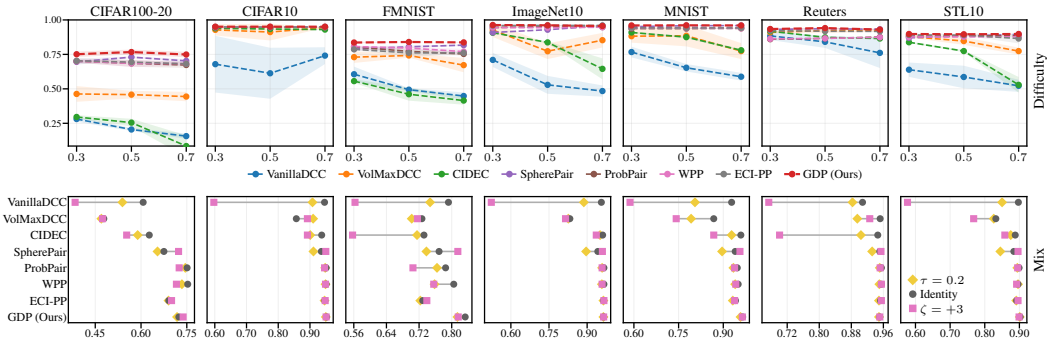

Figure 3: Test ACC under grade recalibration at 10k constraints. Top: Difficulty at $b = 0.3/0.5/0.7$; bottom: Mix at identity, $\tau = 0.2$, and $\zeta = +3$. Gray spans mark method-wise ranges.

Table 2: Hierarchy recovery with 5k constraints (five-run means, %). Mean/Leaf denote mean/finest-level test ACC; best/second-best per dataset/metric are **bold**/underlined.

	CIFAR100-HIER			RCV1-HIER			Aircraft			T200			T608		
Method	Mean↑	Leaf↑	HVR↓	Mean↑	Leaf↑	HVR↓	Mean↑	Leaf↑	HVR↓	Mean↑	Leaf↑	HVR↓	Mean↑	Leaf↑	HVR↓
VanillaDCC	36.5	21.3	48.4	51.7	20.6	19.8	15.7	11.1	63.0	20.7	11.8	65.3	14.0	5.2	70.9
CIDEC	51.2	44.2	22.6	54.2	21.3	16.4	**17.0**	11.3	52.9	32.6	30.4	56.6	22.2	17.2	64.4
SpherePair	57.0	52.4	7.4	79.9	65.8	1.7	14.2	13.8	19.4	33.8	33.5	15.7	20.7	17.6	23.0
ProbPair	61.2	50.9	10.3	80.9	67.4	**1.2**	16.6	**15.0**	30.9	31.8	30.8	23.9	20.6	18.7	33.2
WPP	**61.6**	52.1	10.9	81.5	68.5	**1.2**	16.2	14.8	32.6	32.3	32.4	24.3	21.4	19.2	33.2
FlatGDP	58.0	54.9	20.7	65.2	36.2	10.3	15.6	14.3	59.3	41.3	37.6	42.9	17.7	6.1	73.7
GDP-Hier (Ours)	60.3	**56.1**	**3.8**	**83.4**	**72.0**	2.3	16.2	**15.0**	**11.8**	**47.3**	**52.9**	**4.8**	**29.1**	**24.7**	**6.5**

More importantly, GDP remains consistently strong as grade semantics change. While the strongest baseline shifts across Binary, Difficulty, and Mix, GDP remains near the top, with further cross-source evidence in Appendix D.1. Under Mix, hard-label scores on original samples cannot reveal partial-membership recovery, so Appendix D.2 analyzes recovered mixture composition and shows that GDP tracks the underlying graded membership. Difficulty is intentionally order-separable and serves as a controlled calibration test rather than a model of arbitrary noisy human judgements; Appendix D.3 complements this with CIFAR10H evidence from human-derived grades (Peterson et al., 2019). Across these qualitatively different grade semantics, GDP remains reliable.

Robustness to numerical grade scales. We test numerical-scale robustness: Difficulty grades need not be calibrated probabilities, while Mix tests dependence on its native overlap scale. Table 1 varies Difficulty $b \in \{0.3, 0.5, 0.7\}$ and Mix via $g' = \sigma(\tau \operatorname{logit}(g) + \zeta)$: identity ($\tau = 1, \zeta = 0$), $\tau = 0.2$ makes mixed-membership evidence less decisive, and $\zeta = +3$ shifts predominantly low grades toward the high end. Figure 3 shows that GDP combines high ACC with markedly greater stability across both settings. When Mix grades directly reflect membership overlap, relation-fitting baselines can compete with GDP; after recalibration, these baselines vary more, whereas GDP remains stable. Overall, the results support GDP's reduced dependence on exact grade calibration; see Appendix D.4.

Multi-resolution hierarchy recovery. Table 2 compares GDP-Hier with representative end-to-end, constraint-embedding, and probabilistic pairwise baselines, plus FlatGDP, which trains GDP independently at each resolution. A recovered hierarchy must be accurate within each resolution and coherent across resolutions, since strong levelwise partitions can still contradict one another. VanillaDCC and CIDEC likewise retrain for each target K; SpherePair, ProbPair, and WPP reuse one embedding with K-means, whereas GDP-Hier recovers all resolutions jointly in one run. Overall, GDP-Hier is the strongest hierarchy method, combining high partition accuracy with the lowest HVR on four of five benchmarks. On the most complex T200 and T608 hierarchies, it leads Mean ACC, Leaf ACC, and HVR among the constrained methods in Table 2. Full hierarchy results and additional analyses are in Appendix D.5, with efficiency analyses in Appendix E.

Component and mechanism analysis. We further examine component contributions and underlying mechanisms through targeted ablations, controlled experiments, and supporting diagnostics.

(i) Component ablations. Table 3 examines the main design choices in GDP. Replacing Ranking with BCE treats grades as calibrated numerical targets. The two are comparable when this assumption

Table 3: MNIST component ablation (without autoencoder pretraining; 5k constraints; five-run means). Test ACC/NMI/ARI (%); best/second-best per condition/metric are **bold**/underlined.

Variant	Binary	Diff. $b = 0.3$	Diff. $b = 0.5$	Diff. $b = 0.7$	Mix $\tau = 0.2$	Mix identity	Mix $\zeta = +3$
GDP (Ours)	**95.7 / 89.4 / 90.7**	**94.8 / 87.6 / 89.0**	**94.7 / 87.2 / 88.6**	**94.5 / 87.0 / 88.3**	**91.5 / 81.8 / 82.6**	93.3 / 84.9 / 86.0	**91.6 / 83.3 / 83.7**
Global BCE	95.5 / 89.1 / 90.4	94.0 / 86.0 / 87.3	93.3 / 84.8 / 86.1	93.0 / 84.0 / 85.2	90.0 / 81.3 / 81.1	**93.6 / 85.5 / 86.7**	87.5 / 78.5 / 77.5
w/o LGRC	63.9 / 65.5 / 59.5	63.8 / 56.2 / 51.9	56.4 / 46.3 / 42.7	44.9 / 33.3 / 28.0	47.5 / 35.8 / 28.0	72.1 / 64.2 / 61.1	68.1 / 58.3 / 53.9
w/o $\mathcal{L}_{\text{vertex}}$	92.1 / 82.7 / 83.7	90.5 / 80.6 / 80.9	90.4 / 80.5 / 80.0	88.2 / 78.0 / 75.1	89.6 / 78.4 / 79.0	91.7 / 81.9 / 82.8	89.7 / 78.3 / 79.1
w/o H_{marg}	95.4 / 89.1 / 90.3	92.7 / 85.9 / 86.1	94.3 / 86.8 / 88.0	82.7 / 81.5 / 74.8	90.8 / 80.6 / 81.2	91.0 / 82.9 / 83.1	83.2 / 75.9 / 72.1

matches the supervision, but Ranking is consistently stronger under Difficulty and Mix recalibration, with gains increasing as absolute grade values become less informative. LGRC yields large and consistent gains across conditions, while assignment-gated vertex calibration shows a broadly consistent positive contribution. Beyond the main ablations, marginal entropy prevents degenerate prototype use on an imbalanced hierarchy, while $\mathcal{L}_{\text{fit}}$ improves composite-relation calibration, especially under sparse supervision. Further component and mechanism analyses appear in Appendix D.6.

(ii) Controlled hierarchy recovery. To empirically examine the recovery factors in Theorem 5.4, we construct Synthetic-Hier, a three-level $4 \to 12 \to 48$ hierarchy of Gaussian clusters. Difficulty increases from VeryEasy to Hard as cluster centers move closer under fixed within-leaf variation, increasing overlap. We sample 20k constraints evenly across hierarchy grades, yielding matched settings for theory-aligned analysis. Figure 4 shows that recovery degrades as separability decreases. To quantify its relation to the theoretical condition, we use $\mathcal{E}_{95} = (\delta_{95} + 2\sum_{\ell} \Delta_{\ell}^{\beta} \epsilon_{\ell,95})/\Delta_{\min}^{\beta}$, where the uniform errors in Theorem 5.4 are replaced by their empirical 95th percentiles; smaller values indicate a more favorable recovery regime. Accordingly, larger $\mathcal{E}_{95}$ is associated with lower minimum training ACC across hierarchy levels ($\rho_s = -0.807$) and higher HVR ($\rho_s = 0.572$). Directly reducing $\Delta_{\min}^{\beta}$ from $1/3$ to 0.03 while keeping the generated data and sampled pairs matched likewise deteriorates recovery. Together, these results show that the sufficient-condition factors meaningfully track empirical hierarchy recoverability.

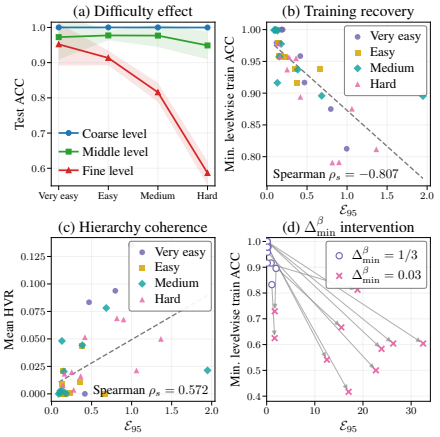

Figure 4: Synthetic-Hier. (a) Levelwise test ACC from VeryEasy to Hard (mean±std). (b,c) $\mathcal{E}_{95}$ versus Min. levelwise train ACC and HVR. (d) Matched $\Delta_{\min}^{\beta}$ intervention from $1/3$ to 0.03; arrows link paired runs.

(iii) Mechanism diagnostics and additional analyses. In Appendix D.6, we examine Local–Global partition agreement, assignment-gate behavior, and relation/ranking structure; these diagnostics support the intended Local–Global role separation and coordination. Separately, broad stability around the shared default configuration and learning efficiency are reported in Appendices D.7 and E.

7 Conclusion

Constrained clustering from graded pairwise supervision involves two distinct but related requirements: preserving graded relations and forming explicit cluster assignments. GDP addresses these requirements through separate but coordinated Local and Global representations, while GDP-Hier extends this principle across hierarchy resolutions. Our theory characterizes the relation–partition design and supports multi-resolution hierarchy recovery. Across diverse benchmarks, GDP achieves consistently strong and frequently leading clustering performance while remaining robust to grade recalibration, and GDP-Hier recovers accurate, coherent partitions across semantic resolutions; targeted mechanism analyses further support the intended roles of the core components.

The current framework has two main limitations. GDP assumes a predefined cluster count, and GDP-Hier additionally requires prescribed hierarchy resolutions and grade breakpoints, leaving automatic discovery of hierarchy structure outside the current scope. GDP-Hier also encourages cross-level coherence without explicitly enforcing nested partitions by construction. Automatic discovery of hierarchy structure from graded supervision is a natural direction for future work.

AI USE STATEMENT

We used generative AI tools for review and feedback on aspects of research methodology and experimental planning, for assistance with formalising and checking technical details in mathematical proofs, for implementation and debugging support, for literature and search assistance, for organising and presenting experimental results, and for language polishing and manuscript refinement. The research ideas and questions originated with the authors; the methodology, theoretical framework and objectives, experimental design and execution, interpretation of results, and scientific conclusions were determined and validated by the authors. All AI-assisted suggestions and outputs were critically reviewed and, where applicable, independently checked against mathematical derivations, code, experimental results, and cited sources by the authors, who take full responsibility for the work.

REPRODUCIBILITY STATEMENT

We provide the information and resources needed to reproduce our results throughout the paper, appendices, and supplementary materials. The proposed methods are specified in Section 4, with algorithmic details in Appendix B.1; the assumptions and complete proofs for the theoretical results are provided in Section 5 and Appendix A. Appendix C documents dataset construction and preprocessing, the compared methods, pairwise-supervision generation, experimental protocols, implementation details, hyperparameter settings, and repeated-run evaluation. Appendix D provides the detailed experimental results, including run-to-run variability, and Appendix E reports empirical runtime measurements. The supplementary materials provide access to the anonymous source code, experimental resources, and instructions required to reproduce the experiments.

REFERENCES

Mona Suliman AlZuhair, Mohamed Maher Ben Ismail, and Ouiem Bchir. Soft semi-supervised deep learning-based clustering. *Applied Sciences*, 13(17):9673, 2023. doi: 10.3390/app13179673.

Violaine Antoine, Benjamin Quost, M-H Masson, and Thierry Denoeux. CEVCLUS: evidential clustering with instance-level constraints for relational data. *Soft Computing*, 18(7):1321–1335, 2014.

Sugato Basu, Arindam Banerjee, and Raymond J Mooney. Active semi-supervision for pairwise constrained clustering. In *Proceedings of the 2004 SIAM International Conference on Data Mining*, pp. 333–344. SIAM, 2004.

Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A. Lord. Making better mistakes: Leveraging class hierarchies with deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12506–12515, 2020.

Mikhail Bilenko, Sugato Basu, and Raymond J Mooney. Integrating constraints and metric learning in semi-supervised clustering. In *Proceedings of the twenty-first International Conference on Machine Learning*, pp. 81–88, 2004.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9912–9924, 2020.

Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. Deep adaptive image clustering. In *Proceedings of the IEEE international conference on computer vision*, pp. 5879–5887, 2017.

Vaggos Chatziafratis, Rad Niazadeh, and Moses Charikar. Hierarchical clustering with structural constraints. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 774–783. PMLR, 2018.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

Ian Davidson and Sugato Basu. A survey of clustering with instance level constraints. *ACM Transactions on Knowledge Discovery from Data*, 1(1):1–41, 2007. doi: 10.1145/1217299.1217300.

Ian Davidson and S. S. Ravi. Using instance-level constraints in agglomerative hierarchical clustering: Theoretical and empirical results. *Data Mining and Knowledge Discovery*, 18(2):257–282, 2009. doi: 10.1007/s10618-008-0103-4.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009.

Irene Diaz-Valenzuela, M Amparo Vila, and Maria J Martin-Bautista. On the use of fuzzy constraints in semisupervised clustering. *IEEE Transactions on Fuzzy Systems*, 24(4):992–999, 2016.

Sharon Fogel, Hadar Averbuch-Elor, Daniel Cohen-Or, and Jacob Goldberger. Clustering-driven deep embedding with pairwise constraints. *IEEE computer graphics and applications*, 39(4):16–27, 2019.

Debarghya Ghoshdastidar, Michaël Perrot, and Ulrike von Luxburg. Foundations of comparison-based hierarchical clustering. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Germán González-Almagro, Daniel Peralta, Eli De Poorter, José-Ramón Cano, and Salvador García. Semi-supervised constrained clustering: an in-depth overview, ranked taxonomy and future research directions. *Artificial Intelligence Review*, 58:157, 2025.

Jann Goschenhofer, Bernd Bischl, and Zsolt Kira. ConstraintMatch for semi-constrained clustering. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–10. IEEE, 2023.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

Xifeng Guo, Long Gao, Xinwang Liu, and Jianping Yin. Improved deep embedded clustering with local structure preservation. In *IJCAI*, volume 17, pp. 1753–1759, 2017.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, and Zsolt Kira. A probabilistic constrained clustering for transfer learning and image category discovery. In *Proceedings of the CVPR 2018 Deep-Vision Workshop*, 2018.

Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, and Zsolt Kira. Multi-class classification without multi-class labels. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SJzR2iRcK7.

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

Brian Kulis, Sugato Basu, Inderjit Dhillon, and Raymond Mooney. Semi-supervised graph clustering: a kernel approach. In *Proceedings of the 22nd International Conference on Machine Learning*, pp. 457–464, 2005.

Martin HC Law, Alexander Topchy, and Anil K Jain. Clustering with soft and group constraints. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pp. 662–670. Springer, 2004.

Martin HC Law, Alexander Topchy, and Anil K Jain. Model-based clustering with probabilistic constraints. In *Proceedings of the 2005 SIAM International Conference on Data Mining*, pp. 641–645. SIAM, 2005.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

Feng Li, Shoumei Li, and Thierry Denœux. k-CEVCLUS: Constrained evidential clustering of large dissimilarity data. *Knowledge-Based Systems*, 142:29–44, 2018.

Zhengdong Lu and Todd Leen. Semi-supervised learning with penalized probabilistic clustering. *Advances in Neural Information Processing Systems*, 17, 2004.

Yucen Luo, Tian Tian, Jiaxin Shi, Jun Zhu, and Bo Zhang. Semi-crowdsourced clustering with deep generative models. *Advances in Neural Information Processing Systems*, 31, 2018.

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

Laura Manduchi, Kieran Chin-Cheong, Holger Michel, Sven Wellmann, and Julia Vogt. Deep conditional Gaussian mixture model for constrained clustering. *Advances in Neural Information Processing Systems*, 34:11303–11314, 2021.

Jian-Ping Mei and Huajiang Lv. Semi-supervised fuzzy c-means regularized with pairwise constraints. In *2018 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, pp. 781–786. IEEE, 2018.

Jian-Ping Mei, Huajiang Lv, Jiuwen Cao, and Weihua Gong. Pairwise constrained fuzzy clustering: Relation, comparison and parallelization. *International Journal of Fuzzy Systems*, 21(6):1938–1949, 2019.

George A. Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38(11): 39–41, 1995. doi: 10.1145/219717.219748.

Tri Nguyen, Shahana Ibrahim, and Xiao Fu. Deep clustering with incomplete noisy pairwise annotations: A geometric regularization approach. In *International Conference on Machine Learning*, pp. 25980–26007. PMLR, 2023.

Abu Quwsar Ohi, Muhammad Firoz Mridha, Farisa Benta Safir, Md Abdul Hamid, and Muhammad Mostafa Monowar. AutoEmbedder: A semi-supervised DNN embedding system for clustering. *Knowledge-Based Systems*, 204:106190, 2020.

Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9617–9626, 2019.

Yazhou Ren, Kangrong Hu, Xinyi Dai, Lili Pan, Steven CH Hoi, and Zenglin Xu. Semi-supervised deep embedded clustering. *Neurocomputing*, 325:121–130, 2019.

Marek Śmieja, Łukasz Struski, and Mário AT Figueiredo. A classification-based approach to semi-supervised clustering with pairwise constraints. *Neural Networks*, 127:193–203, 2020.

Tian Tian, Jie Zhang, Xiang Lin, Zhi Wei, and Hakon Hakonarson. Model-based deep embedding for constrained clustering analysis of single cell RNA-seq data. *Nature Communications*, 12(1):1873, 2021.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine learning*, pp. 1096–1103. ACM, 2008.

Kiri Wagstaff and Claire Cardie. Clustering with instance-level constraints. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 1103–1110. Morgan Kaufmann, 2000.

Kiri Wagstaff, Claire Cardie, Seth Rogers, and Stefan Schrödl. Constrained k-means clustering with background knowledge. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 577–584. Morgan Kaufmann, 2001.

Xiang Wang and Ian Davidson. Flexible constrained spectral clustering. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 563–572. ACM, 2010. doi: 10.1145/1835804.1835877.

Zhen Wang, Shan-Shan Wang, Lan Bai, Wen-Si Wang, and Yuan-Hai Shao. Semisupervised fuzzy clustering with fuzzy pairwise constraints. *IEEE Transactions on Fuzzy Systems*, 30(9):3797–3811, 2022.

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

Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International Conference on Machine Learning*, pp. 478–487. PMLR, 2016.

Hongjing Zhang, Tianyang Zhan, Sugato Basu, and Ian Davidson. A framework for deep constrained clustering. *Data Mining and Knowledge Discovery*, 35:593–620, 2021.

Shaojie Zhang and Ke Chen. Angular constraint embedding via SpherePair loss for constrained clustering. *Advances in Neural Information Processing Systems*, 38:129707–129761, 2025.

Shaojie Zhang and Ke Chen. Uncertainty-aware probabilistic constrained clustering from entangled pairwise supervision. *arXiv preprint arXiv:2608.12027*, 2026.

Lin Zhao, Yuan Yuan, and Feiping Nie. Efficient hierarchical multi-resolution k-means clustering. *Pattern Recognition*, 174:112930, 2026. doi: 10.1016/j.patcog.2025.112930.

Junjing Zheng, Xinyu Zhang, Xiangfeng Qiu, Chengliang Song, and Weidong Jiang. Semi-supervised hyperbolic hierarchical clustering with set-level structural priors. *arXiv preprint arXiv:2606.01525*, 2026.

Li Zheng and Tao Li. Semi-supervised hierarchical clustering. In *2011 IEEE 11th International Conference on Data Mining*, pp. 982–991. IEEE, 2011. doi: 10.1109/ICDM.2011.130.

A Additional theoretical details

We provide complete proofs for the main theoretical results in Section 5, together with formal statements for the two auxiliary results used there.

A.1 Proof of Lemma 5.1

Proof. Let $q_a, q_b \in \Delta^{K-1}$ be the Global assignments of two samples, with $r_{ab}^{\mathrm{glob}} = q_a^\top q_b$. First suppose that the samples belong to the same latent cluster, and let $k^\star$ be the prototype aligned with that cluster. By assumption, $q_{a,k^\star} \geq 1-\epsilon$ and $q_{b,k^\star} \geq 1-\epsilon$. Hence

$$r_{ab}^{\mathrm{glob}} = \sum_{k=1}^{K} q_{a,k} q_{b,k} \geq q_{a,k^\star} q_{b,k^\star} \geq (1-\epsilon)^2.$$

Now suppose that the samples belong to different latent clusters, aligned with distinct prototypes $k_a^\star$ and $k_b^\star$. Since $q_{b,k_b^\star} \geq 1-\epsilon$, the total mass of q_b outside $k_b^\star$ is at most ϵ, and therefore $q_{b,k_a^\star} \leq \epsilon$. Similarly, $\sum_{k \neq k_a^\star} q_{a,k} \leq \epsilon$. Thus

$$\begin{aligned} r_{ab}^{\mathrm{glob}} &= q_{a,k_a^\star} q_{b,k_a^\star} + \sum_{k \neq k_a^\star} q_{a,k} q_{b,k} \\ &\leq q_{b,k_a^\star} + \sum_{k \neq k_a^\star} q_{a,k} \leq 2\epsilon. \end{aligned}$$

This establishes the two claimed bounds. □

A.2 Partition compatibility of zero-margin Ranking

Lemma A.1 (Partition compatibility of zero-margin Ranking)**.** *Let $h_i \in \{0,1\}$ denote latent hard co-membership for constraint i, and let $\mathcal{P} = \{(i,i') : g_i > g_{i'}\}$ be nonempty. If $g_i > g_{i'}$ implies $h_i \geq h_{i'}$, then the hard co-membership vector $\mathbf{h} = (h_i)_{i=1}^{|\mathcal{C}|}$ incurs zero loss under the zero-margin Ranking objective in Eq.* (5)*. In particular, compared constraints with distinct grades but equal hard co-membership incur no penalty. Moreover, any strictly increasing transformation of the grades leaves $\mathcal{P}$, and hence the zero-margin Ranking objective, unchanged.*

Proof. Evaluating Eq. (5) at the hard co-membership vector gives

$$\mathcal{L}_{\mathrm{rank}}(\mathbf{h}) = \frac{1}{|\mathcal{P}|} \sum_{(i,i')\in\mathcal{P}} [h_{i'} - h_i]_+.$$

For every $(i,i') \in \mathcal{P}$, the assumption gives $h_i \geq h_{i'}$, so each summand is zero and therefore $\mathcal{L}_{\mathrm{rank}}(\mathbf{h}) = 0$. This includes the case $h_i = h_{i'}$, for which the corresponding summand remains zero.

If φ is strictly increasing, then $g_i > g_{i'}$ if and only if $\varphi(g_i) > \varphi(g_{i'})$. The comparison set $\mathcal{P}$ is therefore unchanged, and so is the zero-margin Ranking objective. □

A.3 Proof of Theorem 5.2

Proof. As $\lambda \to \infty$, the ranking penalty suppresses order violations, and the quadratic criterion selects the feasible relation vector closest to $\mathbf{r}^{\mathrm{loc}}$ in the limit. For the fixed comparison set $\mathcal{P}$, define the relation-level order violation

$$V_{\mathcal{P}}(\mathbf{r}) := \frac{1}{|\mathcal{P}|} \sum_{(i,i')\in\mathcal{P}} [r_{i'} - r_i]_+.$$

Thus $V_{\mathcal{P}}(\mathbf{r}) = 0$ exactly when $\mathbf{r} \in \mathcal{R}_g$, and the objective in the theorem can be written as

$$F_\lambda(\mathbf{r}) = \frac{1}{2}\|\mathbf{r} - \mathbf{r}^{\mathrm{loc}}\|_2^2 + \lambda V_{\mathcal{P}}(\mathbf{r}), \qquad \mathbf{r} \in [0,1]^{|\mathcal{C}|}.$$

The set $\mathcal{R}_g$ is nonempty because every constant relation vector is feasible. It is closed and convex, being the intersection of $[0,1]^{|\mathcal{C}|}$ with finitely many linear half-spaces, and is therefore compact. Hence the Euclidean projection $\bar{\mathbf{r}}$ exists and is unique. For $\lambda > 0$, F_λ is continuous and strictly convex on the compact box $[0,1]^{|\mathcal{C}|}$, so its minimizer $\mathbf{r}_\lambda$ exists and is unique.

Because $\bar{\mathbf{r}} \in \mathcal{R}_g$, we have $V_{\mathcal{P}}(\bar{\mathbf{r}}) = 0$. Optimality of $\mathbf{r}_\lambda$ gives

$$F_\lambda(\mathbf{r}_\lambda) \leq F_\lambda(\bar{\mathbf{r}}) = \frac{1}{2}\|\bar{\mathbf{r}} - \mathbf{r}^{\mathrm{loc}}\|_2^2.$$

Consequently,

$$V_{\mathcal{P}}(\mathbf{r}_\lambda) \leq \frac{\|\bar{\mathbf{r}} - \mathbf{r}^{\mathrm{loc}}\|_2^2}{2\lambda} \longrightarrow 0 \qquad \text{as } \lambda \to \infty.$$

Suppose, for contradiction, that $\mathbf{r}_\lambda$ does not converge to $\bar{\mathbf{r}}$ as $\lambda \to \infty$. Then there exist $\varepsilon_0 > 0$ and a sequence $\lambda_t \to \infty$ such that $\|\mathbf{r}_{\lambda_t} - \bar{\mathbf{r}}\|_2 \geq \varepsilon_0$ for all t. By compactness, a subsequence satisfies $\mathbf{r}_{\lambda_{t_j}} \to \widetilde{\mathbf{r}}$. Continuity of $V_{\mathcal{P}}$ gives $\widetilde{\mathbf{r}} \in \mathcal{R}_g$, and

$$\frac{1}{2}\|\mathbf{r}_{\lambda_{t_j}} - \mathbf{r}^{\mathrm{loc}}\|_2^2 \leq F_{\lambda_{t_j}}(\mathbf{r}_{\lambda_{t_j}}) \leq \frac{1}{2}\|\bar{\mathbf{r}} - \mathbf{r}^{\mathrm{loc}}\|_2^2.$$

Taking $j \to \infty$ shows that $\widetilde{\mathbf{r}}$ is no farther from $\mathbf{r}^{\mathrm{loc}}$ than $\bar{\mathbf{r}}$. By uniqueness of the Euclidean projection, $\widetilde{\mathbf{r}} = \bar{\mathbf{r}}$, contradicting $\|\mathbf{r}_{\lambda_{t_j}} - \bar{\mathbf{r}}\|_2 \geq \varepsilon_0$. Therefore $\mathbf{r}_\lambda \to \bar{\mathbf{r}}$.

Finally, if $\mathbf{r}^{\mathrm{loc}} \in \mathcal{R}_g$, then its Euclidean projection onto $\mathcal{R}_g$ is itself, so $\bar{\mathbf{r}} = \mathbf{r}^{\mathrm{loc}}$. Indeed, in this case $F_\lambda(\mathbf{r}^{\mathrm{loc}}) = 0$, and therefore $\mathbf{r}_\lambda = \mathbf{r}^{\mathrm{loc}}$ for every $\lambda > 0$. □

A.4 ASSIGNMENT-ADAPTIVE SIMPLEX CALIBRATION

Proposition A.2 (Assignment-adaptive simplex calibration)**.** *Let $\{w_k\}_{k=1}^K$ be the unit regular-simplex prototypes in Section 4.2, let $q \in \Delta^{K-1}$, and define $s(q) := \sum_{k=1}^K q_k w_k$ and $c(q) := (K\sum_{k=1}^K q_k^2 - 1)/(K-1)$. Then $c(q) = \|s(q)\|_2^2$. For the per-sample calibration term*

$$\ell_{\text{vertex}}(u; q) := c(q) \sum_{k=1}^K q_k \|u - w_k\|_2^2,$$

treating q as detached gives $\nabla_u \ell_{\text{vertex}}(u; q) = 2c(q)(u - s(q))$.

Proof. For unit regular-simplex prototypes, $w_k^\top w_{k'} = -1/(K-1)$ whenever $k \neq k'$. Using $\sum_k q_k = 1$,

$$\begin{aligned}
\|s(q)\|_2^2 &= \left\| \sum_{k=1}^K q_k w_k \right\|_2^2 \\
&= \sum_{k=1}^K q_k^2 - \frac{1}{K-1} \sum_{k \neq k'} q_k q_{k'} \\
&= \sum_{k=1}^K q_k^2 - \frac{1 - \sum_{k=1}^K q_k^2}{K-1} \\
&= \frac{K\sum_{k=1}^K q_k^2 - 1}{K-1} = c(q).
\end{aligned}$$

Because the assignments are detached in Eq. (6), q, $c(q)$, and $s(q)$ are fixed when differentiating with respect to u. Therefore,

$$\begin{aligned}
\nabla_u \ell_{\text{vertex}}(u; q) &= 2c(q) \sum_{k=1}^K q_k (u - w_k) \\
&= 2c(q) \left(u - \sum_{k=1}^K q_k w_k \right) \\
&= 2c(q)(u - s(q)).
\end{aligned}$$

□

A.5 PROOF OF PROPOSITION 5.3

Proof. Fix a pair (a, b), and let $\ell^\star \in \{0, \ldots, L\}$ denote its deepest shared hierarchy level, with $\ell^\star = 0$ when the pair already differs at the coarsest level, consistently with Eq. (1). By nestedness,

$$h_{ab}^{(\ell)} = \begin{cases} 1, & \ell \leq \ell^\star, \\ 0, & \ell > \ell^\star. \end{cases}$$

Hence

$$\begin{aligned}
\sum_{\ell=1}^L \Delta_\ell^\beta h_{ab}^{(\ell)} &= \sum_{\ell=1}^{\ell^\star} (\beta_\ell - \beta_{\ell-1}) \\
&= \beta_{\ell^\star} - \beta_0 = \beta_{\ell^\star} = g_{ab},
\end{aligned}$$

where the empty sum for $\ell^\star = 0$ equals zero.

For uniqueness, suppose that additive weights $a_1, \ldots, a_L$ reproduce every hierarchy breakpoint. For the nested indicator pattern associated with breakpoint β_ℓ, reproduction of that breakpoint requires

$$\sum_{j=1}^\ell a_j = \beta_\ell, \qquad \ell = 1, \ldots, L.$$

The case $\ell = 1$ gives $a_1 = \beta_1$. For each $\ell \geq 2$, subtracting the equality at level $\ell - 1$ from that at level ℓ yields $a_\ell = \beta_\ell - \beta_{\ell-1} = \Delta_\ell^\beta$. Thus the breakpoint increments are the unique additive weights reproducing all hierarchy breakpoints. □

A.6 Proof of Theorem 5.4

Proof. For each level ℓ, define the predicted argmax label and its induced hard co-membership by

$$\widehat{k}_a^{(\ell)} := \arg\max_k q_{a,k}^{(\ell)}, \qquad \widehat{h}_{ab}^{(\ell)} := \mathbf{1}\left[\widehat{k}_a^{(\ell)} = \widehat{k}_b^{(\ell)}\right].$$

The proof proceeds by controlling the soft-to-hard relation gap, establishing separation of the induced composite relation, and then recovering the hierarchy from the finest level toward coarser levels.

Soft-to-hard relation control. If $\widehat{k}_a^{(\ell)} = \widehat{k}_b^{(\ell)} = k^\star$, then $q_{a,k^\star}^{(\ell)}, q_{b,k^\star}^{(\ell)} \geq 1 - \epsilon_\ell$, so $r_{ab}^{(\ell)} \geq (1-\epsilon_\ell)^2$. Since $\widehat{h}_{ab}^{(\ell)} = 1$,

$$\left| r_{ab}^{(\ell)} - \widehat{h}_{ab}^{(\ell)} \right| \leq 1 - (1-\epsilon_\ell)^2 \leq 2\epsilon_\ell.$$

If the argmax labels are different, write $k_a = \widehat{k}_a^{(\ell)}$ and $k_b = \widehat{k}_b^{(\ell)}$ with $k_a \neq k_b$. Since $q_{b,k_b}^{(\ell)} \geq 1 - \epsilon_\ell$, we have $q_{b,k_a}^{(\ell)} \leq \epsilon_\ell$; likewise, $\sum_{k \neq k_a} q_{a,k}^{(\ell)} \leq \epsilon_\ell$. Therefore,

$$\begin{aligned} r_{ab}^{(\ell)} &= q_{a,k_a}^{(\ell)} q_{b,k_a}^{(\ell)} + \sum_{k \neq k_a} q_{a,k}^{(\ell)} q_{b,k}^{(\ell)} \\ &\leq q_{b,k_a}^{(\ell)} + \sum_{k \neq k_a} q_{a,k}^{(\ell)} \leq 2\epsilon_\ell. \end{aligned}$$

Here $\widehat{h}_{ab}^{(\ell)} = 0$, so the same bound follows. Thus, for every level,

$$\left| r_{ab}^{(\ell)} - \widehat{h}_{ab}^{(\ell)} \right| \leq 2\epsilon_\ell. \tag{12}$$

Composite separation. Define the argmax-induced hard composite relation

$$\widehat{g}_{ab} := \sum_{\ell=1}^{L} \Delta_\ell^\beta \widehat{h}_{ab}^{(\ell)}.$$

By Eqs. (10) and (12) and the triangle inequality,

$$\left| r_{ab}^{\mathrm{hier}} - \widehat{g}_{ab} \right| \leq 2 \sum_{\ell=1}^{L} \Delta_\ell^\beta \epsilon_\ell.$$

On every connecting pair in the theorem, $|r_{ab}^{\mathrm{hier}} - g_{ab}| \leq \delta$; hence

$$|\widehat{g}_{ab} - g_{ab}| \leq \delta + 2 \sum_{\ell=1}^{L} \Delta_\ell^\beta \epsilon_\ell < \Delta_{\min}^\beta. \tag{13}$$

Finest-level recovery. The composite separation in Eq. (13) permits partition recovery by descending induction from level L to level 1. At the finest level, fix a true cluster $S \in \mathcal{S}^{(L)}$ and consider the graph on its samples whose edges are the observed within-cluster pairs used in the theorem; this graph is connected by assumption. Every such edge (a, b) has $g_{ab} = \beta_L = 1$. If the predicted finest-level partition separated its endpoints, then $\widehat{h}_{ab}^{(L)} = 0$; since every other hard co-membership is at most one,

$$\widehat{g}_{ab} \leq \sum_{j=1}^{L-1} \Delta_j^\beta = 1 - \Delta_L^\beta.$$

Thus $|\widehat{g}_{ab} - g_{ab}| \geq \Delta_L^\beta \geq \Delta_{\min}^\beta$, contradicting Eq. (13). Therefore every edge of the connected graph has equal predicted level-L labels at its endpoints, so all samples in S receive the same predicted label. Each true finest-level cluster is consequently contained in one predicted cluster. This containment defines a map from true clusters to predicted clusters. The map is surjective because every predicted cluster is nonempty: any sample in it belongs to a true cluster, which by containment maps to that predicted cluster. Since both partitions contain exactly $K^{(L)}$ clusters, the map is bijective. Hence the finest-level partition is recovered up to label permutation.

Induction to coarser levels. For the induction step, suppose that the predicted partitions have been recovered at every level $j > \ell$. Fix a true level-ℓ cluster $S \in \mathcal{S}^{(\ell)}$ and an edge (a, b) in its connected within-cluster graph. Let $\ell^\star \geq \ell$ be the deepest level shared by the pair, so $g_{ab} = \beta_{\ell^\star}$. By the induction hypothesis, $\widehat{h}_{ab}^{(j)} = h_{ab}^{(j)}$ for all $j > \ell$, and therefore the contribution of the finer levels to $\widehat{g}_{ab}$ is

$$\sum_{j=\ell+1}^{L} \Delta_j^\beta \widehat{h}_{ab}^{(j)} = \sum_{j=\ell+1}^{\ell^\star} \Delta_j^\beta = \beta_{\ell^\star} - \beta_\ell.$$

Suppose that the predicted level-ℓ partition separated a and b, so $\widehat{h}_{ab}^{(\ell)} = 0$. The total contribution from the coarser levels is at most $\sum_{j=1}^{\ell-1} \Delta_j^\beta = \beta_{\ell-1}$, giving

$$\widehat{g}_{ab} \leq \beta_{\ell-1} + \beta_{\ell^\star} - \beta_\ell = \beta_{\ell^\star} - \Delta_\ell^\beta = g_{ab} - \Delta_\ell^\beta.$$

Hence $|\widehat{g}_{ab} - g_{ab}| \geq \Delta_\ell^\beta \geq \Delta_{\min}^\beta$, again contradicting Eq. (13). Thus every edge in the connected graph of S has equal predicted level-ℓ labels at its endpoints, and connectivity places the entire true cluster S inside one predicted level-ℓ cluster. The same nonempty-cluster counting argument gives a bijection because both partitions contain $K^{(\ell)}$ clusters. Therefore the level-ℓ partition is recovered up to label permutation. Descending induction over $\ell = L, L-1, \ldots, 1$ proves recovery at every hierarchy level. □

B Algorithm and complexity analysis

GDP and GDP-Hier share a compact Local–Global training structure, with the hierarchy extension reusing the same backbone across multiple resolutions; the corresponding algorithms and computational costs are detailed below.

B.1 Algorithm

The training procedures of GDP and GDP-Hier are summarized in Algorithms 1 and 2, respectively. The Global computations in these algorithms use two auxiliary operations defined below.

For a K-way Global head, let $\theta_K = \arccos(-1/(K-1))$ be the angle between regular-simplex prototypes and let $\mathcal{W} = \mathrm{span}\{w_1, \ldots, w_K\}$. Simplex-angle rescaling rescales the angle between a Global feature and a prototype relative to the regular-simplex separation θ_K, so that the regular-simplex inter-prototype angle corresponds to π; the feature norm separately controls the sharpness of the resulting soft assignment. Projection onto the prototype span removes components orthogonal to this span, restricting vertex calibration to the geometry represented by the simplex prototypes. We use

$$\widetilde{\theta}_K(\theta) = \min\left\{ \frac{\pi}{\theta_K}\theta,\, \pi \right\}, \qquad \Pi_{\mathcal{W}}(z) = \mathrm{Proj}_{\mathcal{W}}(z). \tag{14}$$

Accordingly, the assignment score in Eq. (4) is $\rho(z, w_k) = \|z\|_2 \cos\Big(\widetilde{\theta}_K\Big(\arccos\frac{z^\top w_k}{\|z\|_2}\Big)\Big)$. Before evaluating $\arccos$, its normalized inner-product argument is clamped to $[-1 + 10^{-7}, 1 - 10^{-7}]$ for numerical stability. Each w_k is unit-normalized, and the projected representation in Eq. (6) is $u_j = \Pi_{\mathcal{W}}(z_j^{\mathrm{glob}})/R$ for learnable radius $R > 0$.

B.2 Computational complexity analysis

The dominant computation of GDP arises from Local pairwise learning, reconstruction, and Global relation evaluation, while GDP-Hier extends the Global computation across multiple resolutions using the same Local backbone.

GDP. Let κ_{pair} denote the amortized cost of Local pairwise relation learning for one pair, and let κ_{rec} denote the reconstruction cost for one instance. For a K-way Global head, let $\kappa_{\mathrm{G}}(K)$ denote the amortized per-instance cost of the Global mapping, prototype assignment, and vertex calibration. Constant factors arising from evaluating the two instances of a pair are absorbed into these quantities.

Algorithm 1 GDP for graded pairwise constrained clustering

Require: training data $\mathcal{X}$; graded constraints $\mathcal{C} = \{(a_i, b_i, g_i)\}_{i=1}^{|\mathcal{C}|}$; number of clusters K; total training epochs N_{ep}; Global update interval I_{glob}; loss weights $\lambda_{\mathrm{rec}}, \alpha, \gamma, \eta$

Initialize $f_{\phi_{\mathrm{loc}}}$, $f_{\phi_{\mathrm{glob}}}$, d_ψ, the ProbPair parameters (m, T), and radius R
Construct and fix K unit regular-simplex prototypes $\{w_k\}_{k=1}^{K}$
for $n = 1, \ldots, N_{\mathrm{ep}}$ **do**
  for each paired constraint/data mini-batch $(\mathcal{B}_\mathcal{C}, \mathcal{B}_\mathcal{X})$ **do**
    Sample additional pairs from $\mathcal{X}$ and augment $\mathcal{B}_\mathcal{C}$ to form $\mathcal{B}_{\mathrm{LG}}$ as in Eq. (7)
    Compute $z_j^{\mathrm{loc}} = f_{\phi_{\mathrm{loc}}}(x_j)$ for all instances used in $\mathcal{B}_{\mathrm{LG}}$ and $\mathcal{B}_\mathcal{X}$
    Evaluate $\mathcal{L}_{\mathrm{PP}}$ by Eq. (3) on $\mathcal{B}_\mathcal{C}$ and $\mathcal{L}_{\mathrm{rec}}$ on $\mathcal{B}_\mathcal{X}$; set $\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{PP}} + \lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{rec}}$ as in Eq. (8)
    if n is not a scheduled Global update under interval I_{glob} **then**
      Detach the Local representations passed to the Global path
    end if
    Evaluate r_{ab}^{loc} and r_{ab}^{glob} by Eqs. (2) and (4) on $\mathcal{B}_{\mathrm{LG}}$, and compute $\mathcal{L}_{\mathrm{LGRC}}$ by Eq. (7) with $\mathrm{sg}[r_{ab}^{\mathrm{loc}}]$
    $\mathcal{L} \leftarrow \mathcal{L} + \eta\mathcal{L}_{\mathrm{LGRC}}$
    if n is a scheduled Global update under interval I_{glob} **then**
      Compute z_j^{glob} and q_j for the instances used in $\mathcal{B}_\mathcal{C}$ and $\mathcal{B}_\mathcal{X}$ by Eq. (4)
      Construct comparisons (i, i') within $\mathcal{B}_\mathcal{C}$ with $g_i > g_{i'}$, sample $\widetilde{\mathcal{P}}$, and compute $\mathcal{L}_{\mathrm{rank}}$ by Eq. (5)
      Compute H_{marg} from the current Global assignments
      Compute $u_j = \Pi_\mathcal{W}(z_j^{\mathrm{glob}})/R$ and $\mathcal{L}_{\mathrm{vertex}}$ by Eq. (6)
      $\mathcal{L} \leftarrow \mathcal{L} + \mathcal{L}_{\mathrm{rank}} - \alpha H_{\mathrm{marg}} + \gamma\mathcal{L}_{\mathrm{vertex}}$
    end if
    Update all trainable parameters by gradient descent on $\mathcal{L}$
  end for
end for
Compute $\{q_j\}_{j=1}^{|\mathcal{X}|}$ by Eq. (4) and obtain $\mathcal{S} = \{S_k\}_{k=1}^{K}$ by dominant prototype assignments

Over one epoch, the ProbPair and reconstruction objectives cost $O\big(|\mathcal{C}|\kappa_{\mathrm{pair}} + |\mathcal{X}|\kappa_{\mathrm{rec}}\big)$. LGRC additionally evaluates Local and Global relations on the constrained pairs together with sampled pairs. Since the number of sampled pairs grows at most linearly with the constraint budget, this contributes $O\big(|\mathcal{C}|(\kappa_{\mathrm{pair}} + \kappa_{\mathrm{G}}(K))\big)$. The remaining Global objectives are evaluated once every I_{glob} epochs, giving an amortized per-epoch cost $O\big((|\mathcal{C}| + |\mathcal{X}|)\kappa_{\mathrm{G}}(K)/I_{\mathrm{glob}}\big)$. Ranking comparisons are formed within each mini batch and subsampled, so a fixed comparison budget avoids a quadratic dependence on the full constraint set.

The overall training complexity over N_{ep} epochs is therefore

$$O\left(N_{\mathrm{ep}}\left(|\mathcal{C}|\kappa_{\mathrm{pair}} + |\mathcal{X}|\kappa_{\mathrm{rec}} + \left(|\mathcal{C}| + \frac{|\mathcal{C}| + |\mathcal{X}|}{I_{\mathrm{glob}}}\right)\kappa_{\mathrm{G}}(K)\right)\right).$$

For fixed model dimensions, K, I_{glob}, and sampling sizes, GDP therefore scales linearly with both $|\mathcal{C}|$ and $|\mathcal{X}|$. The additional computation is concentrated in the Global relation evaluations required by LGRC and the periodic Global partition updates, increasing the amount of computation per epoch without introducing a higher order dependence on the dataset or constraint-set size.

GDP-Hier. GDP-Hier retains the same Local encoder and decoder and replaces the single Global head with L resolution-specific heads. The corresponding Global computation is therefore governed by $\sum_{\ell=1}^{L}\kappa_{\mathrm{G}}\big(K^{(\ell)}\big)$. Forming the composite hierarchy relation requires $O(L|\mathcal{C}|)$ scalar operations over the pair set, while hierarchy Ranking and the grade fitting loss $\mathcal{L}_{\mathrm{fit}}$ reuse this composite relation at scheduled Global updates.

The resulting training complexity is

$$O\left(N_{\mathrm{ep}}\left(|\mathcal{C}|\kappa_{\mathrm{pair}} + |\mathcal{X}|\kappa_{\mathrm{rec}} + \left(|\mathcal{C}| + \frac{|\mathcal{C}| + |\mathcal{X}|}{I_{\mathrm{glob}}}\right)\sum_{\ell=1}^{L}\kappa_{\mathrm{G}}\Big(K^{(\ell)}\Big) + L|\mathcal{C}|\right)\right).$$

Algorithm 2 GDP-Hier extension for hierarchy-induced relations

Require: GDP procedure in Algorithm 1; hierarchy breakpoints $\{\beta_\ell\}_{\ell=0}^{L}$; levelwise cluster numbers $\{K^{(\ell)}\}_{\ell=1}^{L}$

Retain the Local computations and pair sampling of Algorithm 1, with the same Global update and detach/feedback schedule across all hierarchy heads
Replace the single Global head by L resolution-specific heads with $K^{(\ell)}$ fixed simplex prototypes at level ℓ
for each mini-batch update of Algorithm 1 **do**
    for $\ell = 1, \dots, L$ **do**
        Compute $q_j^{(\ell)}$ by Eq. (4) for all instances required by the current Global path
        $r_{ab}^{(\ell)} \leftarrow q_a^{(\ell)\top} q_b^{(\ell)}, \quad (a,b) \in \mathcal{B}_{\mathrm{LG}}$
    end for
    $r_{ab}^{\mathrm{hier}} \leftarrow \sum_{\ell=1}^{L} (\beta_\ell - \beta_{\ell-1}) r_{ab}^{(\ell)}, \quad (a,b) \in \mathcal{B}_{\mathrm{LG}}$
    Compute $\mathcal{L}_{\mathrm{LGRC}}^{\mathrm{hier}}$ by Eq. (7) using r_{ab}^{hier} as the Global relation
    $\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{loc}} + \eta \mathcal{L}_{\mathrm{LGRC}}^{\mathrm{hier}}$
    if Global update is scheduled **then**
        Compute $\mathcal{L}_{\mathrm{rank}}^{\mathrm{hier}}$ by Eq. (5) and $\mathcal{L}_{\mathrm{fit}}$ on $\mathcal{B}_{\mathcal{C}}$ from the composite relations
        for $\ell = 1, \dots, L$ **do**
            Compute $H_{\mathrm{marg}}^{(\ell)}$ and $\mathcal{L}_{\mathrm{vertex}}^{(\ell)}$ for Global head ℓ
        end for
        $\overline{H}_{\mathrm{marg}} \leftarrow L^{-1} \sum_{\ell=1}^{L} H_{\mathrm{marg}}^{(\ell)}, \quad \overline{\mathcal{L}}_{\mathrm{vertex}} \leftarrow L^{-1} \sum_{\ell=1}^{L} \mathcal{L}_{\mathrm{vertex}}^{(\ell)}$
        $\mathcal{L} \leftarrow \mathcal{L} + \mathcal{L}_{\mathrm{rank}}^{\mathrm{hier}} + \mathcal{L}_{\mathrm{fit}} - \alpha \overline{H}_{\mathrm{marg}} + \gamma \overline{\mathcal{L}}_{\mathrm{vertex}}$
    end if
    Update all trainable parameters by gradient descent on $\mathcal{L}$
end for
Obtain $\{\mathcal{S}^{(\ell)}\}_{\ell=1}^{L}$ from the dominant assignments of the corresponding Global heads

For fixed L, model dimensions, levelwise cluster numbers, I_{glob}, and sampling sizes, GDP-Hier likewise remains linear in $|\mathcal{C}|$ and $|\mathcal{X}|$. Its hierarchy-specific computation scales with the collection of resolution-specific Global heads, while the Local encoder and decoder remain shared across all resolutions.

Practical computational profile. Beyond the asymptotic analysis above, GDP retains a compact network-level training structure: the Local and Global representations share the same encoder backbone, with the Global representation obtained through a lightweight mapping. Its backbone-level computation therefore remains of the same order as standard encoder-based constrained clustering methods such as SpherePair and ProbPair (Zhang & Chen, 2025; 2026), while avoiding the repeated backbone fitting involved in the more elaborate training protocol of ECI-PP (Zhang & Chen, 2026). GDP also does not require the extensive hyperparameter search used by methods such as VolMaxDCC (Nguyen et al., 2023). GDP-Hier further amortizes training across resolutions by producing all prescribed hierarchy partitions within a single run, whereas a single-resolution end-to-end method such as CIDEC (Zhang et al., 2021) must be trained separately for different target cluster numbers. Finally, Appendix D.1 shows that GDP retains strong performance without autoencoder pretraining, indicating limited dependence on autoencoder initialization. Overall wall-clock training costs are reported in Appendix E.

C Details of experimental settings

We provide supplementary information for the experimental settings in Section 6.1, including the benchmark datasets, compared methods, pairwise supervision generation, experimental protocols, and implementation details.

C.1 DATASETS

Flat benchmarks. Following common practice in deep clustering and deep constrained clustering (DCC), we use seven benchmarks that have been widely adopted in prior studies (Xie et al., 2016; Guo et al., 2017; Ren et al., 2019; Zhang et al., 2021; Nguyen et al., 2023; Zhang & Chen, 2025). We additionally include CIFAR10H to evaluate graded supervision derived from human judgments.

- **CIFAR100-20** [1] (Krizhevsky & Hinton, 2009): A 20-class version of CIFAR100 obtained by treating its 20 superclasses as clustering targets. CIFAR100 contains 60,000 32×32 color images from 100 fine classes grouped into 20 superclasses, with 3,000 images per superclass. We use the original CIFAR100 split of 50,000 training and 10,000 test images.
- **CIFAR10** (Krizhevsky & Hinton, 2009): A natural-image dataset containing 60,000 32×32 color images from 10 object categories, with 6,000 images per category. We use a stratified $80/20$ train/test split, yielding 48,000 training and 12,000 test images.
- **FMNIST** [2] (Xiao et al., 2017): A fashion-product image dataset containing 70,000 grayscale 28×28 images from 10 categories. We use its official split of 60,000 training and 10,000 test images.
- **ImageNet10** [3] (Chang et al., 2017): A 10-class subset of ImageNet (Deng et al., 2009), containing 13,000 color images in total, with 1,300 images per class. We use a stratified $80/20$ train/test split, yielding 10,400 training and 2,600 test images.
- **MNIST** [4] (LeCun et al., 1998): A handwritten-digit dataset containing 70,000 grayscale 28×28 images from 10 digit classes. We use its official split of 60,000 training and 10,000 test images.
- **Reuters subset** [5] (Xie et al., 2016; Lewis et al., 2004): A preprocessed subset of the RCV1 newswire corpus using tf–idf features over the 2,000 most frequent words. It contains four root categories, *Corporate/Industrial* (CCAT), *Economics* (ECAT), *Government/Social* (GCAT), and *Markets* (MCAT), with 4,840, 3,470, 2,673, and 1,017 samples, respectively. We use 10,000 samples for training and 2,000 for testing.
- **STL10** [6] (Coates et al., 2011): We use the 13,000 labeled 96×96 color images from STL10, covering 10 object categories with 1,300 labeled images per category. We use a stratified $80/20$ train/test split, yielding 10,400 training and 2,600 test images.
- **CIFAR10H** [7] (Peterson et al., 2019): A human-annotated extension of the 10,000-image CIFAR10 test set, providing aggregate human class-distribution annotations for each image. We include it to evaluate graded supervision derived from human judgments, complementing the synthetically constructed supervision used in our other flat-benchmark experiments. The images remain evenly distributed across the 10 CIFAR10 classes. We use a stratified $80/20$ train/test split, yielding 8,000 training and 2,000 test images. The corresponding human-derived graded supervision is described in Appendix C.3.

Hierarchy benchmarks. We evaluate multi-resolution hierarchy recovery on five real hierarchies spanning text and image domains, two to four semantic resolutions, and substantially different hierarchy sizes and branching structures. We additionally use a controlled synthetic hierarchy for the theory-aligned analysis.

- **CIFAR100-HIER** (Krizhevsky & Hinton, 2009): Uses the native CIFAR100 hierarchy, in which 100 balanced fine classes are grouped into 20 superclasses with five fine classes per superclass, yielding the two resolutions $20 \to 100$. We use the original 50,000/10,000 train/test split.

[1] CIFAR webpage: https://www.cs.toronto.edu/~kriz/cifar.html
[2] FMNIST repository: https://github.com/zalandoresearch/fashion-mnist
[3] ImageNet10 repository: https://github.com/vector-1127/DAC
[4] MNIST webpage: http://yann.lecun.com/exdb/mnist/
[5] Reuters subset: https://github.com/piiswrong/dec
[6] STL10 webpage: https://cs.stanford.edu/~acoates/stl10/
[7] CIFAR10H repository: https://github.com/jcpeterson/cifar-10h

- **RCV1-HIER** [8] (Lewis et al., 2004): A two-resolution hierarchy constructed from the official RCV1 topic taxonomy. We use the four roots CCAT, ECAT, GCAT, and MCAT, and deterministically select five leaf topics under each root with at least 1,000 eligible documents. To obtain a disjoint partition, we retain only documents belonging to exactly one selected fine topic and exactly one of the four roots, with the two labels consistent under the taxonomy. We then sample 1,000 documents without replacement per fine class, yielding 20 fine classes nested under four roots and 20,000 documents in total. The resulting 2,000-dimensional tf–idf representations are split into 16,000 training and 4,000 test samples.
- **FGVC-Aircraft** [9] (Maji et al., 2013): A fine-grained aircraft dataset with variant, family, and manufacturer annotations. The processed benchmark used in our experiments contains 10,000 images organized into 30 manufacturers, 70 families, and 100 variants, giving the three resolutions $30 \to 70 \to 100$. It contains 6,667 training and 3,333 test images.
- **T608 and T200** [10] (Bertinetto et al., 2020): Two evaluation scales constructed from the WordNet (Miller, 1995) hierarchy of tieredImageNet-H. For T608, we retain all 608 leaf classes and obtain the three coarser resolutions by grouping leaves according to their direct parents, second-order ancestors, and third-order ancestors in the released hierarchy, yielding $63 \to 108 \to 201 \to 608$. T608 contains 440,800 training and 15,200 test images. T200 is formed once by uniformly sampling 200 leaf classes from T608 and retaining their inherited ancestor relations, yielding $52 \to 81 \to 130 \to 200$, with 145,000 training and 5,000 test images. The resulting T200 subset and hierarchy are fixed and reused unchanged throughout all experiments. The leaf classes are sample-balanced in both settings, whereas irregular WordNet branching induces substantial imbalance at coarser resolutions. At the coarsest resolution, the largest cluster is 50 times the size of the smallest in T608 and 16 times in T200.
- **Synthetic-Hier**: A balanced synthetic hierarchy with resolutions $4 \to 12 \to 48$, generated from a hierarchical Gaussian latent structure. Each of the 48 leaf classes contains 50 training and 20 test samples, yielding 2,400 training and 960 test instances. The controlled generator, difficulty settings, and theory-aligned interventions are described with the corresponding experimental protocol in Appendix C.4.

C.2 Compared methods

We compare GDP with seven representative baselines selected to span end-to-end assignment learning, deep constraint embedding, probabilistic relation modelling, and graded/weighted relation learning. DC-GMM and SC-DEC are discussed as related soft-constraint methods but are not direct experimental substitutes for unified GDP grades: DC-GMM encodes must-/cannot-link type by sign and certainty by magnitude, whereas SC-DEC relaxes should-link supervision through fuzzy memberships (Manduchi et al., 2021; AlZuhair et al., 2023). Using either with GDP-style grades would require a non-native adaptation. For multi-resolution hierarchy recovery, we use the hierarchy-applicable subset of these baselines and additionally consider FlatGDP, which applies the single-resolution GDP model independently at each prescribed resolution.

VanillaDCC. VanillaDCC (Hsu et al., 2019) is an end-to-end DCC approach based on the Meta Classification Likelihood (MCL) objective. For an instance x_j, let $\boldsymbol{q}_j = (q_{j1}, \ldots, q_{jK})$ denote its soft cluster-assignment vector. MCL interprets $P_i^{\mathrm{co}} = \boldsymbol{q}_{a_i}^\top \boldsymbol{q}_{b_i}$ as the co-assignment probability of a constrained pair and optimizes

$$\mathcal{L}_{\mathrm{MCL}} = -\frac{1}{|\mathcal{C}|} \sum_{i=1}^{|\mathcal{C}|} \left[g_i \log P_i^{\mathrm{co}} + (1 - g_i) \log(1 - P_i^{\mathrm{co}}) \right].$$

MCL was formulated with binary pair labels, but its Bernoulli target enters the objective directly. We therefore evaluate VanillaDCC with the supplied grade $g_i \in [0, 1]$ occupying the same target position.

[8] RCV1-HIER is available in our GitHub repository; the URL is provided in the supplementary materials.

[9] FGVC-Aircraft webpage: https://www.robots.ox.ac.uk/~vgg/data/fgvc-aircraft/

[10] T608 and T200 are available in our GitHub repository; the URL is provided in the supplementary materials.

VolMaxDCC. VolMaxDCC (Nguyen et al., 2023) augments assignment-level pairwise learning with explicit noise modeling and volume maximization. Instead of the raw co-assignment probability used by MCL, it employs the confusion-adjusted relation $P_i'^{\mathrm{co}} = \boldsymbol{q}_{a_i}^\top B \boldsymbol{q}_{b_i}$, where B is obtained from a learnable confusion mechanism. Its objective is

$$\mathcal{L}_{\mathrm{VolMax}} = -\frac{1}{|\mathcal{C}|}\sum_{i=1}^{|\mathcal{C}|}\left[g_i \log P_i'^{\mathrm{co}} + (1-g_i)\log(1-P_i'^{\mathrm{co}})\right] - \lambda_{\mathrm{vol}} \log\det\left(\left(\mathbf{Q}^{\mathrm{VM}}\right)^\top \mathbf{Q}^{\mathrm{VM}}\right),$$

where $\mathbf{Q}^{\mathrm{VM}}$ denotes the assignment matrix and λ_{vol} controls the volume-maximization regularizer. For our graded constraints, the pairwise BCE term is evaluated against the observed grade g_i in place of a binary pair label. This retains the native VolMaxDCC formulation while allowing its noise-aware assignment model to operate on the same graded observations.

CIDEC. CIDEC (Zhang et al., 2021) extends the DEC/IDEC deep embedding clustering framework (Xie et al., 2016; Guo et al., 2017) with pairwise constraints through the MCL loss. An autoencoder maps samples into a latent space, where K K-means-initialized cluster centers induce soft assignments $\boldsymbol{q}_j = (q_{j1}, \ldots, q_{jK})$. DEC sharpens these assignments into a target distribution $\boldsymbol{p}_j$ and refines the clustering by minimizing their KL divergence:

$$\mathcal{L}_{\mathrm{DEC}} = \sum_{j=1}^{|\mathcal{X}|}\sum_{k=1}^{K} p_{jk}\log\frac{p_{jk}}{q_{jk}},$$

where $\boldsymbol{p}_j$ is the sharpened target distribution derived from $\boldsymbol{q}_j$. CIDEC combines this clustering objective with reconstruction and MCL-based pairwise learning. For graded supervision, we use

$$\mathcal{L}_{\mathrm{CIDEC}}^{\mathrm{pair}} = -\frac{1}{|\mathcal{C}|}\sum_{i=1}^{|\mathcal{C}|}\left[g_i \log\left(\boldsymbol{q}_{a_i}^\top \boldsymbol{q}_{b_i}\right) + (1-g_i)\log\left(1-\boldsymbol{q}_{a_i}^\top \boldsymbol{q}_{b_i}\right)\right].$$

Thus, the DEC component shapes the latent clusters, while the MCL term fits their pairwise co-assignment probabilities to the supplied grades g_i.

SpherePair. SpherePair (Zhang & Chen, 2025) is an anchor-free deep constraint embedding method that learns normalized representations through pairwise angular geometry. For a binary constraint $y_i \in \{0,1\}$, let θ_i^{SP} denote the angle between the two constrained representations. Its original formulation defines

$$s_i^{\mathrm{hard}} = \frac{1}{2}\begin{cases}\cos\theta_i^{\mathrm{SP}} + 1, & y_i = 1,\\ \cos\big(\min\{\omega\theta_i^{\mathrm{SP}}, \pi\}\big) + 1, & y_i = 0,\end{cases}$$

and fits this relation to the binary constraint with BCE. The must-link branch attracts pairs toward zero angle, while the cannot-link branch defines a negative angular zone of size π/ω.

We use this original formulation for Binary supervision. For graded constraints $g_i \in [0,1]$, which are not covered by the binary piecewise relation, we follow the continuous extension in (Zhang & Chen, 2026). Specifically, we use $s_i^+ = (\cos\theta_i^{\mathrm{SP}} + 1)/2$ and $s_i^- = (\cos(2\theta_i^{\mathrm{SP}}) + 1)/2 = \cos^2\theta_i^{\mathrm{SP}}$, with

$$\mathcal{L}_{\mathrm{SP}} = -\frac{1}{|\mathcal{C}|}\sum_{i=1}^{|\mathcal{C}|}\left[g_i \log s_i^+ + (1-g_i)\log(1-s_i^-)\right].$$

Here, $g_i = 1$ gives pure angular attraction, $g_i = 0$ favors the orthogonal repulsive boundary at $\theta_i^{\mathrm{SP}} = \pi/2$, and intermediate grades continuously balance the two. SpherePair retains its normalized-embedding reconstruction objective, and K-means is applied to the learned representations for clustering.

ProbPair and Weighted ProbPair. ProbPair (Zhang & Chen, 2026) learns pairwise relations directly in angular representation space. For a constrained pair, it predicts $\hat{g}_i = \sigma((\cos(z_{a_i}, z_{b_i}) - m)/T)$, where the learnable operating point m and temperature $T > 0$ control the location and sharpness of the mapping. The predicted relation is fitted to g_i with BCE and optimized together with the normalized-embedding reconstruction objective.

Weighted ProbPair (WPP) (Zhang & Chen, 2026) uses the same relation model but modulates each pairwise term by an information-based decisiveness weight

$$w_i^{\mathrm{WPP}} = \frac{D_{\mathrm{KL}}(\mathrm{Bern}(g_i) \,\|\, \mathrm{Bern}(\bar{g}))}{(1-g_i)D_{\mathrm{KL}}(\mathrm{Bern}(0) \,\|\, \mathrm{Bern}(\bar{g})) + g_i D_{\mathrm{KL}}(\mathrm{Bern}(1) \,\|\, \mathrm{Bern}(\bar{g}))}, \qquad \bar{g} = |\mathcal{C}|^{-1} \sum_{i=1}^{|\mathcal{C}|} g_i.$$

The weighting emphasizes grades that carry stronger evidence relative to the empirical relation distribution. For Binary supervision, $w_i^{\mathrm{WPP}} = 1$ for the two endpoints, so WPP reduces to ProbPair.

ECI-PP. ECI-PP (Zhang & Chen, 2026) extends ProbPair with an Estimator–Corrector–Integrator pipeline for refining imperfect pairwise supervision. Given an observed grade g_i, out-of-fold Estimators produce a model-side relation belief $\hat{g}_i^{\mathrm{oof}}$. The Corrector predicts a discrepancy $\widehat{d}_i^{\mathrm{ECI}}$ and forms $g_i^{\mathrm{cor}} = \mathrm{softclip}_{\xi}(g_i + \widehat{d}_i^{\mathrm{ECI}})$. Agreement between g_i^{cor} and $\hat{g}_i^{\mathrm{oof}}$ determines a reliability weight w_i^{ECI}, which controls how strongly the corrected relation contributes to the fused target

$$g_i^{\mathrm{BC}} = \frac{g_i + n_0 w_i^{\mathrm{ECI}} g_i^{\mathrm{cor}}}{1 + n_0 w_i^{\mathrm{ECI}}}.$$

The Integrator learns from $(g_i^{\mathrm{BC}}, w_i^{\mathrm{ECI}})$, and its updated relation estimates are fed back to the Estimators for iterative refinement. In our experiments, the supplied grades g_i serve as the observed relations, with all constraints treated as a single supervision channel and without expert identities or corruption indicators.

FlatGDP. For the hierarchy experiments only, we include FlatGDP as a single-resolution reference that applies GDP independently at each prescribed hierarchy resolution. For hierarchy cluster counts $K^{(1)}, \ldots, K^{(L)}$, an independent GDP model is trained for each $K^{(\ell)}$ using the same hierarchy-derived graded constraints, producing one flat partition at that resolution. The partitions are therefore learned without joint optimization across levels. FlatGDP provides a direct reference for separating repeated single-resolution GDP learning from the joint multi-resolution formulation of GDP-Hier.

C.3 Pairwise supervision generation

For flat clustering, we construct pairwise supervision from four sources with different relation-to-clustering interpretations. Binary constraints encode hard class agreement, Difficulty introduces graded evidence while retaining definite class membership, Mix represents genuine partial membership, and CIFAR10H provides human-derived graded relations. For hierarchy recovery, grades encode the deepest shared semantic resolution.

Binary supervision. For a pair of samples (x_a, x_b) with hard labels y_a and y_b, Binary supervision uses

$$g_{ab} = \mathbb{I}[y_a = y_b].$$

Thus, same-class and different-class pairs receive grades 1 and 0, respectively.

Difficulty-based supervision. Difficulty models weak or atypical evidence while keeping the underlying class membership definite. We train a frozen auxiliary classifier on a stratified 1% subset of the training data. Following (Zhang & Chen, 2026), we use a two-layer 512–512 MLP with ReLU activations and dropout 0.2, trained with cross-entropy using Adam with learning rate 10^{-3}, batch size 256, and at most 100 epochs, with early-stopping patience 10 based on training accuracy. Let $\boldsymbol{\pi}_j$ denote its predictive class distribution for x_j. The easiness of an instance is

$$s_j = \frac{\pi_{j,y_j}^2}{\sum_k \pi_{jk}^2}, \qquad s_{uv} = \sqrt{s_u s_v}.$$

For a boundary $b \in [0, 1]$, the pair grade is

$$g_{uv} = \begin{cases} b + (1-b)s_{uv}, & y_u = y_v, \\ b(1 - s_{uv}), & y_u \neq y_v. \end{cases}$$

Easy pairs therefore approach the Binary endpoints, whereas difficult pairs move toward the shared boundary b without changing their hard class relation. We use $b = 0.5$ by default.

Table 4: Hierarchy-grade composition under uniform pair sampling. Percentages are computed exactly from the training-pool hierarchy and follow the order of the listed grade values.

Dataset	Grade values	Proportion (%)
CIFAR100-HIER	$0, \frac{1}{2}, 1$	95.0, 4.0, 1.0
RCV1-HIER	$0, \frac{1}{2}, 1$	75.0, 20.0, 5.0
Aircraft	$0, \frac{1}{3}, \frac{2}{3}, 1$	91.1, 6.7, 1.2, 1.0
T200	$0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1$	96.4, 1.8, 0.9, 0.5, 0.5
T608	$0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1$	96.4, 2.0, 0.9, 0.6, 0.2

Mixed-membership supervision. Mix creates samples with explicit partial membership. For two training samples x_u and x_v from different classes, we draw $\lambda \sim \mathcal{U}(0,1)$ and construct

$$x_{\text{mix}} = \lambda x_u + (1-\lambda)x_v, \qquad \boldsymbol{p}_{\text{mix}} = \lambda \boldsymbol{e}_u + (1-\lambda)\boldsymbol{e}_v,$$

where $\boldsymbol{e}_u$ and $\boldsymbol{e}_v$ are the corresponding one-hot class-membership vectors. Here, x_u and x_v denote the model inputs used for clustering; for benchmarks represented by fixed image features, interpolation is therefore performed in that fixed feature space rather than in raw-image pixel space. Original samples retain their one-hot memberships. For each training split, we generate as many mixed samples as original samples, doubling the constraint-generation pool, and assign a pair (x_a, x_b) the grade

$$g_{ab} = \boldsymbol{p}_a^\top \boldsymbol{p}_b.$$

Hence, intermediate grades have an explicit soft-membership interpretation rather than reflecting uncertain evidence about a definite class.

Human-derived supervision. For CIFAR10H (Peterson et al., 2019), let $\boldsymbol{h}_j \in \Delta^9$ denote the aggregate human class-distribution annotation of image x_j. We define the human-derived grade as

$$g_{ab} = \boldsymbol{h}_a^\top \boldsymbol{h}_b.$$

Constraints are generated only from the 8,000-image training split described in Appendix C.1; the remaining 2,000 images are held out for evaluation. Unlike Mix, these soft memberships are obtained from human judgments rather than synthetically constructed interpolation.

Hierarchy-induced supervision. We use the hierarchy notation introduced in Section 3. For L prescribed semantic resolutions with breakpoints $0 = \beta_0 < \beta_1 < \cdots < \beta_L = 1$, let $h_{ab}^{(\ell)}$ indicate whether x_a and x_b belong to the same cluster at level ℓ, and let $\Delta_\ell^\beta = \beta_\ell - \beta_{\ell-1}$. The hierarchy grade is

$$g_{ab} = \sum_{\ell=1}^{L} \Delta_\ell^\beta h_{ab}^{(\ell)} = \beta_{\ell_{ab}^\star},$$

where $\ell_{ab}^\star$ is the deepest resolution shared by the pair. Thus, intermediate grades represent co-membership at a coarser semantic resolution rather than fractional membership in one flat partition. Unless otherwise stated, we use evenly spaced breakpoints $\beta_\ell = \ell/L$. For example, a two-level hierarchy uses $(0, 0.5, 1)$, assigning grade 0.5 to pairs that share the coarse cluster but differ at the fine level.

Under uniform pair sampling, the hierarchy-grade composition is determined directly by the cluster-size distribution at each resolution. Let N denote the size of the training pool, let $n_k^{(\ell)}$ be the size of cluster k at resolution ℓ, and define $M_\ell = \sum_k \binom{n_k^{(\ell)}}{2}$. Among the $\binom{N}{2}$ possible unordered pairs, grades 0, β_ℓ for $1 \le \ell < L$, and 1 occur with counts $\binom{N}{2} - M_1$, $M_\ell - M_{\ell+1}$, and M_L, respectively. Table 4 reports the resulting exact proportions for all five hierarchy benchmarks.

C.4 Experimental protocol

We summarize the protocols for flat clustering, robustness to numerical grade scales, hierarchy recovery, component analyses, controlled Synthetic-Hier experiments, and hyperparameter sensitivity.

Common evaluation protocol. For all real-data experiments, pairwise constraints are uniformly sampled from the corresponding training pool. We report ACC, NMI, and ARI against the available ground-truth labels. Assignment-based methods use their native cluster predictions, whereas embedding-based methods apply K-means to the learned representations. For Mix, clustering scores are computed on the original, non-mixed samples. All real-data results are averaged over five independent runs.

Main flat-clustering evaluation. On the seven standard flat benchmarks, we evaluate Binary, Difficulty, and Mix supervision with 1k, 5k, and 10k constraints. The main setting uses pretrained initialization; for methods using autoencoder pretraining, we also evaluate the same budgets from random initialization. CIFAR10H complements these experiments with human-derived graded supervision under the 5k setting.

Hard-label clustering scores under Mix do not directly measure recovery of partial membership. We therefore additionally examine GDP's recovered mixture composition. For a mixed sample generated from source classes u and v with coefficient λ, after permutation alignment we estimate $\hat{\lambda} = q_{\mathrm{mix},u}/(q_{\mathrm{mix},u} + q_{\mathrm{mix},v})$ from the corresponding GDP assignment probabilities and compare it with the generating λ.

Robustness to numerical grade scales. Numerical grades need not have universally calibrated absolute semantics: the same relation ordering can be represented on different numerical scales. We therefore vary the grade scale while preserving the underlying relation semantics and ordering. With pretrained initialization, a sufficiently large budget of 10k constraints is used to reduce the influence of supervision sparsity and isolate the effect of grade calibration.

For Difficulty, we vary the boundary as $b \in \{0.3, 0.5, 0.7\}$ around the default $b = 0.5$. These settings shift the numerical separation between same-class and different-class grades without changing their ordering. Because different-class pairs dominate under uniform sampling on typical multiclass data, the grade distribution is naturally concentrated toward lower values; varying b moves this distribution across substantially different numerical regimes.

For Mix, we use the endpoint-preserving monotone transformation

$$g' = \begin{cases} 0, & g = 0, \\ \sigma(\tau\ \mathrm{logit}(g) + \zeta)\,, & 0 < g < 1, \\ 1, & g = 1. \end{cases}$$

Besides the identity setting $(\tau, \zeta) = (1, 0)$, we consider $\tau = 0.2$, which compresses interior grades toward 0.5, and $\zeta = +3$, which shifts them toward larger values. The two transformations alter the spread and location of the numerical scale, respectively, while preserving endpoints and relation ordering.

Multi-resolution hierarchy recovery. We evaluate hierarchy recovery on the five real hierarchy benchmarks using pretrained models and 5k hierarchy constraints. The evaluation follows the prescribed nested (laminar) setting of Section 3, where each finer cluster has a single parent, rather than a more general multiple-parent taxonomy represented by a DAG. The comparison includes representative methods from assignment-level DCC, deep constraint embedding, and probabilistic pairwise learning, without duplicating family-level extensions whose additional machinery is orthogonal to multi-resolution partitioning.

VanillaDCC and CIDEC are trained independently at each prescribed cluster count, and FlatGDP similarly trains one single-resolution GDP model per level. SpherePair, ProbPair, and WPP learn one representation and apply K-means independently at each prescribed cluster count, so cross-level coherence is evaluated rather than imposed by hierarchy-enforcing post-processing. GDP-Hier jointly learns all prescribed resolutions.

We report ACC, NMI, and ARI at each resolution, together with mean ACC, finest-level (Leaf) ACC, and hierarchy violation rate (HVR). For adjacent levels ℓ and $\ell + 1$, each predicted finer-level cluster is assigned to the coarser cluster containing the majority of its samples. The transition violation rate

v_ℓ is the fraction of samples inconsistent with this majority parent, and

$$\mathrm{HVR} = \frac{1}{L-1}\sum_{\ell=1}^{L-1} v_\ell.$$

Thus, v_ℓ measures coherence at one hierarchy transition, whereas HVR averages coherence violations across the hierarchy.

We additionally examine hierarchy-grade breakpoints on CIFAR100-HIER and RCV1-HIER using pretrained models and 10k constraints, varying the intermediate breakpoint as $\beta_1 \in \{0.3, 0.5, 0.7\}$ while preserving the hierarchy and semantic ordering.

Component and mechanism analyses. We conduct the main component ablation without autoencoder pretraining to expose differences among model variants more clearly. We use 5k constraints on all seven standard benchmarks under seven supervision conditions: Binary; Difficulty with $b \in \{0.3, 0.5, 0.7\}$; and Mix with $\tau = 0.2$, the identity mapping, and $\zeta = +3$. We compare full GDP with Global BCE, w/o LGRC, w/o gated vertex, and w/o marginal entropy. Global BCE replaces the ordinal Global ranking objective with direct BCE fitting of the Global relation score to the numerical grade; the remaining variants remove the indicated component individually.

Grade fitting is examined separately by comparing GDP-Hier with its w/o grade-fitting variant on CIFAR100-HIER and RCV1-HIER under 1k, 5k, and 10k constraints without autoencoder pretraining. We also stress-test marginal entropy on T200, whose coarsest resolution has a 16:1 largest-to-smallest cluster-size ratio, using the pretrained 5k setting with and without marginal entropy.

Controlled Synthetic-Hier analysis. Synthetic-Hier allows hierarchy separability and theory-relevant quantities to be varied directly. The $4 \to 12 \to 48$ hierarchy is generated from a 64-dimensional hierarchical Gaussian tree. For a node v at level ℓ, its center is constructed as

$$\boldsymbol{\mu}_v = \boldsymbol{\mu}_{\mathrm{parent}(v)} + \Delta_\ell^{\mathrm{sep}} \frac{\boldsymbol{d}_v}{\|\boldsymbol{d}_v\|_2}, \qquad \boldsymbol{d}_v \sim \mathcal{N}(\mathbf{0}, I),$$

with unit isotropic Gaussian noise at the leaves, followed by a fixed nonlinear $64 \to 256 \to 512$ observation map with Gaussian observation noise of standard deviation 0.1.

The tuple $(\Delta_1^{\mathrm{sep}}, \Delta_2^{\mathrm{sep}}, \Delta_3^{\mathrm{sep}})$ specifies the levelwise center separation at the coarse, middle, and fine resolutions, respectively. We use four difficulty profiles:

$$\begin{aligned} \text{VeryEasy} &: (16, 10, 8), \quad \text{Easy} : (10, 7, 4), \\ \text{Medium} &: (9, 6, 3), \quad \text{Hard} : (6, 4, 2). \end{aligned}$$

Larger offsets increase separation relative to the fixed noise level and therefore produce easier recovery. Within each matched repeat, the hierarchy directions, sampled noise, observation map, and sample identities are shared across profiles; only the three separation parameters change. The main analysis trains GDP-Hier for 100 epochs with 20k hierarchy constraints evenly distributed across the four grade levels, using 10 runs per profile.

We summarize the empirical recovery regime by

$$\mathcal{E}_{95} = \frac{\delta_{95} + 2\sum_{\ell=1}^{L} \Delta_\ell^\beta \epsilon_{\ell,95}}{\Delta_{\min}^\beta},$$

where δ_{95} and $\epsilon_{\ell,95}$ are empirical 95th-percentile relation-fit and assignment errors, respectively. This is the empirical counterpart of the error-to-gap quantity appearing in the recovery analysis.

We further reduce the minimum hierarchy-grade gap from $1/3$ to 0.03. Starting from $(0, \frac{1}{3}, \frac{2}{3}, 1)$, the first increment is fixed at 0.03 and the remaining interval is divided equally between the two later levels, giving

$$(0,\ 0.03,\ 0.515,\ 1).$$

This changes the grade separation while preserving the hierarchy topology and, within each matched run, the generated data and sampled constraints.

Hyperparameter sensitivity. We vary one factor at a time around the shared GDP defaults, considering α, γ, η, λ_{rec}, and the synchronized Global update/feedback interval. The analysis uses pretrained GDP with Binary supervision on the seven standard benchmarks under 1k, 5k, and 10k constraint budgets. The tested values are reported with the corresponding sensitivity results.

C.5 IMPLEMENTATION

We provide the implementation details[11] for the models evaluated in our experiments, including network architectures, pretraining, and optimization settings.

Data preprocessing and input features. For MNIST and FMNIST, we use flattened 784-dimensional pixel vectors, while Reuters and RCV1-HIER use 2,000-dimensional tf–idf representations. For CIFAR100-20, CIFAR10, ImageNet10, STL10, CIFAR10H, and CIFAR100-HIER, we use fixed 512-dimensional features extracted with BYOL (Grill et al., 2020) using a ResNet-34 backbone; Aircraft uses 2,048-dimensional BYOL features from a ResNet-50 backbone. For CIFAR10, CIFAR100, and ImageNet10, BYOL pretraining is transductive with respect to the downstream split, using the complete image collections before those splits; CIFAR10H reuses the CIFAR10 representation, STL10 uses its official unlabeled split, and Aircraft uses the official training and validation images only, with its ResNet-50 initialized from ImageNet-pretrained weights. For T200 and T608, we use fixed 2,048-dimensional features extracted with a SwAV ResNet-50 (Caron et al., 2020) using the official 800-epoch self-supervised ImageNet checkpoint.[12] In all cases, the resulting input features are fixed and shared across all compared methods before constrained-clustering training, while pairwise supervision and clustering optimization use only the designated training split. The train/test splits follow Appendix C.1.

Network architectures. The encoder-based methods use fully connected architectures consistent with prior DCC studies (Ren et al., 2019; Zhang et al., 2021; Zhang & Chen, 2025).

(i) *Flat clustering.* CIDEC, SpherePair, ProbPair, WPP, and ECI-PP use an autoencoder with encoder widths 500–500–2000–D and a symmetric decoder. We set $D = 20$ for CIFAR100-20 and $D = 10$ for the remaining flat benchmarks. GDP uses the same 500–500–2000 encoder stem, with the 2,000-dimensional hidden layer serving as the Local representation and a trainable linear projection producing the D-dimensional Global representation.

(ii) *Hierarchy recovery.* CIDEC and FlatGDP are trained independently at each resolution, with $D = K^{(\ell)}$ except for the coarse RCV1-HIER resolution, where $D = 10$. SpherePair, ProbPair, and WPP instead learn a single embedding with $D = K^{(L)}$, where $K^{(L)}$ is the finest-level cluster count, and apply K-means to the same representation at the required resolutions. GDP-Hier shares one 500–500–2000 Local encoder across all resolutions, with the 2,000-dimensional hidden layer serving as the shared Local representation. Each resolution ℓ has a separate Global projection with dimension $D_\ell = K^{(\ell)}$.

(iii) *Native assignment networks.* VanillaDCC and VolMaxDCC retain their native assignment MLPs with two hidden layers of size 512–512.

Autoencoder pretraining. In our experiments, pretraining denotes autoencoder-based initialization of the clustering-network backbone. Pretrain-capable methods use stacked denoising autoencoder (SDAE) pretraining (Hinton & Salakhutdinov, 2006; Vincent et al., 2008). Each hidden layer is pretrained for 300 epochs with 20% random masking corruption and MSE reconstruction, after which the complete autoencoder is fine-tuned for 500 epochs under the same corruption and reconstruction objective. Both stages use SGD with learning rate 0.1, momentum 0.9, and batch size 256 on the training split. CIDEC uses the standard SDAE initialization. SpherePair, ProbPair, WPP, ECI-PP, GDP, and GDP-Hier use Sphere-SDAE initialization, where reconstruction is performed from normalized embeddings (Zhang & Chen, 2025). For GDP and GDP-Hier, the pretrained weights initialize the shared Local encoder, whereas the Global heads are initialized separately as described below. The no-pretraining experiments use random initialization.

[11] **Our source code is available on GitHub, and the URL is included in the supplementary materials.**

[12] Official SwAV checkpoint: https://dl.fbaipublicfiles.com/deepcluster/swav_800ep_pretrain.pth.tar

GDP implementation. GDP retains the 2,000-dimensional Local representation and instantiates the Global branch as follows.

(i) *Global geometry.* The Local representation is mapped to z^{glob} by a trainable linear projection. For K clusters, the K unit prototypes are fixed at the vertices of a regular simplex constructed in a $(K-1)$-dimensional subspace and zero-padded when the Global dimension is larger. This construction requires the Global dimension to satisfy $D_{\mathrm{glob}} \geq K-1$. The prototype weights remain fixed throughout training. The simplex radius R is initialized to 1 and constrained to remain positive through a softplus parameterization.

Vertex calibration uses an additional orthogonal projection onto the fixed prototype span. Let B denote an orthonormal basis of this span, computed once from the prototype matrix by SVD; the projected representation is $\Pi_{\mathcal{W}}(z) = BB^{\top}z$. This projection is used only in the vertex term. Global assignments in Eq. (4) are computed directly from the unprojected z^{glob} and the fixed prototypes.

(ii) *Pretrained initialization.* For pretrained GDP, spherical K-means is first applied to the normalized Local representations. The Local-to-Global projection is then derived analytically from the resulting K-means centroids and the fixed simplex prototypes, such that the K-means hard partition is preserved under the initial Global assignments. The projection is subsequently calibrated to match the mean assignment entropy of the initial Global head while retaining this hard partition. The Global projection and Local encoder use a 100-epoch linear learning-rate warm-up. Without pretraining, the Global head is randomly initialized and the K-means-based initialization is omitted.

(iii) *Optimization.* GDP is optimized with Adam using a base learning rate of 10^{-3} and batch size 256. The default loss weights are $\alpha = 0.2$, $\gamma = 1$, $\eta = 2$, and $\lambda_{\mathrm{rec}} = 0.02$. For marginal entropy, we separately compute the assignment marginal over the two endpoints of the current constraint mini-batch and over a standard training-data mini-batch, and sum their entropies. The radius uses learning rate 0.2, while the Local ProbPair parameters (m, T) use learning rate 10^{-2}. The Local-to-Global objectives and Global-to-Local feedback are optimized every five epochs, whereas LGRC is optimized at every epoch. For Ranking, we use at most 5,000 randomly sampled grade-ordered comparisons. For LGRC, we use an additional random-pair budget of $\min(10|\mathcal{C}|, 50{,}000)$. GDP is trained for 500 epochs by default; for pretrained runs with 5k or 10k constraints, training is reduced to 100 epochs.

GDP-Hier implementation. GDP-Hier extends GDP by jointly optimizing multiple resolution-specific Global heads over a shared Local representation. The pretrained initialization, loss weights, and optimization settings follow those of GDP, with the same initialization procedure applied independently to each Global head. For the real hierarchy benchmarks, training uses 3,000 epochs when the finest-level cluster count is at least 100, and 1,000 epochs otherwise.

Baseline implementation. For the compared baselines, we follow their reported or established implementation procedures whenever available.

- **VanillaDCC.** VanillaDCC (Hsu et al., 2019) optimizes the MCL objective using Adam with learning rate 10^{-3} and batch size 256. Training runs for at most 500 epochs.
- **VolMaxDCC.** For VolMaxDCC, we follow the original optimization and validation-based hyperparameter-selection procedure (Nguyen et al., 2023). The learnable confusion matrix is parameterized as $B = \sigma(B')$, with diagonal entries of B' initialized to 1 and off-diagonal entries to -1. SGD is used with learning rate 0.5 for the network parameters and 0.1 for B', with batch size 128 and a maximum of 500 epochs. The geometric regularization weight λ_{vol} is selected from $\{0, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ using a validation subset reserved from the training data; ground-truth labels of this subset are used to compute mean validation ACC over three runs for model selection, following the search procedure of the original method.
- **CIDEC.** For CIDEC, we follow the recommended configuration in Zhang et al. (2021), with the DEC clustering coefficient set to 0.1. Cluster anchors are initialized by K-means, and optimization uses Adam with learning rate 10^{-3} and batch size 256. Training runs for at most 500 epochs, with early stopping based on the relative change in soft cluster assignments.

- **SpherePair.** SpherePair follows its original angular-embedding configuration (Zhang & Chen, 2025). We use reconstruction weight $\lambda_{\text{rec}} = 0.02$ and fix the negative-zone factor to its canonical value $\omega = 2$ for Binary supervision; graded supervision follows the continuous extension described in Appendix C.2. Optimization uses Adam with learning rate 10^{-3} and batch size 256, and training runs for 500 epochs.
- **ProbPair and WPP.** ProbPair and WPP follow the implementation introduced in Zhang & Chen (2026). Both use $\lambda_{\text{rec}} = 0.02$ and Adam with base learning rate 10^{-3} and batch size 256. The probabilistic readout parameters (m, T) are initialized as $(0, 0.1)$ and optimized with learning rate 10^{-2}. WPP additionally applies the information-based pair weighting defined in Appendix C.2. Both methods are trained for 500 epochs.
- **ECI-PP.** ECI-PP uses the default configuration reported in Zhang & Chen (2026), with five Estimator folds, a 64–16 Corrector MLP, $\lambda_{\text{cor}} = 0.5$, $\gamma = 10$, $n_0 = 10$, and soft-clipping sharpness $\xi = 20$. The Estimator and Integrator use Adam with base learning rate 10^{-3} and batch size 256, while their probabilistic readout parameters use learning rate 10^{-2}. Training uses the information-based warm-up and runs for 500 epochs.
- **FlatGDP.** FlatGDP applies the pretrained single-resolution GDP implementation independently at each prescribed hierarchy resolution and otherwise retains the GDP initialization and optimization settings above. Training uses 3,000 epochs when the corresponding cluster count is at least 100 and 1,000 epochs otherwise.

Software and hardware. All model training code is implemented in PyTorch 2.3.0 with CUDA 12.1 [13]. Each training run is executed on a single NVIDIA L40S 48GB or NVIDIA H200 141GB GPU.

D Additional experimental results

Additional results complement the main experiments in Section 6.2 with complete comparisons and further analyses that substantiate the main empirical findings.

D.1 Main clustering effectiveness

We extend the main comparison in Table 1 in five directions: (i) complete training- and test-set results with uncertainty statistics; (ii) results across constraint budgets; (iii) results without autoencoder pretraining; (iv) comparison with an unsupervised clustering reference; and (v) cross-source and aggregate comparisons.

Complete statistics for the main comparison. Tables 5–7 provide the complete training- and test-set results with standard deviations for the main comparison in Table 1. The training-set results largely mirror the test-set comparisons, with GDP retaining the same overall advantage across the three supervision sources. GDP also shows generally small standard deviations, indicating stable performance across repeated runs. These complete statistics therefore support the conclusions drawn from the test-set means reported in the main text.

Performance across constraint budgets. Tables 8–10 report the pretrained comparisons with 1k, 5k, and 10k constraints. GDP remains broadly leading across all three budgets, ranking first in 48, 51, and 48 of the 63 dataset–source–metric comparisons, respectively. The effect of increasing supervision is particularly visible on the more challenging datasets. On CIFAR100-20, for example, GDP's ACC increases from 58.2 to 71.0 to 76.1 under Binary, from 58.6 to 69.9 to 76.7 under Difficulty, and from 56.7 to 64.0 to 72.4 under Mix as the budget increases. In contrast, at 10k constraints, several strong methods already approach similar high performance on datasets such as CIFAR10 and ImageNet10, leaving less separation between them. The intermediate budget consequently exposes methodological differences more clearly than either the sparsest or several near-saturated settings.

Across these regimes, GDP also shows a comparatively consistent advantage rather than relying on one particular supervision source. This is especially clear under Difficulty, where it ranks first in

[13] PyTorch 2.3.0: https://github.com/pytorch/pytorch/releases/tag/v2.3.0

Table 5: Training (blue) and test (black) ACC/NMI/ARI (%; five-run mean±std) under Binary supervision at 5k constraints. Best/second-best test means are **bold**/underlined; WPP reduces to ProbPair and is omitted.

Dataset	Metric	Vanilla DCC	VolMax DCC	CIDEC	SpherePair	ProbPair	WPP	ECI-PP	**GDP**
CIFAR100-20	ACC	48.9±3.9 48.9±4.1	44.8±5.2 44.6±5.4	57.7±1.8 57.6±1.8	69.5±1.8 69.4±1.7	66.6±1.8 66.6±1.8	–	64.2±2.6 63.9±2.6	71.2±1.4 **71.0**±1.5
	NMI	51.0±2.6 51.7±2.8	47.2±2.4 47.8±2.3	59.9±1.0 60.5±0.9	63.7±1.0 64.5±0.8	63.2±1.3 64.0±1.3	–	62.7±0.7 63.2±0.7	65.6±1.2 **66.2**±1.2
	ARI	35.0±3.4 35.0±3.5	29.5±4.1 29.5±4.2	43.9±1.2 43.6±1.2	53.3±2.1 53.4±1.9	50.6±1.7 50.7±1.6	–	49.5±1.7 49.3±1.8	54.9±1.7 **54.7**±1.9
CIFAR10	ACC	88.2±6.7 88.1±6.6	89.0±7.5 89.1±7.1	94.9±0.2 94.8±0.2	94.9±0.1 94.7±0.1	94.8±0.2 94.7±0.2	–	94.7±0.2 94.6±0.2	95.0±0.1 **94.9**±0.1
	NMI	84.6±4.1 84.4±4.1	84.5±5.5 84.5±5.3	89.2±0.2 89.0±0.2	89.1±0.1 88.8±0.3	88.9±0.2 88.8±0.3	–	88.9±0.2 88.9±0.2	89.4±0.1 **89.3**±0.2
	ARI	81.7±7.3 81.3±7.1	82.2±8.6 82.1±8.4	89.4±0.4 89.1±0.3	89.4±0.1 89.0±0.3	89.1±0.3 88.9±0.4	–	89.0±0.4 88.9±0.4	89.6±0.1 **89.4**±0.2
FMNIST	ACC	77.7±1.2 76.7±1.0	69.4±5.6 68.6±5.6	68.4±7.8 67.6±8.1	80.6±0.9 79.5±1.0	77.8±1.7 76.9±1.8	–	70.9±3.9 70.2±3.7	82.8±1.5 **81.8**±1.3
	NMI	68.3±0.5 67.3±0.7	63.7±3.5 62.9±3.5	64.3±3.7 63.6±3.9	72.3±0.5 71.2±0.7	69.6±1.0 68.7±1.3	–	66.7±1.1 66.0±0.9	74.4±0.9 **73.4**±0.9
	ARI	62.6±1.1 61.3±1.0	55.5±4.6 54.4±4.6	54.7±7.2 53.8±7.3	66.7±0.9 65.2±1.0	62.6±1.8 61.3±2.1	–	57.3±1.9 56.5±1.6	69.5±1.4 **68.1**±1.4
ImageNet10	ACC	88.6±6.7 88.7±6.8	77.4±8.9 77.5±8.9	94.8±0.1 95.5±0.3	95.2±0.2 95.8±0.2	95.4±0.4 95.8±0.3	–	95.4±0.2 **96.3**±0.1	95.6±0.2 96.1±0.4
	NMI	82.1±5.6 82.6±5.8	70.9±8.2 71.6±8.4	88.1±0.2 89.9±0.5	88.6±0.4 90.2±0.4	89.1±0.7 90.2±0.9	–	89.2±0.4 **91.5**±0.2	89.6±0.4 91.0±0.9
	ARI	80.4±8.3 80.3±8.8	66.4±10.4 66.7±10.5	89.0±0.2 90.4±0.5	89.7±0.5 90.9±0.5	90.1±0.8 90.9±0.7	–	90.1±0.4 **92.1**±0.2	90.6±0.3 91.7±0.8
MNIST	ACC	81.6±2.2 82.1±2.1	80.8±5.7 81.1±5.8	97.4±0.1 96.6±0.3	96.6±0.4 96.4±0.3	95.1±1.8 94.8±1.4	–	93.4±0.6 93.5±0.8	96.9±0.3 **97.1**±0.3
	NMI	73.6±1.8 74.8±1.7	72.3±4.6 73.1±5.0	93.3±0.2 91.8±0.5	91.2±0.8 90.9±0.6	88.2±3.3 87.8±2.7	–	85.6±0.8 85.9±1.1	91.8±0.7 **92.5**±0.7
	ARI	70.7±2.4 71.5±2.3	70.4±6.2 71.0±6.5	94.5±0.2 92.7±0.7	92.7±0.7 92.2±0.5	89.6±3.5 88.9±2.9	–	86.3±1.1 86.5±1.5	93.3±0.6 **93.8**±0.7
Reuters	ACC	96.1±0.3 95.0±0.3	94.4±0.2 93.4±0.2	83.2±1.0 82.3±0.8	96.1±0.3 94.5±0.5	96.1±0.2 94.9±0.2	–	95.6±0.3 94.8±0.3	96.3±0.3 **95.2**±0.7
	NMI	84.6±0.8 81.2±0.7	79.8±0.6 76.7±0.5	70.8±0.5 69.5±0.7	84.6±1.0 79.9±1.3	84.6±0.5 81.0±0.6	–	83.3±0.6 81.3±0.6	85.3±0.8 **82.1**±2.4
	ARI	90.9±0.5 87.9±0.7	87.5±0.3 84.3±0.4	79.6±0.5 77.6±0.6	91.0±0.7 86.6±1.1	90.8±0.4 87.7±0.4	–	89.7±0.5 88.0±0.5	91.2±0.6 **88.5**±1.9
STL10	ACC	83.4±9.6 82.2±9.0	80.2±6.3 78.9±6.3	90.6±0.2 88.9±0.2	90.9±0.3 89.3±0.4	90.4±0.4 88.7±0.2	–	90.0±0.5 88.9±0.5	91.6±0.3 **89.9**±0.3
	NMI	76.4±6.2 74.8±5.7	72.4±5.8 71.4±5.7	81.8±0.4 79.6±0.3	82.2±0.4 80.3±0.5	81.7±0.4 79.6±0.1	–	81.2±0.6 79.9±0.7	83.2±0.5 **81.2**±0.4
	ARI	72.9±9.8 70.7±8.9	64.7±9.2 62.6±8.7	80.9±0.4 77.8±0.4	81.6±0.6 78.7±0.7	80.5±0.7 77.5±0.3	–	80.0±0.8 78.0±0.9	82.8±0.6 **79.7**±0.4

18/21 comparisons at both 1k and 5k constraints and in all 21 comparisons at 10k, while the closest competing methods vary more across datasets, sources, and budgets. Overall, the budget sweep shows that GDP's advantage persists from sparse to dense supervision, while the magnitude of the observed differences naturally changes with the available pairwise information and the difficulty of the dataset.

Results without autoencoder pretraining. Tables 11–13 repeat the comparison with random initialization for pretrain-capable methods; VanillaDCC and VolMaxDCC remain unchanged because they have no pretraining variant. GDP† remains broadly leading across all three budgets, ranking first in 49/63, 58/63, and 48/63 dataset–source–metric comparisons at 1k, 5k, and 10k constraints, respectively. At the main 5k budget, GDP† is within the top two in all 63 comparisons. Removing pretraining therefore does not qualitatively change the overall comparison, although the effect varies across datasets and supervision sources.

The comparative advantage of GDP becomes more pronounced without autoencoder pretraining. At the main 5k budget, its number of first-place results increases from 51/63 with pretraining to 58/63 with random initialization, while its top-two coverage increases from 59/63 to all 63 comparisons. The change is particularly clear on CIFAR100-20 under Mix supervision: with pretraining, GDP reaches 64.0% ACC and trails the strongest competitor by 4.5 points, whereas GDP† reaches 65.0% without pretraining and instead leads the strongest competitor by 2.9 points. Thus, although pretraining can

Table 6: Training (blue) and test (black) ACC/NMI/ARI (%; five-run mean±std) under Difficulty supervision at 5k constraints. Best/second-best test means are **bold**/underlined.

Dataset	Metric	Vanilla DCC	VolMax DCC	CIDEC	SpherePair	ProbPair	WPP	ECI-PP	**GDP**
CIFAR100-20	ACC	17.3±0.7	45.3±3.6	29.1±2.9	68.4±1.3	60.7±2.4	63.0±3.2	64.5±2.2	70.3±1.2
		17.4±0.7	45.4±3.5	25.9±2.6	68.0±1.4	60.8±2.4	63.2±3.3	64.3±2.1	**69.9**±1.1
	NMI	14.9±1.0	48.1±2.3	31.6±3.2	62.7±1.1	58.7±1.1	60.4±1.8	62.1±1.1	64.1±0.8
		15.3±1.0	49.0±2.2	29.7±3.0	62.9±1.1	59.5±1.3	61.5±1.8	62.6±1.1	**64.5**±0.8
	ARI	4.1±0.3	30.4±3.0	4.5±0.5	51.1±1.9	42.2±1.6	45.3±3.1	48.0±2.5	53.4±1.3
		4.1±0.3	30.6±3.0	2.5±0.3	50.8±2.0	42.3±1.6	45.7±3.2	47.9±2.4	**52.9**±1.1
CIFAR10	ACC	74.1±6.9	86.9±6.9	93.2±0.2	94.6±0.2	94.5±0.2	94.4±0.2	94.3±0.2	94.9±0.2
		74.3±6.9	87.1±6.8	93.2±0.4	94.7±0.2	94.5±0.2	94.4±0.2	94.3±0.2	**94.9**±0.1
	NMI	69.6±6.0	82.8±4.2	84.8±0.3	88.3±0.3	88.5±0.2	88.4±0.2	88.3±0.3	89.0±0.3
		70.2±6.1	83.4±4.4	85.7±0.5	88.6±0.2	88.6±0.3	88.5±0.2	88.4±0.2	**89.1**±0.2
	ARI	62.9±8.4	80.6±6.7	85.8±0.5	88.7±0.3	88.5±0.3	88.4±0.3	88.2±0.3	89.2±0.3
		63.3±8.5	80.9±6.9	85.9±0.7	88.8±0.3	88.5±0.4	88.3±0.3	88.2±0.3	**89.3**±0.3
FMNIST	ACC	47.5±2.6	71.1±4.2	46.9±2.0	78.6±1.4	70.0±3.7	73.1±2.1	73.5±1.5	82.0±1.6
		47.7±2.6	70.2±3.7	45.7±1.5	77.4±1.4	69.4±3.8	72.6±1.9	72.9±1.2	**81.1**±1.3
	NMI	42.6±2.6	63.1±2.6	42.7±1.6	68.8±1.2	65.9±1.2	65.9±0.7	68.1±0.8	73.7±0.8
		43.2±2.8	62.1±2.1	42.7±2.4	67.7±1.3	65.5±1.2	65.7±0.8	67.3±0.8	**72.8**±0.5
	ARI	27.0±2.2	55.6±3.1	21.5±2.8	63.5±1.3	56.3±2.0	57.5±1.2	59.6±1.1	68.7±1.2
		27.1±1.9	54.2±2.4	19.8±3.0	62.0±1.4	55.5±2.1	56.8±1.0	58.5±0.9	**67.4**±0.9
ImageNet10	ACC	39.9±3.0	73.6±10.1	79.9±0.7	92.9±0.4	93.6±0.4	93.5±0.2	94.7±0.1	95.5±0.2
		40.7±3.0	74.3±10.6	81.1±1.5	93.7±0.8	94.2±0.5	94.3±0.6	95.7±0.4	**96.2**±0.3
	NMI	32.2±1.1	66.4±8.6	63.8±0.7	84.2±0.7	86.1±0.6	85.6±0.2	88.0±0.3	89.5±0.3
		34.1±1.6	68.3±9.3	67.7±1.4	86.2±1.3	87.7±0.9	87.8±0.7	90.3±0.7	**91.3**±0.5
	ARI	23.0±1.2	60.9±12.1	60.0±1.7	85.1±0.8	86.3±0.9	86.1±0.4	88.6±0.2	90.4±0.4
		23.6±1.7	62.4±12.7	61.4±3.8	86.6±1.5	87.5±1.0	87.8±1.2	90.7±0.8	**91.9**±0.6
MNIST	ACC	56.5±2.1	74.6±4.4	85.4±1.1	94.3±0.7	92.0±0.8	92.3±0.5	91.8±0.3	96.1±0.1
		57.3±2.0	75.6±4.5	84.8±1.5	94.3±0.7	92.6±0.7	92.6±0.5	92.3±0.4	**96.3**±0.1
	NMI	49.7±2.4	68.1±3.4	76.8±0.6	87.0±1.0	82.3±1.1	83.1±0.8	82.7±0.5	89.9±0.3
		51.6±2.3	69.6±3.6	78.0±0.8	87.3±1.0	83.7±1.0	83.9±0.9	83.7±0.8	**90.6**±0.3
	ARI	40.0±4.3	63.0±4.5	68.1±1.3	88.1±1.3	83.3±1.5	83.9±1.0	83.2±0.5	91.6±0.3
		41.3±4.1	64.4±5.0	66.5±1.8	88.0±1.4	84.6±1.3	84.6±1.1	84.0±0.9	**92.0**±0.3
Reuters	ACC	88.9±0.7	91.4±1.4	89.0±0.3	92.9±0.4	85.3±0.4	85.6±0.6	90.3±0.4	92.4±0.4
		88.3±0.9	91.9±1.7	87.9±0.5	**93.9**±0.2	86.7±0.3	86.9±0.3	91.0±0.4	93.5±0.4
	NMI	68.6±1.6	75.1±2.3	68.1±0.4	77.9±0.7	70.1±0.9	70.6±0.6	75.5±0.5	77.3±0.7
		68.4±2.1	75.1±3.8	66.3±1.1	**78.8**±0.7	70.3±0.4	70.6±1.0	75.7±0.7	78.7±1.0
	ARI	77.7±1.2	81.7±2.2	78.4±0.5	84.1±0.7	71.5±0.8	71.4±0.9	79.9±0.8	83.7±0.6
		76.2±1.7	82.5±3.3	75.3±1.1	**85.7**±0.7	73.3±0.5	73.6±0.5	81.1±0.8	85.6±0.7
STL10	ACC	52.0±4.6	83.4±2.3	76.2±0.9	87.6±1.6	88.1±0.8	84.7±3.5	88.8±0.4	91.1±0.3
		52.8±4.5	83.5±2.4	76.9±0.8	86.7±1.5	87.1±0.6	83.8±3.4	87.7±0.3	**89.4**±0.3
	NMI	48.4±4.3	69.8±2.3	58.2±1.5	77.1±1.7	78.4±0.8	76.4±0.8	79.8±0.4	82.5±0.4
		52.2±4.5	71.2±2.4	61.6±1.3	76.5±1.5	77.7±0.5	75.9±0.9	78.8±0.3	**80.5**±0.5
	ARI	31.6±2.9	68.6±3.2	54.9±1.9	75.9±2.2	76.2±1.2	72.4±3.0	77.4±0.8	81.8±0.5
		32.6±3.2	68.7±3.2	56.8±1.6	74.5±2.0	74.6±0.9	70.9±2.6	75.4±0.7	**78.8**±0.6

improve absolute performance, GDP depends on it less strongly than competing methods, making its relative advantage more evident when all pretrain-capable methods start from random initialization.

Training from random initialization also removes the separate autoencoder-pretraining stage and its associated computational cost. Pretraining generally improves absolute performance, especially when pairwise supervision is sparse, but the no-pretraining results show that GDP† can retain strong relative performance without this additional stage. This makes the same framework applicable when computational resources are limited or a separate unsupervised pretraining phase is undesirable.

Comparison with unsupervised clustering. We additionally apply K-means directly to the benchmark input representations as an unsupervised reference. K-means is fitted on the training set with three initializations and evaluated over five repeats, with test samples assigned to the nearest training centroid. Table 14 shows that the input representations already contain substantial clustering structure, but GDP obtains clear further gains from pairwise supervision. At the main 5k budget, both GDP and GDP† exceed K-means in all 63 dataset–source–metric comparisons. The improvement is also substantial in absolute terms: K-means obtains 50.6% and 51.5% ACC on CIFAR100-20 and FMNIST, whereas GDP† reaches 63.2–71.3% and 77.2–81.2%, respectively, across the three supervision sources.

The comparison further shows that the availability of pairwise constraints alone does not guarantee an improvement over unsupervised clustering. Under Difficulty supervision at 5k, for example,

Table 7: Training (blue) and test (black) ACC/NMI/ARI (%; five-run mean±std) under Mix supervision at 5k constraints. Best/second-best test means are **bold**/underlined.

Dataset	Metric	Vanilla DCC	VolMax DCC	CIDEC	SpherePair	ProbPair	WPP	ECI-PP	**GDP**
CIFAR100-20	ACC	48.2±3.4 48.3±3.3	41.9±4.3 42.0±4.1	59.3±1.9 59.4±2.0	64.3±3.3 63.9±3.3	68.5±1.3 **68.5**±1.2	65.7±3.9 65.5±3.4	62.8±2.9 62.6±2.9	64.0±2.1 64.0±2.2
	NMI	48.4±1.6 49.4±1.7	45.9±2.4 46.6±2.3	58.9±1.3 59.8±1.5	60.7±1.5 61.0±1.4	63.0±1.0 **63.7**±1.1	61.9±1.5 62.5±1.4	61.1±1.1 61.6±1.1	61.4±1.7 62.3±1.7
	ARI	33.9±2.1 34.1±2.1	27.2±3.4 27.3±3.3	44.8±2.2 44.9±2.2	47.7±2.4 47.3±2.3	51.9±1.7 **51.9**±1.6	49.7±2.6 49.6±2.3	48.0±1.4 47.8±1.5	47.5±3.0 47.7±3.0
CIFAR10	ACC	90.8±2.2 90.7±2.2	86.1±7.3 86.2±7.1	92.5±0.6 92.9±0.5	92.9±1.0 92.9±1.0	94.6±0.4 94.6±0.3	94.7±0.2 94.6±0.2	94.1±0.5 94.0±0.5	95.2±0.1 **95.1**±0.2
	NMI	85.5±1.4 85.6±1.3	81.7±5.7 81.9±5.4	85.6±0.8 86.3±0.7	86.6±1.1 86.8±1.2	88.7±0.5 88.8±0.4	88.8±0.1 88.8±0.2	88.3±0.6 88.3±0.5	89.5±0.1 **89.4**±0.2
	ARI	83.6±2.4 83.5±2.3	77.4±9.0 77.5±8.7	84.8±1.0 85.4±0.9	85.9±1.6 85.9±1.6	88.8±0.6 88.8±0.5	89.0±0.3 88.8±0.3	87.9±0.9 87.8±0.8	89.9±0.2 **89.7**±0.4
FMNIST	ACC	73.3±4.6 72.8±4.7	67.6±2.3 67.0±2.5	71.1±3.0 70.6±2.9	74.2±2.0 73.4±1.9	70.5±3.0 70.0±2.9	72.8±2.8 71.8±2.8	70.0±3.7 69.3±3.7	80.5±2.2 **79.3**±2.0
	NMI	64.8±2.2 64.4±2.3	61.9±1.4 61.4±1.6	65.2±1.8 64.8±1.6	66.2±1.1 65.4±1.4	64.9±1.4 64.4±1.4	66.2±1.3 65.4±1.5	66.3±1.2 65.6±1.3	72.2±1.3 **71.3**±1.1
	ARI	58.5±3.1 57.7±3.2	53.2±2.1 52.6±2.2	57.0±2.5 56.4±2.4	58.7±1.9 57.7±1.9	56.0±2.2 55.2±2.3	57.8±2.2 56.7±2.4	56.5±2.1 55.6±2.0	66.6±1.9 **65.1**±1.5
ImageNet10	ACC	84.5±6.0 84.9±6.0	67.3±5.8 68.4±6.1	94.9±0.2 95.5±0.3	92.8±0.8 93.6±0.9	95.5±0.2 96.2±0.3	95.4±0.2 96.0±0.4	95.0±0.2 96.0±0.3	95.9±0.2 **96.5**±0.1
	NMI	78.4±5.6 80.1±5.1	61.1±6.7 62.9±6.5	88.1±0.5 89.8±0.6	84.1±1.1 86.1±1.5	89.3±0.3 90.9±0.7	89.0±0.4 90.5±0.7	88.4±0.5 90.7±0.8	90.1±0.3 **91.7**±0.3
	ARI	75.7±7.8 76.6±7.5	54.3±8.1 55.4±8.0	89.1±0.5 90.3±0.7	84.7±1.5 86.4±1.8	90.4±0.3 91.7±0.7	90.1±0.5 91.3±0.8	89.4±0.5 91.3±0.7	91.2±0.4 **92.4**±0.3
MNIST	ACC	84.4±3.9 85.0±4.0	67.6±5.5 68.3±5.7	94.4±0.5 94.0±0.7	89.2±2.8 89.7±2.8	92.2±0.2 92.5±0.2	93.1±0.5 93.2±0.5	91.3±0.7 91.6±0.7	95.6±0.2 **95.8**±0.2
	NMI	72.6±3.1 73.8±3.3	60.5±4.2 61.2±4.5	88.1±0.6 87.3±0.8	81.7±1.6 82.5±1.9	82.7±0.3 83.5±0.3	84.3±0.7 84.8±0.8	82.0±1.0 82.8±0.9	89.0±0.4 **89.6**±0.4
	ARI	71.5±4.5 72.5±4.8	54.0±6.5 54.6±6.8	88.3±0.9 87.1±1.4	80.3±3.2 80.9±3.6	83.7±0.4 84.4±0.3	85.4±1.0 85.8±0.9	82.2±1.3 82.7±1.3	90.6±0.4 **91.0**±0.4
Reuters	ACC	93.2±3.9 92.2±3.5	93.3±0.7 93.1±0.6	91.7±1.8 90.1±2.1	94.6±0.4 94.3±0.2	95.5±0.2 **94.7**±0.4	95.8±0.3 **94.7**±0.4	94.4±0.5 94.1±0.5	95.1±0.2 94.5±0.5
	NMI	78.5±5.3 75.7±4.6	77.3±1.7 76.8±1.5	74.7±3.0 71.0±3.3	80.7±1.2 80.0±0.8	82.6±0.4 **80.6**±1.2	83.5±0.8 80.4±1.2	80.3±1.0 79.6±1.2	82.1±0.5 80.2±1.3
	ARI	86.3±3.8 83.6±3.5	85.1±1.2 84.3±1.2	84.1±2.0 80.2±2.6	87.6±0.9 86.8±0.5	88.8±0.3 **87.1**±0.7	89.6±0.6 87.0±1.1	86.7±0.9 86.4±1.1	88.6±0.5 87.0±1.0
STL10	ACC	86.0±4.9 84.2±4.9	71.6±3.0 71.9±2.8	89.8±0.4 88.1±0.7	86.0±2.5 85.3±2.6	90.5±0.4 88.9±0.2	90.2±0.3 88.7±0.1	88.7±0.3 87.7±0.2	91.5±0.4 **89.8**±0.5
	NMI	77.4±3.9 75.3±3.9	62.6±3.2 63.7±2.8	80.6±0.6 78.6±0.9	76.1±1.6 75.9±1.9	81.4±0.5 79.7±0.3	81.1±0.4 79.3±0.1	79.3±0.3 78.4±0.2	83.1±0.6 **81.0**±0.7
	ARI	75.2±5.7 72.2±5.8	50.7±4.2 52.0±3.5	79.4±0.7 76.4±1.2	73.9±2.9 72.7±3.1	80.7±0.8 77.9±0.4	80.3±0.5 77.6±0.2	77.6±0.5 75.9±0.4	82.6±0.7 **79.6**±0.9

VanillaDCC remains below K-means on all three metrics for five of the seven datasets, whereas both GDP and GDP† remain above the unsupervised reference throughout. Together with the no-pretraining results, this indicates that GDP effectively converts the additional pairwise information into improved clustering performance, rather than relying primarily on favorable input features or autoencoder initialization.

Cross-source and aggregate consistency. Figure 5 compares test ACC across Binary, Difficulty, and Mix supervision for every dataset and constraint budget. GDP remains consistently at or near the best across supervision sources, whereas the strongest competing method changes with the dataset and source. At 5k constraints, GDP achieves the best ACC in 17/21 dataset–source cases; the remaining leading positions are distributed among SpherePair, ProbPair, WPP, and ECI-PP rather than being dominated by a single alternative. The same overall pattern persists at 1k and 10k constraints, indicating that GDP's advantage is not tied to one particular supervision source.

Figure 6 provides a broader view by measuring, for each metric and budget, the gap from each method to the best result in each of the 21 dataset–source cases. At 5k constraints, GDP is best in 17/21 cases for each of ACC, NMI, and ARI, and lies within one percentage point of the best in 20/21 cases for every metric. This concentration near the optimum persists across budgets: at 1k, GDP is within one percentage point of the best in 19/21, 17/21, and 18/21 cases for ACC, NMI, and ARI, respectively, while at 10k the corresponding count is 20/21 for all three metrics. Competing methods can lead

Table 8: Test ACC/NMI/ARI (%; five-run means) at 1k constraints. Best/second-best are **bold**/underlined per dataset/source/metric; WPP reduces to ProbPair under Binary and is omitted.

Method	CIFAR100-20	CIFAR10	FMNIST	ImageNet10	MNIST	Reuters	STL10
Binary							
VanillaDCC	26.3 / 28.7 / 12.5	59.4 / 56.8 / 43.9	59.3 / 56.0 / 44.3	41.4 / 34.3 / 23.3	52.7 / 46.7 / 36.2	64.1 / 34.4 / 37.6	62.3 / 59.4 / 47.2
VolMaxDCC	41.8 / 47.3 / 26.8	74.1 / 72.2 / 63.2	60.0 / 56.4 / 44.2	54.5 / 47.3 / 36.0	62.9 / 58.7 / 49.0	81.8 / 60.5 / 66.9	59.4 / 54.6 / 37.9
CIDEC	56.0 / **60.7** / 42.7	90.9 / 84.7 / 82.2	57.0 / 60.5 / 44.5	94.5 / 88.3 / 88.2	88.1 / 86.3 / 82.6	77.2 / 59.5 / 66.6	87.6 / 78.5 / 75.6
SpherePair	58.1 / 58.5 / 42.4	92.8 / 86.5 / 85.5	72.6 / 63.7 / 55.3	94.1 / 87.7 / 87.4	93.5 / 85.5 / 86.1	**91.9 / 73.7 / 82.0**	87.1 / 77.4 / 74.9
ProbPair	53.3 / 55.0 / 38.1	92.7 / 86.7 / 85.5	52.4 / 51.9 / 35.8	94.1 / 87.5 / 87.5	90.8 / 81.6 / 81.3	85.0 / 66.6 / 74.8	85.9 / 76.2 / 73.0
WPP	–	–	–	–	–	–	–
ECI-PP	54.9 / 57.5 / 40.1	88.9 / 84.5 / 81.2	65.1 / 62.1 / 50.3	94.9 / 88.8 / 88.9	80.3 / 72.5 / 67.2	89.9 / 70.4 / 78.9	85.3 / 75.9 / 72.0
GDP (Ours)	**58.2** / 59.3 / **43.3**	**94.0 / 88.2 / 87.7**	**73.3 / 67.5 / 58.8**	**95.4 / 89.6 / 90.0**	**95.8 / 89.7 / 91.0**	91.3 / 73.6 / 79.0	**88.7 / 79.5 / 77.6**
Difficulty							
VanillaDCC	16.9 / 15.2 / 5.3	47.2 / 44.5 / 29.7	43.5 / 39.6 / 24.8	26.5 / 16.9 / 8.2	45.0 / 38.2 / 28.3	67.1 / 36.9 / 40.2	36.6 / 34.7 / 20.5
VolMaxDCC	42.0 / 47.5 / 27.5	78.0 / 75.4 / 67.6	63.9 / 58.6 / 47.6	59.0 / 52.6 / 42.5	56.3 / 52.3 / 40.7	75.8 / 53.3 / 59.6	70.4 / 65.5 / 57.2
CIDEC	42.1 / 46.3 / 20.4	92.0 / 85.7 / 84.0	47.6 / 46.5 / 23.0	94.2 / 87.6 / 87.7	88.8 / 85.9 / 82.6	76.5 / 54.8 / 60.5	86.6 / 76.6 / 73.9
SpherePair	56.5 / 58.5 / 41.2	92.8 / 86.8 / 85.7	69.6 / 63.6 / 54.0	93.1 / 85.9 / 85.4	89.8 / 81.6 / 80.1	**91.9 / 73.5 / 81.4**	86.3 / 76.2 / 73.4
ProbPair	53.5 / 55.0 / 37.4	90.9 / 84.7 / 82.6	53.9 / 51.7 / 36.7	91.4 / 83.7 / 82.2	83.5 / 73.3 / 70.5	88.2 / 69.3 / 75.0	83.7 / 73.8 / 69.5
WPP	54.3 / 56.2 / 38.4	92.5 / 86.2 / 85.0	53.3 / 51.7 / 36.5	91.8 / 83.9 / 82.9	87.8 / 77.1 / 75.7	89.8 / 70.2 / 78.3	84.0 / 74.4 / 70.2
ECI-PP	53.6 / 56.8 / 38.3	91.0 / 85.4 / 83.0	63.6 / 60.9 / 49.2	93.6 / 87.2 / 86.5	78.6 / 70.3 / 65.0	89.5 / 69.9 / 77.5	85.1 / 75.6 / 71.5
GDP (Ours)	**58.6 / 59.2 / 43.2**	**94.2 / 88.3 / 88.1**	**75.6 / 69.1 / 61.4**	**95.6 / 89.9 / 90.4**	**94.7 / 90.0 / 90.3**	88.7 / 70.6 / 75.4	**88.8 / 79.8 / 77.6**
Mix							
VanillaDCC	31.0 / 32.9 / 16.4	52.4 / 52.1 / 39.3	55.2 / 52.8 / 41.8	43.9 / 32.6 / 22.4	54.5 / 45.7 / 36.9	65.3 / 32.6 / 39.0	61.2 / 57.0 / 45.8
VolMaxDCC	38.1 / 45.1 / 24.8	68.5 / 65.6 / 54.7	59.8 / 55.8 / 44.0	49.9 / 41.4 / 30.1	56.2 / 50.9 / 40.1	67.2 / 41.6 / 44.0	63.8 / 56.3 / 43.1
CIDEC	55.6 / **59.4** / **41.6**	91.0 / 84.6 / 82.3	57.0 / 58.0 / 44.0	94.1 / 87.5 / 87.5	**91.3 / 86.3 / 84.5**	75.6 / 56.0 / 62.8	86.3 / 76.8 / 73.3
SpherePair	56.5 / 58.1 / 41.2	91.8 / 85.2 / 83.7	61.6 / 58.5 / 47.0	91.6 / 83.7 / 82.6	84.4 / 77.7 / 73.9	**89.8 / 69.1 / 77.7**	84.6 / 74.5 / 70.7
ProbPair	51.3 / 53.5 / 35.9	90.7 / 84.0 / 81.9	50.9 / 50.5 / 35.9	91.7 / 83.1 / 82.6	82.5 / 73.2 / 69.5	89.2 / 68.3 / 76.6	84.9 / 74.6 / 71.2
WPP	51.1 / 53.2 / 35.4	91.3 / 84.7 / 82.8	51.5 / 50.5 / 35.8	91.3 / 82.5 / 81.9	85.1 / 75.2 / 73.1	86.7 / 66.6 / 73.3	84.8 / 74.4 / 70.9
ECI-PP	52.8 / 56.4 / 38.1	90.7 / 83.9 / 81.6	65.8 / 62.1 / 50.6	92.7 / 85.3 / 84.7	75.4 / 68.7 / 62.2	86.6 / 65.8 / 72.1	84.0 / 74.3 / 70.1
GDP (Ours)	**56.7** / 58.3 / 41.4	**93.8 / 87.8 / 87.3**	**71.3 / 65.2 / 56.5**	**95.4 / 89.6 / 90.0**	91.0 / 85.4 / 84.3	81.3 / 55.8 / 60.1	**88.8 / 79.6 / 77.6**

Table 9: Test ACC/NMI/ARI (%; five-run means) at 5k constraints. Best/second-best are **bold**/underlined per dataset/source/metric; WPP reduces to ProbPair under Binary and is omitted.

Method	CIFAR100-20	CIFAR10	FMNIST	ImageNet10	MNIST	Reuters	STL10
Binary							
VanillaDCC	48.9 / 51.7 / 35.0	88.1 / 84.4 / 81.3	76.7 / 67.3 / 61.3	88.7 / 82.6 / 80.3	82.1 / 74.8 / 71.5	95.0 / 81.2 / 87.9	82.2 / 74.8 / 70.7
VolMaxDCC	44.6 / 47.8 / 29.5	89.1 / 84.5 / 82.1	68.6 / 62.9 / 54.4	77.5 / 71.6 / 66.7	81.1 / 73.1 / 71.0	93.4 / 76.7 / 84.3	78.9 / 71.4 / 62.6
CIDEC	57.6 / 60.5 / 43.6	94.8 / 89.0 / 89.1	67.6 / 63.6 / 53.8	95.5 / 89.9 / 90.4	96.6 / 91.8 / 92.7	82.3 / 69.5 / 77.6	88.9 / 79.6 / 77.8
SpherePair	69.4 / 64.5 / 53.4	94.7 / 88.8 / 89.0	79.5 / 71.2 / 65.2	95.8 / 90.2 / 90.9	96.4 / 90.9 / 92.2	94.5 / 79.9 / 86.6	89.3 / 80.3 / 78.7
ProbPair	66.6 / 64.0 / 50.7	94.7 / 88.8 / 88.9	76.9 / 68.7 / 61.3	95.8 / 90.2 / 90.9	94.8 / 87.8 / 88.9	94.9 / 81.0 / 87.7	88.7 / 79.6 / 77.5
WPP	–	–	–	–	–	–	–
ECI-PP	63.9 / 63.2 / 49.3	94.6 / 88.9 / 88.9	70.2 / 66.0 / 56.5	**96.3 / 91.5 / 92.1**	93.5 / 85.9 / 86.5	94.8 / 81.3 / 88.0	88.9 / 79.9 / 78.0
GDP (Ours)	**71.0 / 66.2 / 54.7**	**94.9 / 89.3 / 89.4**	**81.8 / 73.4 / 68.1**	96.1 / 91.0 / 91.7	**97.1 / 92.5 / 93.8**	**95.2 / 82.1 / 88.5**	**89.9 / 81.2 / 79.7**
Difficulty							
VanillaDCC	17.4 / 15.3 / 4.1	74.3 / 70.2 / 63.3	47.7 / 43.2 / 27.1	40.7 / 34.1 / 23.6	57.3 / 51.6 / 41.3	88.3 / 68.4 / 76.2	52.8 / 52.2 / 32.6
VolMaxDCC	45.4 / 49.0 / 30.6	87.1 / 83.4 / 80.9	70.2 / 62.1 / 54.2	74.3 / 68.3 / 62.4	75.6 / 69.6 / 64.4	91.9 / 75.1 / 82.5	83.5 / 71.2 / 68.7
CIDEC	25.9 / 29.7 / 2.5	93.2 / 85.7 / 85.9	45.7 / 42.7 / 19.8	81.1 / 67.7 / 61.4	84.8 / 78.0 / 66.5	87.9 / 66.3 / 75.3	76.9 / 61.6 / 56.8
SpherePair	68.0 / 62.9 / 50.8	94.7 / 88.6 / 88.8	77.4 / 67.7 / 62.0	93.7 / 86.2 / 86.6	94.3 / 87.3 / 88.0	**93.9 / 78.8 / 85.7**	86.7 / 76.5 / 74.5
ProbPair	60.8 / 59.5 / 42.3	94.5 / 88.6 / 88.5	69.4 / 65.5 / 55.5	94.2 / 87.7 / 87.5	92.6 / 83.7 / 84.6	86.7 / 70.3 / 73.3	87.1 / 77.7 / 74.6
WPP	63.2 / 61.5 / 45.7	94.4 / 88.5 / 88.3	72.6 / 65.7 / 56.8	94.3 / 87.8 / 87.8	92.6 / 83.9 / 84.6	86.9 / 70.6 / 73.6	83.8 / 75.9 / 70.9
ECI-PP	64.3 / 62.6 / 47.9	94.3 / 88.4 / 88.2	72.9 / 67.3 / 58.5	95.7 / 90.3 / 90.7	92.3 / 83.7 / 84.0	91.0 / 75.7 / 81.1	87.7 / 78.8 / 75.4
GDP (Ours)	**69.9 / 64.5 / 52.9**	**94.9 / 89.1 / 89.3**	**81.1 / 72.8 / 67.4**	**96.2 / 91.3 / 91.9**	**96.3 / 90.6 / 92.0**	93.5 / 78.7 / 85.6	**89.4 / 80.5 / 78.8**
Mix							
VanillaDCC	48.3 / 49.4 / 34.1	90.7 / 85.6 / 83.5	72.8 / 64.4 / 57.7	84.9 / 80.1 / 76.6	85.0 / 73.8 / 72.5	92.2 / 75.7 / 83.6	84.2 / 75.3 / 72.2
VolMaxDCC	42.0 / 46.6 / 27.3	86.2 / 81.9 / 77.5	67.0 / 61.4 / 52.6	68.4 / 62.9 / 55.4	68.3 / 61.2 / 54.6	93.1 / 76.8 / 84.3	71.9 / 63.7 / 52.0
CIDEC	59.4 / 59.8 / 44.9	92.9 / 86.3 / 85.4	70.6 / 64.8 / 56.4	95.5 / 89.8 / 90.3	94.0 / 87.3 / 87.1	90.1 / 71.0 / 80.2	88.1 / 78.6 / 76.4
SpherePair	63.9 / 61.0 / 47.3	92.9 / 86.8 / 85.9	73.4 / 65.4 / 57.7	93.6 / 86.1 / 86.4	89.7 / 82.5 / 80.9	94.3 / 80.0 / 86.8	85.3 / 75.9 / 72.7
ProbPair	**68.5 / 63.7 / 51.9**	94.6 / 88.8 / 88.8	70.0 / 64.4 / 55.2	96.2 / 90.9 / 91.7	92.5 / 83.5 / 84.4	**94.7 / 80.6 / 87.1**	88.9 / 79.7 / 77.9
WPP	65.5 / 62.5 / 49.6	94.6 / 88.8 / 88.8	71.8 / 65.4 / 56.7	96.0 / 90.5 / 91.3	93.2 / 84.8 / 85.8	**94.7** / 80.4 / 87.0	88.7 / 79.3 / 77.6
ECI-PP	62.6 / 61.6 / 47.8	94.0 / 88.3 / 87.8	69.3 / 65.6 / 55.6	96.0 / 90.7 / 91.3	91.6 / 82.8 / 82.7	94.1 / 79.6 / 86.4	87.7 / 78.4 / 75.9
GDP (Ours)	64.0 / 62.3 / 47.7	**95.1 / 89.4 / 89.7**	**79.3 / 71.3 / 65.1**	**96.5 / 91.7 / 92.4**	**95.8 / 89.6 / 91.0**	94.5 / 80.2 / 87.0	**89.8 / 81.0 / 79.6**

in individual conditions, but none maintains a similarly consistent near-best profile across datasets, supervision sources, metrics, and constraint budgets.

Table 10: Test ACC/NMI/ARI (%; five-run means) at 10k constraints. Best/second-best are **bold**/underlined per dataset/source/metric; WPP reduces to ProbPair under Binary and is omitted.

Method	CIFAR100-20	CIFAR10	FMNIST	ImageNet10	MNIST	Reuters	STL10
Binary							
VanillaDCC	63.8 / 61.6 / 49.0	94.7 / 88.6 / 88.9	77.5 / 69.0 / 62.5	95.8 / 90.3 / 91.0	94.1 / 86.0 / 87.6	94.2 / 79.1 / 86.0	89.6 / 80.3 / 79.0
VolMaxDCC	45.1 / 49.2 / 31.8	94.5 / 88.4 / 88.5	75.4 / 67.6 / 61.1	95.4 / 89.5 / 90.2	84.1 / 78.0 / 75.6	94.8 / 80.8 / 87.6	84.4 / 76.9 / 71.6
CIDEC	60.8 / 61.9 / 46.4	95.0 / 89.3 / 89.5	74.2 / 68.3 / 60.9	96.4 / 91.6 / 92.2	97.2 / 92.9 / 94.0	94.6 / 82.2 / 88.4	89.2 / 80.0 / 78.4
SpherePair	75.8 / 68.1 / 59.8	95.1 / 89.4 / 89.6	83.6 / 74.5 / 70.1	96.1 / 90.8 / 91.6	97.1 / 92.3 / 93.7	**95.9** / 84.0 / 89.9	**90.1** / 81.2 / 80.0
ProbPair	71.8 / 67.0 / 56.1	95.0 / 89.3 / 89.4	81.8 / 72.5 / 67.5	**96.6** / 91.7 / 92.5	95.6 / 89.4 / 90.6	95.3 / 82.3 / 88.5	89.7 / 80.8 / 79.2
WPP	–	–	–	–	–	–	–
ECI-PP	69.5 / 66.9 / 55.0	94.9 / 89.2 / 89.3	76.6 / 69.7 / 62.0	**96.6 / 91.9 / 92.6**	95.9 / 90.0 / 91.2	**95.9 / 84.1 / 90.0**	89.7 / 80.9 / 79.3
GDP (Ours)	**76.1 / 69.9 / 60.8**	**95.2 / 89.6 / 89.9**	**83.9 / 75.2 / 70.6**	96.4 / 91.4 / 92.1	**97.5 / 93.4 / 94.6**	95.8 / 83.8 / 89.6	**90.1 / 81.5 / 80.1**
Difficulty							
VanillaDCC	20.5 / 19.1 / 4.7	61.3 / 53.7 / 44.8	49.4 / 44.6 / 28.5	52.9 / 46.0 / 32.1	65.3 / 59.8 / 50.3	84.1 / 59.3 / 66.9	58.6 / 54.1 / 35.3
VolMaxDCC	45.8 / 48.9 / 30.0	91.2 / 86.5 / 85.3	74.2 / 66.9 / 59.6	77.3 / 72.3 / 65.6	88.5 / 82.1 / 81.3	92.3 / 75.8 / 83.3	84.7 / 73.5 / 71.6
CIDEC	25.5 / 29.1 / 3.1	93.4 / 85.8 / 86.1	46.1 / 46.8 / 24.3	83.7 / 69.3 / 66.7	87.5 / 80.1 / 71.0	87.4 / 65.8 / 74.0	77.4 / 62.4 / 58.5
SpherePair	73.0 / 66.0 / 56.6	94.9 / 88.8 / 89.2	80.5 / 70.4 / 65.8	92.9 / 84.6 / 85.1	95.8 / 89.7 / 91.0	93.6 / 78.9 / 85.7	88.4 / 78.2 / 76.9
ProbPair	68.5 / 64.4 / 46.9	94.7 / 88.8 / 88.9	77.2 / 69.3 / 60.7	95.0 / 88.7 / 89.1	93.7 / 85.5 / 86.7	86.6 / 70.7 / 73.2	88.4 / 79.0 / 76.6
WPP	68.5 / 64.2 / 49.1	94.8 / 88.9 / 89.1	79.6 / 69.9 / 63.5	94.4 / 87.7 / 88.1	94.0 / 86.0 / 87.3	86.6 / 70.9 / 73.0	87.7 / 77.6 / 75.4
ECI-PP	69.4 / 65.7 / 51.7	94.7 / 88.9 / 88.9	76.1 / 69.3 / 61.6	96.0 / 90.7 / **91.4**	94.2 / 86.6 / 87.8	91.8 / 77.6 / 83.1	89.0 / 80.0 / 78.0
GDP (Ours)	**76.7 / 69.1 / 60.9**	**95.1 / 89.4 / 89.7**	**84.0 / 74.7 / 70.4**	**96.1 / 90.8 / 91.4**	**96.1 / 90.1 / 91.6**	**94.1 / 79.5 / 86.6**	**89.6 / 80.9 / 79.1**
Mix							
VanillaDCC	60.7 / 58.6 / 46.5	94.6 / 88.7 / 88.8	79.2 / 69.8 / 64.7	95.6 / 89.8 / 90.5	92.8 / 83.4 / 84.9	90.8 / 71.2 / 79.2	89.7 / 80.3 / 79.2
VolMaxDCC	47.8 / 50.5 / 33.1	85.7 / 82.8 / 78.9	72.7 / 65.8 / 58.8	83.2 / 78.4 / 74.9	86.8 / 79.0 / 77.9	95.2 / 82.2 / 88.5	83.2 / 75.0 / 68.4
CIDEC	62.7 / 62.7 / 49.6	93.7 / 87.1 / 86.9	73.2 / 68.9 / 60.8	96.1 / 91.0 / 91.6	95.9 / 90.2 / 91.2	94.6 / 80.3 / 87.0	88.8 / 79.5 / 77.6
SpherePair	67.5 / 62.5 / 50.8	93.6 / 87.4 / 87.0	76.9 / 67.6 / 61.2	94.2 / 86.8 / 87.6	94.0 / 87.1 / 87.7	95.1 / 82.0 / 88.5	88.4 / 78.6 / 76.9
ProbPair	75.1 / **68.4 / 59.3**	95.1 / 89.5 / 89.7	78.5 / 69.9 / 63.4	**96.6 / 91.8 / 92.6**	94.7 / 87.2 / 88.7	**95.5 / 83.1 / 89.0**	89.9 / 80.8 / 79.6
WPP	**75.3** / 68.0 / 58.9	95.1 / 89.4 / 89.7	80.5 / 71.0 / 65.5	**96.6** / 91.7 / **92.6**	95.0 / 87.9 / 89.3	**95.5** / 82.9 / 88.9	89.6 / 80.4 / 79.1
ECI-PP	69.1 / 65.2 / 53.2	94.7 / 89.0 / 88.9	72.9 / 66.9 / 58.1	96.5 / 91.7 / 92.4	94.1 / 86.6 / 87.5	95.3 / 82.6 / 88.8	89.2 / 80.1 / 78.4
GDP (Ours)	72.4 / 66.5 / 55.9	**95.2 / 89.6 / 90.0**	**83.3 / 74.3 / 69.6**	96.5 / 91.6 / 92.4	**96.5 / 91.0 / 92.4**	95.1 / 82.2 / 88.3	**90.2 / 81.6 / 80.2**

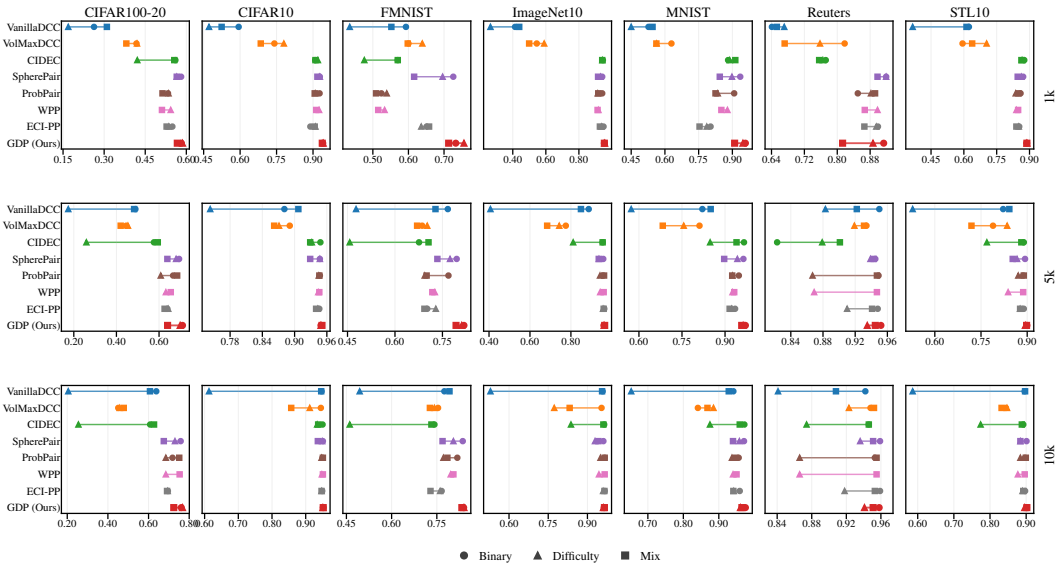

Figure 5: Cross-source test ACC at 1k, 5k, and 10k constraints. Markers denote Binary, Difficulty, and Mix supervision, with their span reflecting performance variation across sources. WPP is omitted for Binary, for which it coincides with ProbPair.

D.2 Structured mixed-membership recovery

Under Mix supervision, ACC/NMI/ARI on the original hard-labeled samples measure clustering quality but do not reveal whether the assignment of a mixed sample preserves its underlying partial membership. We therefore evaluate held-out mixed samples directly in the aligned Global assignment space. For a sample generated from source classes u and v with mixing coefficient λ, let q_u and q_v denote the assignment mass aligned to these two classes. Source-class mass, $q_u + q_v$, measures how

Table 11: Test ACC/NMI/ARI (%; five-run means) at 1k constraints. † denotes the random-initialization variant without autoencoder pretraining; VanillaDCC and VolMaxDCC are unchanged. Best/second-best are **bold**/underlined per dataset/source/metric; WPP† reduces to ProbPair† under Binary and is omitted.

Method	CIFAR100-20	CIFAR10	FMNIST	ImageNet10	MNIST	Reuters	STL10
Binary							
VanillaDCC	26.3 / 28.7 / 12.5	59.4 / 56.8 / 43.9	59.3 / 56.0 / 44.3	41.4 / 34.3 / 23.3	52.7 / 46.7 / 36.2	64.1 / 34.4 / 37.6	62.3 / 59.4 / 47.2
VolMaxDCC	41.8 / 47.3 / 26.8	74.1 / 72.2 / 63.2	60.0 / 56.4 / 44.2	54.5 / 47.3 / 36.0	62.9 / 58.7 / 49.0	81.8 / 60.5 / 66.9	59.4 / 54.6 / 37.9
CIDEC†	38.0 / 43.8 / 24.0	67.2 / 66.2 / 55.1	51.7 / 53.5 / 37.6	71.2 / 67.2 / 58.1	61.4 / 61.2 / 49.2	67.5 / 44.7 / 47.3	63.4 / 60.5 / 47.3
SpherePair†	54.8 / 57.4 / 40.4	90.5 / 84.2 / 82.0	64.3 / 60.4 / 49.4	90.6 / 81.2 / 80.6	83.5 / 74.5 / 70.9	88.9 / 67.7 / 76.3	84.4 / 74.1 / 70.6
ProbPair†	36.8 / 38.6 / 20.9	84.4 / 76.0 / 72.0	55.3 / 53.4 / 38.7	63.8 / 52.4 / 45.0	69.0 / 59.5 / 51.1	67.4 / 46.5 / 48.4	69.8 / 60.5 / 52.4
WPP†	–	–	–	–	–	–	–
ECI-PP†	49.4 / 53.5 / 34.5	90.1 / 83.4 / 80.6	60.6 / 58.6 / 45.3	85.3 / 74.9 / 71.3	71.3 / 63.0 / 55.4	81.3 / 60.0 / 66.4	83.9 / 74.5 / 69.7
GDP† (Ours)	**55.3 / 57.7 / 41.2**	**94.1 / 88.3 / 88.0**	**69.5 / 64.3 / 54.8**	**93.5 / 86.7 / 86.3**	**85.8 / 79.0 / 76.5**	**90.0 / 69.5 / 76.8**	**86.1 / 77.7 / 74.3**
Difficulty							
VanillaDCC	16.9 / 15.2 / 5.3	47.2 / 44.5 / 29.7	43.5 / 39.6 / 24.8	26.5 / 16.9 / 8.2	45.0 / 38.2 / 28.3	67.1 / 36.9 / 40.2	36.6 / 34.7 / 20.5
VolMaxDCC	42.0 / 47.5 / 27.5	78.0 / 75.4 / 67.6	63.9 / 58.6 / 47.6	59.0 / 52.6 / 42.5	56.3 / 52.3 / 40.7	75.8 / 53.3 / 59.6	70.4 / 65.5 / 57.2
CIDEC†	24.0 / 28.3 / 8.6	66.2 / 66.0 / 54.3	41.1 / 39.6 / 20.9	76.2 / 70.2 / 63.3	65.2 / 65.1 / 52.6	74.2 / 48.8 / 54.2	59.7 / 60.4 / 45.2
SpherePair†	**53.1 / 55.7 / 38.2**	87.2 / 80.5 / 77.2	62.6 / 59.3 / 47.6	84.6 / 74.0 / 71.3	78.5 / 68.3 / 64.0	**88.2 / 66.8 / 75.7**	83.4 / 73.0 / 68.7
ProbPair†	41.3 / 43.1 / 25.0	79.1 / 70.5 / 63.7	57.3 / 55.2 / 41.4	72.8 / 59.7 / 53.3	64.4 / 54.5 / 47.4	77.9 / 53.7 / 57.7	71.8 / 62.8 / 54.2
WPP†	42.4 / 43.6 / 25.9	79.1 / 70.4 / 64.6	56.9 / 53.8 / 40.2	78.5 / 66.0 / 61.1	67.1 / 56.4 / 49.0	79.3 / 56.9 / 61.8	77.4 / 68.0 / 61.7
ECI-PP†	51.1 / 54.6 / 35.9	88.4 / 83.4 / 80.2	61.1 / 59.1 / 46.2	**88.8 / 79.6 / 77.4**	73.7 / 65.4 / 58.9	80.1 / 61.6 / 65.1	**83.6** / 74.8 / 69.3
GDP† (Ours)	52.1 / 54.1 / 36.2	**90.9 / 86.0 / 83.8**	**66.1 / 59.8 / 49.0**	82.8 / 74.8 / 70.5	**80.1 / 73.0 / 68.3**	86.3 / 66.0 / 73.6	81.9 / **75.3 / 70.7**
Mix							
VanillaDCC	31.0 / 32.9 / 16.4	52.4 / 52.1 / 39.3	55.2 / 52.8 / 41.8	43.9 / 32.6 / 22.4	54.5 / 45.7 / 36.9	65.3 / 32.6 / 39.0	61.2 / 57.0 / 45.8
VolMaxDCC	38.1 / 45.1 / 24.8	68.5 / 65.6 / 54.7	59.8 / 55.8 / 44.0	49.9 / 41.4 / 30.1	56.2 / 50.9 / 40.1	67.2 / 41.6 / 44.0	63.8 / 56.3 / 43.1
CIDEC†	30.9 / 34.6 / 16.2	59.9 / 59.3 / 44.9	50.7 / 48.6 / 33.5	59.3 / 57.3 / 44.4	56.0 / 52.7 / 40.1	71.3 / 45.7 / 51.5	57.0 / 56.0 / 42.4
SpherePair†	49.1 / 51.4 / 33.1	81.7 / 77.0 / 71.2	58.3 / 56.4 / 43.1	67.3 / 57.0 / 49.9	68.8 / 60.8 / 53.5	**74.1** / 50.6 / **54.0**	76.1 / 67.9 / 61.3
ProbPair†	35.8 / 37.1 / 20.2	73.3 / 64.9 / 56.6	56.4 / 54.6 / 40.6	60.2 / 47.5 / 38.8	62.7 / 52.0 / 43.9	**74.1** / 46.0 / 51.7	74.6 / 63.8 / 56.9
WPP†	36.2 / 38.4 / 21.1	77.9 / 70.2 / 63.8	57.2 / 53.6 / 41.0	50.7 / 39.5 / 30.0	58.4 / 50.2 / 40.6	61.3 / 38.5 / 38.2	69.5 / 59.1 / 51.3
ECI-PP†	45.4 / 50.6 / 31.6	87.0 / 81.6 / 77.5	58.5 / **58.2** / 44.3	80.1 / 68.7 / 63.6	67.4 / 59.4 / 51.2	71.9 / **51.2** / 52.6	81.3 / 71.2 / 65.6
GDP† (Ours)	**53.5 / 56.3 / 38.9**	**90.9 / 86.4 / 84.6**	**63.7** / 58.1 / **46.8**	**89.2 / 82.0 / 80.2**	**72.3 / 65.5 / 58.6**	73.7 / 47.4 / 50.0	**86.0 / 77.1 / 73.9**

strongly the learned assignment concentrates on the generating classes, while source top-2 hit rate records whether they form the two largest assignment components. Their corresponding uniform and random-pair references provide natural baselines for source-identity recovery. We further estimate the relative source contribution as

$$\hat{\lambda} = \frac{q_u}{q_u + q_v},$$

whose Pearson correlation with the generating λ measures recovery of the continuous mixture composition. Table 15 reports the complete values, while Figure 7 summarizes the same diagnostics across datasets and constraint budgets.

The learned assignments retain clear information about both the generating source classes and their relative composition. Source-class mass remains above the uniform reference for every dataset and budget, and the stricter source top-2 hit rate likewise remains above its random-pair reference throughout. These two diagnostics consistently identify structure associated with the generating source pair, while the normalized composition estimate shows a clear dependence on the amount of supervision. Across the seven datasets, the correlation between $\hat{\lambda}$ and λ ranges from 0.570–0.841 at 1k constraints, increases to 0.643–0.866 at 5k, and reaches 0.705–0.876 at 10k, with all seven datasets improving monotonically across the three budgets. Thus, GDP not only preserves the source identities underlying mixed samples, but increasingly recovers the continuous membership proportions encoded by Mix supervision as additional pairwise information becomes available.

D.3 Human-derived graded supervision

Unlike the controlled Binary, Difficulty, and Mix constructions, CIFAR10H provides naturally occurring graded relations derived from human label distributions, spanning near-deterministic agreement and varying degrees of class ambiguity. We construct the corresponding pairwise grades from these distributions following the procedure in Appendix C.3. Table 16 reports the 5k results with and without autoencoder pretraining. The strongest methods remain closely grouped, while GDP

Table 12: Test ACC/NMI/ARI (%; five-run means) at 5k constraints. † denotes the random-initialization variant without autoencoder pretraining; VanillaDCC and VolMaxDCC are unchanged. Best/second-best are **bold**/underlined per dataset/source/metric; WPP† reduces to ProbPair† under Binary and is omitted.

Method	CIFAR100-20	CIFAR10	FMNIST	ImageNet10	MNIST	Reuters	STL10
Binary							
VanillaDCC	48.9 / 51.7 / 35.0	88.1 / 84.4 / 81.3	76.7 / 67.3 / 61.3	88.7 / 82.6 / 80.3	82.1 / 74.8 / 71.5	95.0 / 81.2 / 87.9	82.2 / 74.8 / 70.7
VolMaxDCC	44.6 / 47.8 / 29.5	89.1 / 84.5 / 82.1	68.6 / 62.9 / 54.4	77.5 / 71.6 / 66.7	81.1 / 73.1 / 71.0	93.4 / 76.7 / 84.3	78.9 / 71.4 / 62.6
CIDEC†	46.1 / 48.5 / 30.7	90.1 / 85.5 / 83.3	72.3 / 66.0 / 57.7	87.6 / 81.9 / 79.1	93.1 / 87.5 / 87.3	87.6 / 68.4 / 76.6	82.5 / 74.2 / 69.2
SpherePair†	66.9 / 63.6 / 51.9	94.7 / 89.0 / 89.0	**79.3** / 70.4 / 64.4	95.4 / 89.3 / 90.1	95.1 / 88.3 / 89.6	95.1 / 81.6 / 88.0	89.0 / 79.6 / 78.0
ProbPair†	62.8 / 61.4 / 47.6	94.5 / 88.5 / 88.5	74.2 / 66.7 / 59.0	95.8 / 90.2 / 90.9	89.8 / 81.1 / 80.6	93.8 / 78.3 / 86.0	88.3 / 78.8 / 77.0
WPP†	–	–	–	–	–	–	–
ECI-PP†	62.3 / 62.5 / 47.5	94.1 / 88.3 / 87.9	71.7 / 65.7 / 56.4	95.7 / 90.2 / 90.7	88.6 / 80.1 / 78.4	94.7 / 80.7 / 87.4	88.5 / 79.4 / 77.3
GDP† (Ours)	**71.3 / 67.0 / 56.2**	**94.9 / 89.3 / 89.3**	79.1 / **70.9 / 64.6**	**95.9 / 90.4 / 91.1**	**95.8 / 89.6 / 90.9**	**95.2 / 81.9 / 88.2**	**89.5 / 80.7 / 79.0**
Difficulty							
VanillaDCC	17.4 / 15.3 / 4.1	74.3 / 70.2 / 63.3	47.7 / 43.2 / 27.1	40.7 / 34.1 / 23.6	57.3 / 51.6 / 41.3	88.3 / 68.4 / 76.2	52.8 / 52.2 / 32.6
VolMaxDCC	45.4 / 49.0 / 30.6	87.1 / 83.4 / 80.9	70.2 / 62.1 / 54.2	74.3 / 68.3 / 62.4	75.6 / 69.6 / 64.4	91.9 / 75.1 / 82.5	83.5 / 71.2 / 68.7
CIDEC†	15.2 / 15.5 / 2.2	79.4 / 76.7 / 70.4	40.3 / 36.8 / 19.0	64.8 / 60.0 / 40.9	56.5 / 59.6 / 42.2	71.9 / 49.4 / 53.1	49.0 / 48.8 / 28.0
SpherePair†	63.0 / 60.6 / **47.5**	94.5 / 88.4 / 88.5	73.3 / 65.2 / 58.4	89.5 / 80.6 / 79.4	91.3 / 83.0 / 82.9	**93.5 / 78.3** / 85.3	87.0 / 76.7 / 74.8
ProbPair†	51.5 / 50.2 / 32.9	93.4 / 87.0 / 86.4	72.1 / 64.5 / 54.3	88.9 / 81.6 / 78.7	86.1 / 75.5 / 73.6	86.7 / 68.5 / 73.5	85.3 / 77.3 / 72.3
WPP†	54.0 / 53.4 / 36.1	94.0 / 87.6 / 87.5	70.7 / 64.5 / 54.8	91.9 / 83.9 / 83.0	88.5 / 77.9 / 77.1	86.2 / 68.8 / 72.9	84.2 / 74.9 / 70.1
ECI-PP†	61.7 / **61.1** / 44.8	93.8 / 88.0 / 87.4	72.7 / 67.1 / 57.7	94.1 / 87.7 / 87.4	89.3 / 80.0 / 79.0	91.3 / 75.8 / 81.8	86.7 / 78.0 / 74.0
GDP† (Ours)	**63.2** / 60.8 / 47.4	**95.0 / 89.3 / 89.5**	**81.2 / 71.6 / 65.9**	**95.7 / 90.2 / 90.6**	**94.7 / 87.4 / 88.7**	**93.5** / 78.0 / **85.5**	**89.1 / 80.2 / 78.0**
Mix							
VanillaDCC	48.3 / 49.4 / 34.1	90.7 / 85.6 / 83.5	72.8 / 64.4 / 57.7	84.9 / 80.1 / 76.6	85.0 / 73.8 / 72.5	92.2 / 75.7 / 83.6	84.2 / 75.3 / 72.2
VolMaxDCC	42.0 / 46.6 / 27.3	86.2 / 81.9 / 77.5	67.0 / 61.4 / 52.6	68.4 / 62.9 / 55.4	68.3 / 61.2 / 54.6	93.1 / 76.8 / 84.3	71.9 / 63.7 / 52.0
CIDEC†	42.9 / 45.8 / 28.3	83.2 / 77.6 / 73.1	66.4 / 58.9 / 49.3	78.8 / 77.6 / 71.0	76.0 / 71.6 / 65.6	71.0 / 43.7 / 49.6	73.3 / 70.4 / 62.3
SpherePair†	54.4 / 54.9 / 39.5	88.7 / 83.4 / 80.9	67.1 / 61.8 / 53.0	86.1 / 75.9 / 73.8	78.4 / 71.0 / 66.5	93.6 / 77.7 / 85.1	86.1 / 76.0 / 73.3
ProbPair†	56.8 / 54.7 / 41.2	93.9 / 87.9 / 87.5	68.1 / 62.4 / 53.4	93.8 / 86.3 / 86.8	85.1 / 73.8 / 72.0	87.3 / 69.9 / 77.0	87.3 / 77.9 / 75.4
WPP†	62.1 / 57.0 / 44.2	93.7 / 87.8 / 87.3	70.2 / 63.6 / 54.8	93.5 / 85.9 / 86.3	86.9 / 76.7 / 75.5	86.0 / 67.9 / 74.5	87.5 / 77.7 / 75.5
ECI-PP†	60.7 / **60.4** / 45.6	93.6 / 87.7 / 87.0	69.8 / 65.3 / 55.5	94.8 / 88.5 / 88.9	88.3 / 78.4 / 77.0	93.9 / 79.0 / 85.6	86.9 / 77.9 / 75.1
GDP† (Ours)	**65.0** / 59.1 / **46.7**	**94.9 / 89.2 / 89.4**	**77.2 / 69.3 / 62.8**	**95.9 / 90.5 / 91.2**	**93.3 / 84.8 / 86.0**	**94.8 / 81.1 / 87.8**	**89.5 / 80.6 / 79.0**

attains the highest test mean on ACC, NMI, and ARI in both initialization regimes. With pretraining, GDP achieves 94.2%, 88.6%, and 87.9%, respectively; from random initialization, GDP† reaches 94.0%, 88.2%, and 87.5%. These results extend the evaluation beyond controlled grade constructions to supervision derived from actual human labeling behavior.

D.4 Robustness to numerical grade scales

The recalibration experiments examine sensitivity to the numerical grade scale while preserving the underlying relational ordering. Following the protocol in Appendix C.4, Difficulty varies the boundary $b \in \{0.3, 0.5, 0.7\}$, while Mix uses the identity scale together with two endpoint-preserving monotone recalibrations, $\tau = 0.2$ and $\zeta = +3$. The main text summarizes these results through ACC in Figure 3; Figures 8–10 present the corresponding ACC, NMI, and ARI results under the same pretrained 10k protocol.

Two complementary stress tests are considered.

(i) **Difficulty.** This is a controlled, order-separable recalibration: varying b changes the numerical placement of the grades without changing the underlying same-class/different-class semantics or their ordering. This controlled construction is distinct from human-derived graded supervision, which is evaluated separately on CIFAR10H in Appendix D.3. Multi-class co-membership is naturally sparse: on balanced MNIST, for example, two randomly sampled examples belong to the same class with probability $1/10$, giving an expected binary co-membership of 0.1. Larger values of b therefore place the graded supervision progressively farther from this low-co-membership geometry while preserving the same relational order. GDP maintains both high performance and a stable response across this sweep. On FMNIST, its ACC remains within 83.6–84.0% across $b \in \{0.3, 0.5, 0.7\}$, whereas VanillaDCC and CIDEC decrease markedly as b increases. The corresponding NMI and ARI results exhibit the same overall pattern.

Table 13: Test ACC/NMI/ARI (%; five-run means) at 10k constraints. † denotes the random-initialization variant without autoencoder pretraining; VanillaDCC and VolMaxDCC are unchanged. Best/second-best are **bold**/underlined per dataset/source/metric; WPP† reduces to ProbPair† under Binary and is omitted.

Method	CIFAR100-20	CIFAR10	FMNIST	ImageNet10	MNIST	Reuters	STL10
Binary							
VanillaDCC	63.8 / 61.6 / 49.0	94.7 / 88.6 / 88.9	77.5 / 69.0 / 62.5	95.8 / 90.3 / 91.0	94.1 / 86.0 / 87.6	94.2 / 79.1 / 86.0	89.6 / 80.3 / 79.0
VolMaxDCC	45.1 / 49.2 / 31.8	94.5 / 88.4 / 88.5	75.4 / 67.6 / 61.1	95.4 / 89.5 / 90.2	84.1 / 78.0 / 75.6	94.8 / 80.8 / 87.6	84.4 / 76.9 / 71.6
CIDEC†	55.8 / 56.0 / 40.9	94.6 / 88.8 / 88.7	76.5 / 69.2 / 62.2	93.6 / 87.8 / 87.4	96.0 / 90.2 / 91.4	84.7 / 66.6 / 74.9	86.9 / 78.5 / 75.9
SpherePair†	74.8 / 68.4 / 60.0	95.0 / 89.3 / 89.5	82.4 / 73.1 / **68.3**	95.9 / 90.5 / 91.2	96.3 / 90.5 / 92.0	**96.2 / 85.0 / 90.6**	89.7 / 80.6 / 79.3
ProbPair†	74.6 / 67.8 / 57.9	94.9 / 89.0 / 89.3	80.7 / 71.1 / 65.8	96.2 / 91.1 / 91.8	95.0 / 88.1 / 89.4	95.6 / 83.2 / 89.2	**90.0 / 81.2** / 79.8
WPP†	–	–	–	–	–	–	–
ECI-PP†	65.4 / 65.5 / 52.7	**95.1 / 89.5 / 89.8**	77.8 / 69.6 / 62.4	**96.5 / 91.6 / 92.4**	94.9 / 88.1 / 89.1	95.8 / 83.9 / 89.7	**90.0 / 81.2** / 79.8
GDP† (Ours)	**75.5 / 69.8 / 61.1**	**95.1 / 89.5** / 89.7	**82.5 / 73.5 / 68.3**	96.3 / 91.2 / 91.9	**96.5 / 91.1 / 92.6**	95.5 / 83.0 / 89.0	**90.0** / 81.1 / **79.9**
Difficulty							
VanillaDCC	20.5 / 19.1 / 4.7	61.3 / 53.7 / 44.8	49.4 / 44.6 / 28.5	52.9 / 46.0 / 32.1	65.3 / 59.8 / 50.3	84.1 / 59.3 / 66.9	58.6 / 54.1 / 35.3
VolMaxDCC	45.8 / 48.9 / 30.0	91.2 / 86.5 / 85.3	74.2 / 66.9 / 59.6	77.3 / 72.3 / 65.6	88.5 / 82.1 / 81.3	92.3 / 75.8 / 83.3	84.7 / 73.5 / 71.6
CIDEC†	10.8 / 8.0 / 1.1	85.5 / 81.0 / 78.1	41.9 / 39.3 / 21.7	50.0 / 50.1 / 35.6	59.1 / 61.3 / 44.5	82.5 / 59.5 / 65.0	42.1 / 45.5 / 28.4
SpherePair†	69.5 / 63.8 / 53.6	94.8 / 88.6 / 89.0	80.1 / 69.6 / 65.3	91.6 / 82.6 / 82.9	94.1 / **87.0** / 88.0	**93.4 / 78.8 / 85.4**	87.6 / 77.2 / 75.8
ProbPair†	60.6 / 57.4 / 40.9	94.3 / 88.2 / 87.9	76.3 / 68.6 / 59.8	94.8 / 88.4 / 88.7	91.3 / 82.3 / 81.6	86.3 / 69.5 / 72.8	88.2 / 78.8 / 75.9
WPP†	65.2 / 60.8 / 45.1	94.5 / 88.3 / 88.3	79.7 / 69.4 / 63.0	94.0 / 87.2 / 87.1	92.9 / 84.3 / 85.1	86.8 / 70.8 / 73.9	87.2 / 77.4 / 74.4
ECI-PP†	67.1 / 64.4 / 49.6	94.8 / 88.9 / 89.0	77.7 / 70.0 / 62.7	**95.9 / 90.6 / 91.2**	92.9 / 84.8 / 85.4	92.0 / 77.3 / 83.2	86.3 / 78.9 / 74.1
GDP† (Ours)	**76.2 / 70.0 / 61.0**	**95.2 / 89.4 / 89.8**	**82.8 / 72.7 / 68.0**	95.6 / 89.9 / 90.4	**94.5** / 86.9 / **88.4**	93.2 / 77.5 / 84.9	**89.4 / 80.4 / 78.6**
Mix							
VanillaDCC	60.7 / 58.6 / 46.5	94.6 / 88.7 / 88.8	79.2 / 69.8 / 64.7	95.6 / 89.8 / 90.5	92.8 / 83.4 / 84.9	90.8 / 71.2 / 79.2	89.7 / 80.3 / 79.2
VolMaxDCC	47.8 / 50.5 / 33.1	85.7 / 82.8 / 78.9	72.7 / 65.8 / 58.8	83.2 / 78.4 / 74.9	86.8 / 79.0 / 77.9	95.2 / 82.2 / 88.5	83.2 / 75.0 / 68.4
CIDEC†	49.8 / 51.1 / 35.4	84.9 / 80.3 / 76.8	76.0 / 67.3 / 60.7	86.3 / 83.2 / 79.4	89.2 / 82.7 / 81.5	80.3 / 57.4 / 63.3	80.6 / 74.6 / 69.1
SpherePair†	60.0 / 57.5 / 44.4	92.4 / 86.5 / 85.4	68.7 / 63.0 / 54.9	91.9 / 83.9 / 83.6	85.6 / 78.2 / 75.7	95.1 / 81.9 / 88.4	86.4 / 76.8 / 74.7
ProbPair†	71.9 / 64.7 / 54.9	94.7 / 89.0 / 89.0	78.0 / 68.7 / 62.7	96.1 / 90.7 / 91.5	92.3 / 83.1 / 84.0	94.8 / 81.3 / 87.7	89.5 / 80.4 / 79.0
WPP†	73.7 / 65.4 / 56.3	94.8 / 88.9 / 89.1	79.4 / 70.1 / 64.3	96.4 / 91.5 / 92.2	93.4 / 85.0 / 86.1	95.2 / 82.3 / 88.6	89.8 / 80.6 / 79.5
ECI-PP†	66.4 / 64.4 / 52.1	94.8 / 89.1 / 89.1	73.8 / 67.8 / 59.5	96.3 / 91.3 / 92.0	91.5 / 82.9 / 82.8	95.1 / 82.1 / 88.3	89.3 / 80.4 / 78.7
GDP† (Ours)	**73.9 / 65.9 / 56.8**	**95.2 / 89.6 / 89.9**	**81.6 / 72.2 / 67.4**	**96.6 / 91.8 / 92.5**	**95.1 / 88.1 / 89.5**	**95.5 / 83.2 / 89.2**	**90.3 / 81.7 / 80.4**

Table 14: K-means clustering on the benchmark input representations. Test ACC/NMI/ARI (%; five-run mean±std).

Dataset	K	ACC	NMI	ARI
CIFAR100-20	20	50.6±2.5	56.1±1.2	35.7±1.6
CIFAR10	10	78.7±4.3	78.9±1.9	69.6±2.7
FMNIST	10	51.5±4.8	51.6±1.4	35.6±3.1
ImageNet10	10	80.3±10.1	82.3±5.6	72.0±13.7
MNIST	10	53.5±1.6	49.8±0.7	36.9±0.4
Reuters	4	61.6±9.2	40.3±8.0	33.8±12.7
STL10	10	78.7±6.8	73.1±2.6	64.3±6.6

(ii) **Mix.** Mix begins from a grade that directly reflects overlap between the two soft membership vectors. Its two recalibrations deliberately stress different aspects of this numerical representation: $\tau = 0.2$ moves intermediate grades toward 0.5, creating a more ambiguous relation regime, whereas $\zeta = +3$ assigns substantially larger numerical values to the same ordered relations, deliberately moving them away from the low co-membership geometry of a multiclass partition. GDP remains stable under both Mix recalibrations; on FMNIST, its ACC stays within 81.4–83.3%, with NMI and ARI showing the same overall behavior.

Taken together, the Difficulty and Mix sweeps show that strong clustering performance does not depend on one particular numerical realization of the graded relations. This behavior is consistent with the role of the Global ranking objective in GDP: the Global branch organizes relations primarily through their relative order rather than forcing the learned co-assignment scores to reproduce the raw grade values. The results therefore provide empirical support for the intended reduction in

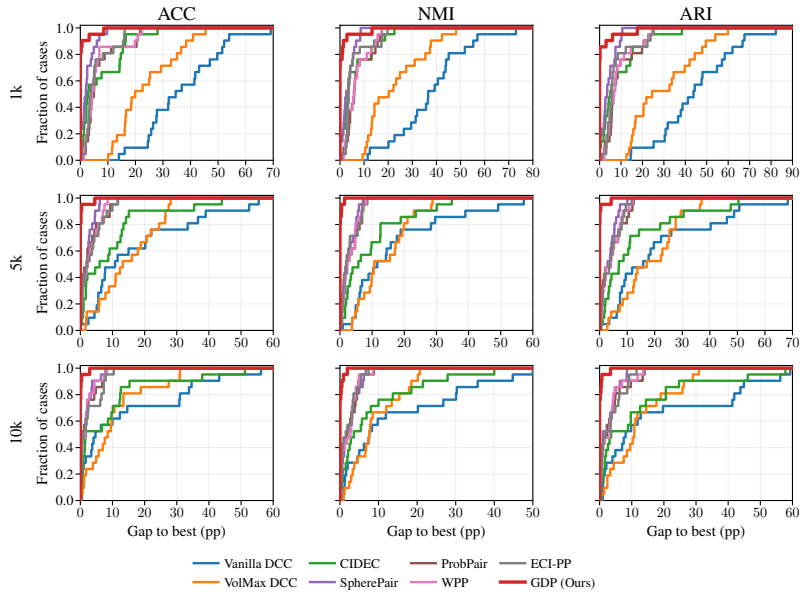

Figure 6: Cumulative gap to the best method over the 21 dataset–supervision cases for each metric and constraint budget. Each curve reports the fraction of cases within a given performance gap, so a sharper concentration near zero indicates more consistently near-best performance.

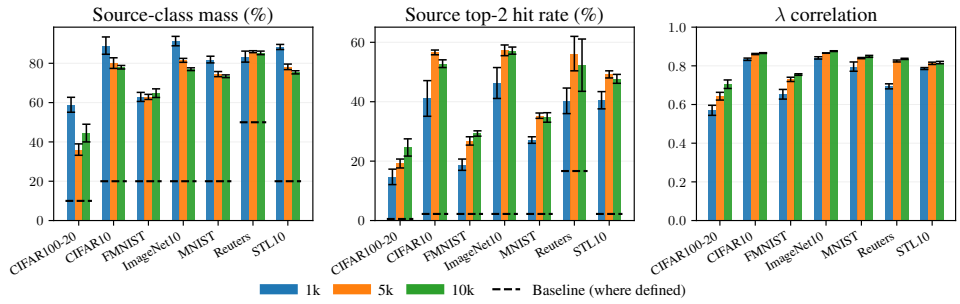

Figure 7: Structured mixed-membership recovery across the seven flat benchmarks and three constraint budgets. Bars show five-run means with standard-deviation error bars; dashed lines denote reference baselines where defined.

dependence on exact grade calibration, while the remaining value-sensitive components can still exploit the graded information available in the supervision.

D.5 Multi-resolution hierarchy recovery

Beyond the aggregate hierarchy comparison in Table 2, we examine multi-resolution recovery through levelwise partition quality, joint versus separate optimization across resolutions, transition-level coherence, an input-representation reference, agglomerative hierarchy post-processing, and robustness to hierarchy-grade spacing.

Complete levelwise recovery. Table 17 expands the partition-quality view to ACC, NMI, and ARI at every prescribed resolution. The levelwise results show that the advantage is not confined to the aggregate metrics. On T200 and T608, GDP-Hier attains the highest ACC, NMI, and ARI at every reported resolution among the constrained methods in this comparison, showing that its gains are distributed across the hierarchy rather than concentrated at the finest level. The same levelwise advantage holds on RCV1-HIER. On CIFAR100-HIER, GDP-Hier leads all three leaf

Table 15: Structured mixed-membership recovery on held-out mixed samples (five-run mean±std). Reference baselines are shown where defined.

Dataset	Metric	Baseline	1k	5k	10k
CIFAR100-20	Source-class mass (%)	10.0	58.9 ± 3.8	36.1 ± 2.9	44.5 ± 4.5
	Source top-2 hit rate (%)	0.5	14.7 ± 2.6	19.2 ± 1.5	24.6 ± 2.9
	λ correlation	–	0.570 ± 0.026	0.643 ± 0.020	0.705 ± 0.022
CIFAR10	Source-class mass (%)	20.0	89.0 ± 4.4	80.1 ± 2.7	78.0 ± 0.9
	Source top-2 hit rate (%)	2.2	41.1 ± 6.0	56.6 ± 0.8	52.8 ± 1.3
	λ correlation	–	0.834 ± 0.006	0.861 ± 0.003	0.866 ± 0.002
FMNIST	Source-class mass (%)	20.0	62.9 ± 2.3	62.9 ± 1.3	64.8 ± 2.2
	Source top-2 hit rate (%)	2.2	18.8 ± 1.9	26.8 ± 1.4	29.3 ± 0.9
	λ correlation	–	0.653 ± 0.025	0.730 ± 0.011	0.755 ± 0.004
ImageNet10	Source-class mass (%)	20.0	91.3 ± 2.4	81.5 ± 1.0	77.0 ± 0.7
	Source top-2 hit rate (%)	2.2	46.3 ± 5.2	57.3 ± 1.8	57.2 ± 1.2
	λ correlation	–	0.841 ± 0.006	0.866 ± 0.001	0.876 ± 0.002
MNIST	Source-class mass (%)	20.0	81.9 ± 1.7	74.5 ± 1.3	73.4 ± 0.7
	Source top-2 hit rate (%)	2.2	27.1 ± 1.1	35.3 ± 0.9	34.7 ± 1.6
	λ correlation	–	0.796 ± 0.024	0.840 ± 0.003	0.849 ± 0.005
Reuters	Source-class mass (%)	50.0	83.4 ± 2.8	85.9 ± 0.5	85.3 ± 0.9
	Source top-2 hit rate (%)	16.7	40.3 ± 4.3	56.2 ± 5.8	52.3 ± 8.8
	λ correlation	–	0.694 ± 0.013	0.825 ± 0.005	0.836 ± 0.003
STL10	Source-class mass (%)	20.0	88.3 ± 1.3	78.2 ± 1.4	75.4 ± 0.8
	Source top-2 hit rate (%)	2.2	40.5 ± 2.9	49.2 ± 1.2	47.7 ± 1.5
	λ correlation	–	0.786 ± 0.005	0.813 ± 0.006	0.817 ± 0.006

Table 16: CIFAR10H results with human-derived graded supervision at 5k constraints (five-run mean±std, %). Blue/black denote training/test results; best/second-best test means within each block are **bold**/underlined. VanillaDCC and VolMaxDCC are unchanged across blocks.

Setting	Metric	Vanilla DCC	VolMax DCC	CIDEC	SpherePair	ProbPair	WPP	ECI-PP	**GDP**
Pretrained	ACC	89.0±4.7 87.9 ± 4.5	91.1±8.2 90.1 ± 7.6	95.3±0.2 94.1 ± 0.4	93.8±0.4 93.0 ± 0.4	95.2±0.2 94.0 ± 0.3	95.0±0.1 94.0 ± 0.4	94.6±0.3 93.8 ± 0.3	95.7±0.1 **94.2** ± 0.2
	NMI	85.1±3.7 84.1 ± 3.1	86.5±6.0 85.1 ± 5.5	90.1±0.3 88.2 ± 0.3	87.6±0.6 86.4 ± 0.4	89.9±0.1 88.3 ± 0.4	89.6±0.2 88.1 ± 0.5	89.2±0.3 88.1 ± 0.4	90.7±0.2 **88.6** ± 0.3
	ARI	82.5±5.4 80.5 ± 4.9	84.9±9.4 82.8 ± 8.5	90.2±0.4 87.6 ± 0.6	87.2±0.7 85.5 ± 0.7	89.9±0.2 87.5 ± 0.5	89.7±0.2 87.5 ± 0.7	88.8±0.5 87.1 ± 0.6	90.9±0.2 **87.9** ± 0.4
No pretrain	ACC	89.0±4.7 87.9 ± 4.5	91.1±8.2 90.1 ± 7.6	83.2±7.9 82.5 ± 7.1	93.3±1.5 93.0 ± 1.3	94.4±0.8 93.4 ± 1.0	94.5±0.3 93.7 ± 0.4	94.1±0.4 93.5 ± 0.5	95.4±0.2 **94.0** ± 0.5
	NMI	85.1±3.7 84.1 ± 3.1	86.5±6.0 85.1 ± 5.5	81.7±5.6 80.4 ± 5.0	86.8±2.0 86.6 ± 1.9	88.9±0.6 87.4 ± 0.7	88.9±0.6 87.8 ± 0.3	88.5±0.7 87.6 ± 0.6	90.3±0.3 **88.2** ± 0.4
	ARI	82.5±5.4 80.5 ± 4.9	84.9±9.4 82.8 ± 8.5	75.9±8.4 73.8 ± 7.7	86.4±2.5 85.7 ± 2.4	88.6±1.3 86.4 ± 1.5	88.7±0.6 86.9 ± 0.6	88.0±0.9 86.6 ± 0.8	90.4±0.4 **87.5** ± 0.9

metrics together with the lowest HVR; on Aircraft, it matches the best leaf ACC while reducing HVR to 11.8%, compared with 19.4% for the next-best method.

Joint versus separate recovery. The comparison with FlatGDP directly tests whether optimizing the prescribed resolutions separately is sufficient. FlatGDP trains one GDP model per resolution, whereas GDP-Hier learns all resolutions jointly through the composite hierarchy relation. The distinction is especially clear on the deeper hierarchies. On T200, joint recovery raises leaf ACC from 37.6% to 52.9% and reduces HVR from 42.9% to 4.8%. On T608, mean ACC increases from 17.7% to 29.1% while HVR falls from 73.7% to 6.5%. Strong single-resolution partitions therefore

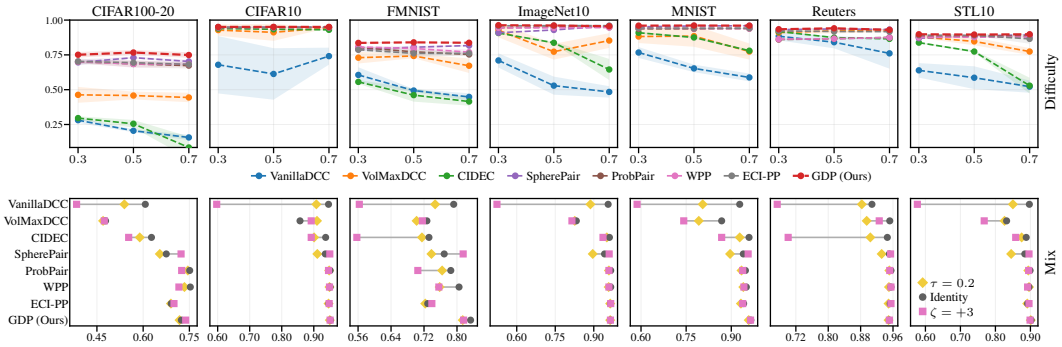

Figure 8: Test ACC under targeted grade recalibration at 10k constraints. Difficulty varies $b \in \{0.3, 0.5, 0.7\}$; Mix uses $\tau = 0.2$, the identity mapping, and $\zeta = +3$.

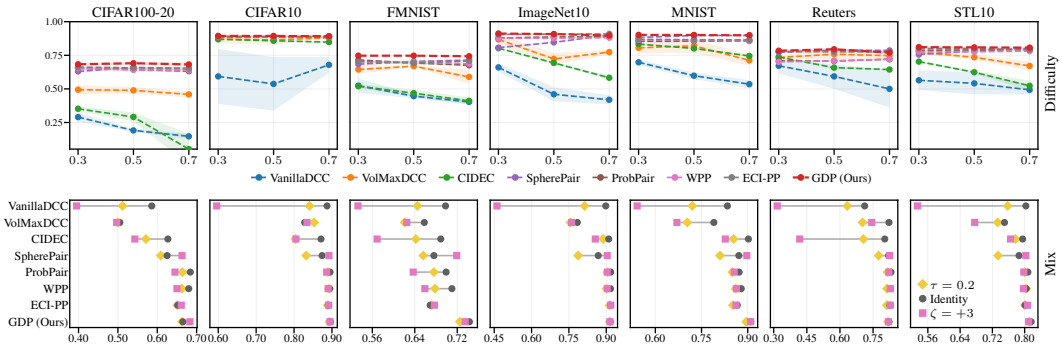

Figure 9: Test NMI under targeted grade recalibration at 10k constraints. Difficulty varies $b \in \{0.3, 0.5, 0.7\}$; Mix uses $\tau = 0.2$, the identity mapping, and $\zeta = +3$.

do not by themselves produce a coherent hierarchy; the joint formulation coordinates their recovery through the composite hierarchy relation.

The same joint formulation also avoids repeating the full end-to-end optimization at every prescribed resolution. VanillaDCC, CIDEC, and FlatGDP require a separate training run for each target K, whereas GDP-Hier optimizes all resolution-specific heads jointly and directly returns the prescribed partitions from a single trained model. This avoids resolution-wise retraining; relative to FlatGDP, the resulting computational benefit becomes particularly pronounced on deeper hierarchies, as quantified in Appendix E.

Cross-level coherence. Table 18 resolves the aggregate HVR into transition-level diagnostics of the recovered parent–child structure. For a transition from level ℓ to the finer level $\ell + 1$, let $\mathcal{I}$ index the nonempty child clusters, let n denote the number of evaluated samples, and define

$$m_k = \max_h |S_k^{(\ell+1)} \cap S_h^{(\ell)}|, \qquad k \in \mathcal{I},$$

as the number of samples in child cluster k that fall into its majority parent. Child-to-parent purity and transition violation are then

$$\text{Purity}_\ell = \frac{1}{|\mathcal{I}|} \sum_{k \in \mathcal{I}} \frac{m_k}{|S_k^{(\ell+1)}|}, \qquad v_\ell = 1 - \frac{1}{n} \sum_{k \in \mathcal{I}} m_k.$$

Purity measures how consistently nonempty child clusters map to their dominant parents, while v_ℓ measures the corresponding sample-level inconsistency; HVR averages v_ℓ across adjacent transitions, as defined in Appendix C.4. The nonempty-child ratio further records how much of the prescribed finer partition is actually utilized. This becomes increasingly informative as the hierarchy grows finer: Table 4 shows that the available supervision becomes strongly concentrated on low hierarchy grades, most notably on T608, where 96.4% of pairs have grade 0 and only 0.2% have grade 1.

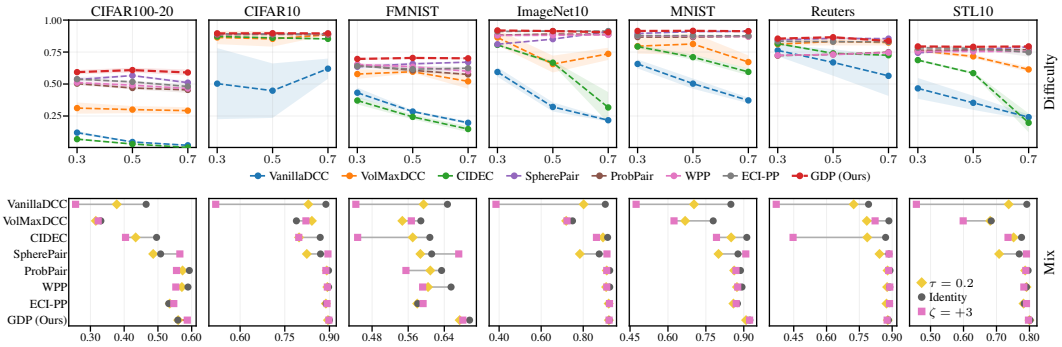

Figure 10: Test ARI under targeted grade recalibration at 10k constraints. Difficulty varies $b \in \{0.3, 0.5, 0.7\}$; Mix uses $\tau = 0.2$, the identity mapping, and $\zeta = +3$.

The transition results show that GDP-Hier preserves both cross-level consistency and broad cluster utilization under this increasingly sparse fine-grained supervision. On the shallower hierarchies, the recovered transitions are already close to clean parent–child mappings. On T200, child-cluster utilization remains near complete across all three transitions while purity stays high throughout. T608 presents the most demanding case: despite its 608-way finest partition and highly sparse direct leaf-level evidence, GDP-Hier keeps 76.9% of the leaf clusters nonempty while maintaining an overall HVR of 6.5%. These diagnostics therefore complement the aggregate HVR by showing that coherent hierarchy recovery is accompanied by substantial utilization of the prescribed finer partitions rather than being achieved through a heavily collapsed solution.

Input-representation reference. To distinguish hierarchy recovery from structure already present in the input representations, we further compare GDP-Hier with K-means applied independently at each prescribed resolution. Table 19 shows that GDP-Hier achieves higher mean ACC and lower HVR on all five hierarchy benchmarks. On T200, mean ACC increases from 38.2% to 47.3% while HVR decreases from 21.8% to 4.8%; on T608, mean ACC increases from 27.7% to 29.1% while HVR decreases from 26.7% to 6.5%. At the T608 leaf level, however, K-means on the fixed SwAV input features attains 34.1% ACC versus 24.7% for GDP-Hier, showing that the input representation already contains strong fine-level separability. Relative to independent K-means, GDP-Hier achieves higher mean ACC across resolutions and substantially lower HVR, indicating that its advantage lies in organizing the available structure into more compatible partitions across the prescribed hierarchy resolutions rather than simply maximizing the finest-level clustering quality.

Agglomerative hierarchy post-processing. To test whether generic hierarchy-enforcing post-processing can account for the observed hierarchy recovery, we conduct a separate five-run analysis on the input representations and on SpherePair, ProbPair, and WPP embeddings trained with 5k hierarchy-induced constraints. Finest-level training K-means centroids are ℓ_2-normalized and clustered using average linkage with cosine distance; cuts of the resulting dendrogram provide the prescribed coarser partitions, which test samples inherit through their finest-level clusters. Table 20 reports the levelwise results. Because post-processing starts from the same finest-level K-means partition, the leaf result is unchanged; HVR is zero by construction rather than evidence of hierarchy learned by the representation.

Against the existing GDP-Hier results in Table 2, the post-processed references do not provide a uniformly competitive alternative. On T200, the strongest post-processed reference reaches only 41.0% mean ACC, compared with 47.3% for GDP-Hier, with similarly clear gaps in mean NMI and ARI. On T608, the fixed input representation attains higher mean ACC after agglomeration (31.2% versus 29.1%), but GDP-Hier retains higher mean NMI (59.3% versus 57.1%) and nearly identical mean ARI, so the ACC gain does not translate into uniformly better agreement with the semantic hierarchy. On some settings such as Aircraft, post-processing can further expose useful coarse structure from the learned representation. Overall, however, these gains are dataset-dependent, whereas GDP-Hier provides consistently strong multi-resolution recovery by learning the prescribed semantic partitions jointly rather than relying on hierarchy construction after representation learning.

Table 17: Levelwise hierarchy recovery with 5k constraints (five-run means, %). Each resolution reports ACC/NMI/ARI; Mean averages the levelwise metrics, and lower HVR is better. Best/second-best values are **bold**/underlined.

CIFAR100-HIER

Method	HVR	fine (leaf)	super	Mean
VanillaDCC	48.4	21.3 / 51.3 / 16.1	51.7 / 50.8 / 35.0	36.5 / 51.0 / 25.6
CIDEC	22.6	44.2 / 65.1 / 29.8	58.2 / 61.3 / 44.4	51.2 / 63.2 / 37.1
SpherePair	7.4	52.4 / 69.5 / 40.0	61.7 / 63.4 / 45.5	57.0 / 66.5 / 42.8
ProbPair	10.3	50.9 / 68.4 / 38.8	**71.6** / **66.7** / **56.1**	61.2 / 67.6 / **47.5**
WPP	10.9	52.1 / 68.9 / 39.3	71.1 / 66.6 / 55.3	**61.6** / **67.8** / 47.3
FlatGDP	20.7	54.9 / 68.9 / 40.8	61.1 / 59.8 / 45.5	58.0 / 64.3 / 43.1
GDP-Hier	**3.8**	**56.1** / **69.9** / **41.2**	64.4 / 63.4 / 49.9	60.3 / 66.7 / 45.5

RCV1-HIER

Method	HVR	fine (leaf)	super	Mean
VanillaDCC	19.8	20.6 / 41.8 / 19.1	82.7 / 58.2 / 62.0	51.7 / 50.0 / 40.5
CIDEC	16.4	21.3 / 46.9 / 22.1	87.1 / 69.8 / 73.4	54.2 / 58.3 / 47.8
SpherePair	1.7	65.8 / 67.3 / 52.1	94.1 / 80.3 / 85.1	79.9 / 73.8 / 68.6
ProbPair	**1.2**	67.4 / 66.1 / 52.2	94.4 / 80.7 / 85.6	80.9 / 73.4 / 68.9
WPP	**1.2**	68.5 / 67.4 / 52.7	94.5 / 81.2 / 86.1	81.5 / 74.3 / 69.4
FlatGDP	10.3	36.2 / 47.7 / 25.3	94.2 / 80.4 / 85.3	65.2 / 64.1 / 55.3
GDP-Hier	2.3	**72.0** / **69.7** / **57.3**	**94.8** / **82.2** / **86.8**	**83.4** / **76.0** / **72.0**

Aircraft

Method	HVR	variant (leaf)	family	manufacturer	Mean
VanillaDCC	63.0	11.1 / 34.0 / 3.9	15.4 / 32.8 / **7.6**	20.5 / 27.7 / 11.3	15.7 / 31.5 / 7.6
CIDEC	52.9	11.3 / 34.3 / 2.5	15.6 / 32.6 / 6.9	**24.1** / 26.4 / **19.5**	**17.0** / 31.1 / **9.6**
SpherePair	19.4	13.8 / 39.4 / 6.0	13.7 / 33.4 / 5.1	15.2 / 22.7 / 4.7	14.2 / 31.8 / 5.3
ProbPair	30.9	**15.0** / **41.0** / 5.6	**16.0** / **36.7** / 6.4	18.9 / **29.6** / 8.9	16.6 / **35.8** / 7.0
WPP	32.6	14.8 / 40.8 / 5.3	15.7 / **36.7** / 6.2	18.2 / 29.4 / 8.5	16.2 / 35.6 / 6.7
FlatGDP	59.3	14.3 / 38.9 / 4.7	15.0 / 33.5 / 4.8	17.4 / 25.2 / 6.0	15.6 / 32.5 / 5.2
GDP-Hier	**11.8**	**15.0** / 40.6 / **6.1**	15.7 / 35.5 / 6.0	17.8 / 26.1 / 6.6	16.2 / 34.1 / 6.2

T200

Method	HVR	leaf	parent	ancestor2	ancestor3	Mean
VanillaDCC	65.3	11.8 / 45.1 / 6.4	19.8 / 43.2 / 12.1	24.1 / 40.5 / 17.2	27.1 / 38.4 / 17.4	20.7 / 41.8 / 13.3
CIDEC	56.6	30.4 / 63.0 / 16.3	33.7 / 60.3 / 20.7	34.1 / 57.0 / 21.7	32.4 / 52.1 / 20.0	32.6 / 58.1 / 19.7
SpherePair	15.7	33.5 / 66.9 / 21.8	35.6 / 63.7 / 26.5	33.4 / 58.4 / 25.2	32.6 / 52.9 / 21.7	33.8 / 60.5 / 23.8
ProbPair	23.9	30.8 / 63.5 / 18.5	33.4 / 59.7 / 23.8	32.2 / 54.0 / 24.3	30.7 / 48.8 / 20.6	31.8 / 56.5 / 21.8
WPP	24.3	32.4 / 64.2 / 19.5	34.0 / 60.2 / 24.3	32.4 / 54.6 / 23.4	30.5 / 49.4 / 20.3	32.3 / 57.1 / 21.9
FlatGDP	42.9	37.6 / 67.0 / 26.5	47.2 / 71.5 / 35.0	42.4 / 66.3 / 30.0	38.1 / 58.3 / 23.3	41.3 / 65.8 / 28.7
GDP-Hier	**4.8**	**52.9** / **77.3** / **40.5**	**50.7** / **75.0** / **40.7**	**45.9** / **68.9** / **34.7**	**39.5** / **61.3** / **26.6**	**47.3** / **70.6** / **35.6**

T608

Method	HVR	leaf	parent	ancestor2	ancestor3	Mean
VanillaDCC	70.9	5.2 / 39.7 / 2.0	12.1 / 32.1 / 6.6	16.8 / 30.4 / 10.9	21.8 / 29.5 / 12.5	14.0 / 32.9 / 8.0
CIDEC	64.4	17.2 / 60.1 / 7.4	21.7 / 51.7 / 12.6	24.4 / 49.1 / 15.3	25.7 / 45.1 / 15.8	22.2 / 51.5 / 12.8
SpherePair	23.0	17.6 / 60.7 / 7.9	19.0 / 50.1 / 12.5	22.0 / 45.7 / 15.6	24.3 / 41.8 / 14.7	20.7 / 49.6 / 12.7
ProbPair	33.2	18.7 / 60.9 / 7.9	19.2 / 49.2 / 11.3	21.6 / 44.4 / 14.5	22.9 / 40.7 / 14.0	20.6 / 48.8 / 11.9
WPP	33.2	19.2 / 61.3 / 8.2	19.5 / 49.9 / 11.8	22.2 / 45.3 / 15.1	24.5 / 41.9 / 15.4	21.4 / 49.6 / 12.6
FlatGDP	73.7	6.1 / 39.6 / 1.5	15.4 / 38.2 / 7.9	29.9 / 56.4 / 20.0	19.3 / 27.2 / 8.1	17.7 / 40.4 / 9.4
GDP-Hier	**6.5**	**24.7** / **67.1** / **15.8**	**30.2** / **61.6** / **20.3**	**31.5** / **57.2** / **22.6**	**30.0** / **51.3** / **18.1**	**29.1** / **59.3** / **19.2**

Robustness to hierarchy-grade spacing. Finally, we examine whether hierarchy recovery depends on the precise numerical spacing of the hierarchy grades. Figure 11 varies the intermediate breakpoint in the two-level setting: the default $\beta_1 = 0.5$ gives equal grade increments and $\Delta^{\beta}_{\min} = 0.5$, whereas moving the breakpoint to 0.3 or 0.7 reduces $\Delta^{\beta}_{\min}$ to 0.3 while preserving the same semantic hierarchy and relation ordering. GDP-Hier remains stable under this change. On CIFAR100-HIER, leaf ACC stays within 56.0–57.7% and HVR within 3.4–4.5%; on RCV1-HIER, leaf ACC remains within 72.5–75.2% while HVR stays below 2.5%. Since $\Delta^{\beta}_{\min}$ enters directly into the sufficient recovery condition of Theorem 5.4, the stability under these less separated grade scales complements the theoretical analysis and shows that the observed recovery is not tied to the maximally separated default breakpoint within the tested range.

Table 18: Transition-level hierarchy diagnostics for GDP-Hier (five-run mean±std, %). Purity and nonempty-child ratio are higher-is-better; violation is lower-is-better.

Dataset	Transition	Purity ↑	Violation ↓	Nonempty ↑
CIFAR100-HIER	super → fine	95.6±0.6	3.8±0.6	100.0±0.0
RCV1-HIER	super → fine	97.7±1.4	2.3±1.5	98.0±2.7
Aircraft	manufacturer → family	84.2±3.4	13.8±2.8	99.4±1.3
	family → variant	87.2±1.3	9.8±1.0	100.0±0.0
T200	ancestor3 → ancestor2	90.7±2.4	9.6±2.6	99.3±1.1
	ancestor2 → parent	97.3±0.7	2.7±0.8	94.9±1.4
	parent → leaf	97.9±0.7	2.1±0.9	95.7±1.3
T608	ancestor3 → ancestor2	95.4±1.5	3.9±1.1	97.0±3.5
	ancestor2 → parent	97.1±0.7	2.7±0.7	99.3±0.6
	parent → leaf	90.5±0.8	12.8±1.1	76.9±2.6

Table 19: K-means on the hierarchy-benchmark input representations at the prescribed resolutions (five-run mean±std, %). Mean averages the levelwise ACC/NMI/ARI values; lower HVR is better.

Dataset	Resolution	K	ACC	NMI	ARI	HVR↓
CIFAR100-HIER	fine (leaf)	100	53.8±0.4	69.9±0.3	38.2±0.6	–
	super	20	50.6±2.5	56.1±1.2	35.7±1.6	–
	Mean	–	52.2±1.3	63.0±0.6	37.0±0.9	13.2±1.1
RCV1-HIER	fine (leaf)	20	35.6±4.0	51.3±3.7	24.4±3.0	–
	super	4	59.1±10.3	38.8±6.2	28.3±5.9	–
	Mean	–	47.4±5.3	45.1±4.2	26.3±3.5	10.4±5.0
Aircraft	variant (leaf)	100	13.3±0.3	38.4±0.3	4.5±0.1	–
	family	70	13.6±0.2	32.4±0.3	4.3±0.1	–
	manufacturer	30	15.0±0.6	19.3±0.4	4.2±0.3	–
	Mean	–	14.0±0.2	30.0±0.1	4.4±0.2	27.0±0.6
T200	leaf	200	46.7±0.2	71.1±0.2	26.5±0.5	–
	parent	130	40.4±0.7	65.9±0.3	24.0±0.8	–
	ancestor2	81	34.9±0.1	58.6±0.2	20.3±0.2	–
	ancestor3	52	31.0±0.3	51.6±0.5	16.9±0.5	–
	Mean	–	38.2±0.2	61.8±0.1	21.9±0.3	21.8±1.0
T608	leaf	608	34.1±0.3	68.8±0.1	17.0±0.2	–
	parent	201	25.9±0.4	55.9±0.2	13.6±0.3	–
	ancestor2	108	25.3±0.5	50.0±0.3	13.9±0.4	–
	ancestor3	63	25.6±0.4	45.2±0.5	13.5±0.4	–
	Mean	–	27.7±0.2	55.0±0.2	14.5±0.1	26.7±0.9

Taken together, the levelwise results, joint-versus-separate comparison, transition diagnostics, independent K-means reference, agglomerative post-processing analysis, and breakpoint sweep support the joint multi-resolution formulation of GDP-Hier. The recovered partitions remain strong at individual resolutions while forming a coherent cross-level structure, and this behavior persists across hierarchy depths and the tested hierarchy-grade calibrations.

D.6 Component and mechanism analysis

Beyond the representative analysis in the main text, we examine the principal components across all seven flat benchmarks and use targeted stress tests and diagnostics to clarify the mechanisms they are designed to support.

Component ablations. Table 21 extends the main component comparison to all seven datasets under Binary, three Difficulty calibrations, and three Mix calibrations. Global BCE is an objective-form ablation that retains the same Global co-assignment $r_i^{\mathrm{glob}} = q_{a_i}^{\top} q_{b_i}$ but replaces the order-based Ranking objective with direct numerical fitting of the supplied grade:

$$\mathcal{L}_{\mathrm{BCE}}^{\mathrm{glob}} = -\frac{1}{|\mathcal{C}|} \sum_{i=1}^{|\mathcal{C}|} \left[g_i \log r_i^{\mathrm{glob}} + (1 - g_i) \log(1 - r_i^{\mathrm{glob}}) \right].$$

Table 20: Hierarchy recovery with agglomerative post-processing (five-run means, %). Each resolution reports ACC/NMI/ARI; Mean averages the levelwise metrics. Input features denote the fixed benchmark representations. The finest-level K-means partition is preserved, while all coarser partitions are obtained by cutting a single dendrogram; HVR is therefore 0 by construction and is omitted.

CIFAR100-HIER			
Representation	fine (leaf)	super	Mean
Input features	53.8 / 69.9 / 38.2	48.4 / 60.0 / 34.2	51.1 / 64.9 / 36.2
SpherePair	53.6 / 69.9 / 40.9	61.8 / 64.7 / 45.8	57.7 / 67.3 / 43.3
ProbPair	52.5 / 68.4 / 39.2	63.9 / 65.9 / 50.4	58.2 / 67.2 / 44.8
WPP	51.9 / 68.6 / 39.1	64.9 / 66.8 / 51.9	58.4 / 67.7 / 45.5
RCV1-HIER			
Representation	fine (leaf)	super	Mean
Input features	35.6 / 51.3 / 24.4	69.5 / 48.5 / 46.1	52.6 / 49.9 / 35.3
SpherePair	67.5 / 67.9 / 51.3	93.7 / 79.3 / 84.0	80.6 / 73.6 / 67.6
ProbPair	67.4 / 66.7 / 51.9	94.3 / 80.8 / 85.6	80.9 / 73.7 / 68.7
WPP	70.5 / 68.0 / 54.6	94.4 / 80.8 / 85.6	82.4 / 74.4 / 70.1
Aircraft			
Representation	variant (leaf)	family	manufacturer	Mean
Input features	13.3 / 38.4 / 4.5	13.4 / 31.5 / 4.1	16.2 / 18.2 / 3.4	14.3 / 29.4 / 4.0
SpherePair	13.5 / 39.1 / 5.6	13.4 / 32.7 / 5.0	15.5 / 22.5 / 5.1	14.1 / 31.4 / 5.2
ProbPair	14.7 / 40.7 / 5.4	16.2 / 36.9 / 7.3	29.5 / 31.8 / 26.9	20.1 / 36.4 / 13.2
WPP	14.9 / 41.1 / 5.7	16.1 / 37.4 / 7.5	28.2 / 32.2 / 24.5	19.7 / 36.9 / 12.6
T200					
Representation	leaf	parent	ancestor2	ancestor3	Mean
Input features	46.7 / 71.1 / 26.5	43.1 / 67.7 / 29.4	38.0 / 61.5 / 26.6	36.1 / 55.4 / 23.1	41.0 / 63.9 / 26.4
SpherePair	35.2 / 67.6 / 22.8	34.0 / 63.5 / 24.7	34.2 / 58.9 / 26.9	35.1 / 54.0 / 24.1	34.6 / 61.0 / 24.6
ProbPair	29.8 / 63.0 / 18.1	31.0 / 59.1 / 23.3	34.1 / 55.1 / 30.5	37.2 / 51.2 / 29.2	33.0 / 57.1 / 25.3
WPP	32.2 / 64.2 / 19.6	33.0 / 60.6 / 25.2	35.3 / 56.5 / 31.0	35.4 / 52.4 / 29.5	34.0 / 58.4 / 26.3
T608					
Representation	leaf	parent	ancestor2	ancestor3	Mean
Input features	34.2 / 68.8 / 17.1	29.4 / 58.4 / 19.4	30.4 / 53.3 / 22.6	31.0 / 47.9 / 19.0	31.2 / 57.1 / 19.5
SpherePair	17.5 / 60.9 / 7.9	20.4 / 50.7 / 15.4	25.8 / 47.1 / 20.5	29.3 / 43.7 / 19.1	23.3 / 50.6 / 15.7
ProbPair	18.1 / 60.5 / 7.5	20.1 / 49.5 / 15.1	25.5 / 45.6 / 21.6	31.7 / 42.8 / 25.8	23.8 / 49.6 / 17.5
WPP	18.8 / 61.0 / 8.1	20.2 / 49.8 / 14.8	26.2 / 46.7 / 22.9	31.1 / 43.4 / 25.5	24.1 / 50.3 / 17.8

The comparison therefore isolates whether Global supervision should preserve relative grade ordering or directly reproduce the numerical grade values. Under Difficulty, GDP exceeds Global BCE on ACC for six of the seven datasets at both $b = 0.3$ and $b = 0.5$, and for all seven at $b = 0.7$; NMI and ARI show the same broad pattern. Under Mix identity, Global BCE can remain competitive because the supplied grade already has a direct soft-overlap interpretation, whereas this competitiveness weakens under the targeted recalibrations, particularly for $\zeta = +3$. Together with the results in Appendix D.4, this supports the intended role of Ranking in reducing reliance on a particular numerical realization of the graded relations.

Under the same CIFAR100-20 Mix setting with 5k constraints and no autoencoder pretraining, Table 12 reports a test ACC of 65.0%, whereas the full GDP result in Table 21 is 59.8%. These results come from independently executed five-run experiment families rather than reused runs, yielding approximately $65.0 \pm 2.2\%$ and $59.8 \pm 2.2\%$ test ACC, respectively. Our audit found no discrepancy in data construction, aggregation, or nominal experimental configuration. The difference therefore indicates appreciable between-family variability in this setting; the available evidence does not isolate a single causal source for this variation. We retain both results because each is the corresponding five-run summary for its own matched experimental family.

The other component removals show broadly consistent directional effects. Removing LGRC lowers ACC, NMI, and ARI in all 49 dataset–condition cases, with many of the resulting degradations being substantial. Removing the assignment-gated vertex term lowers each metric in 47/49 cases, indicating a broadly consistent positive contribution, although some individual margins are comparatively small.

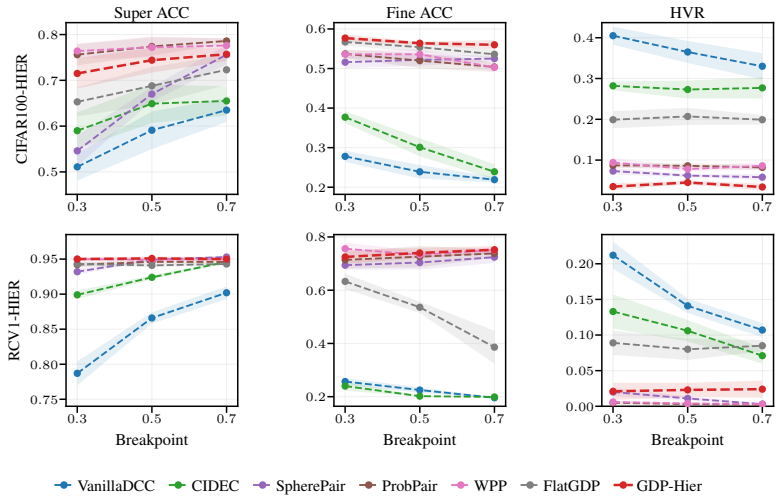

Figure 11: Hierarchy-grade breakpoint robustness on CIFAR100-HIER and RCV1-HIER (10k constraints; five-run mean±std). The intermediate breakpoint varies over $\beta_1 \in \{0.3, 0.5, 0.7\}$.

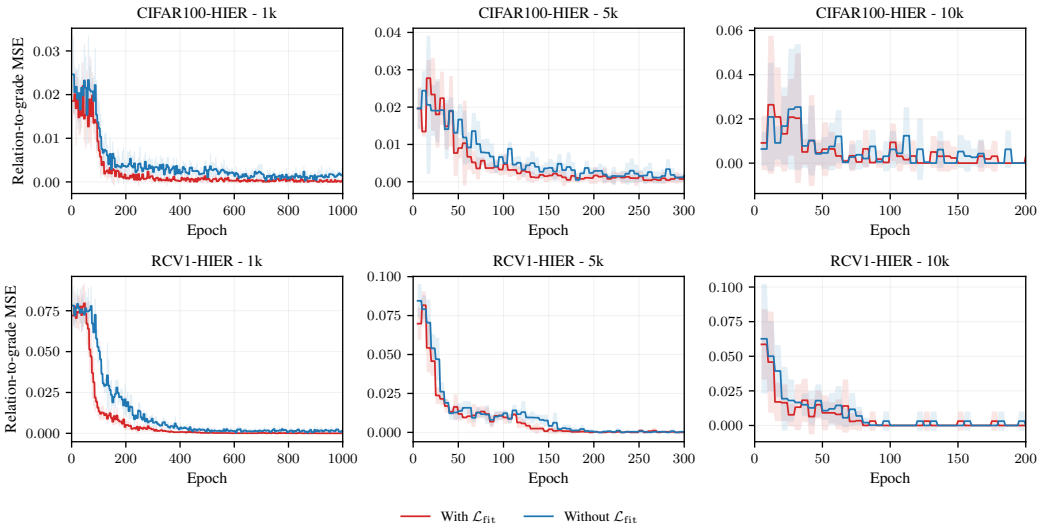

Figure 12: Composite relation-to-grade calibration with and without $\mathcal{L}_{\text{fit}}$ on CIFAR100-HIER and RCV1-HIER across three constraint budgets. Curves report relation-to-grade MSE over training.

Marginal entropy is more condition-dependent on the standard flat benchmarks, motivating a targeted stress test on T200 (Table 22), whose irregular hierarchy is better suited to assessing whether broad prototype utilization can become harmful. The results provide clear evidence for its stabilizing role: removing H_{marg} reduces the nonempty-child ratios from 94.9–99.3% to 6.5–20.0% across the three transitions, while leaf ACC falls from 52.9% to 6.4%. Thus, rather than imposing a harmful balance bias on this irregular hierarchy, marginal entropy prevents progressive prototype under-utilization as the resolution becomes finer. For hierarchy recovery, $\mathcal{L}_{\text{fit}}$ directly calibrates the composite relation to the prescribed hierarchy grades. As shown in Figure 12, $\mathcal{L}_{\text{fit}}$ consistently reduces relation-to-grade MSE, with the clearest effect under sparse supervision.

Local–Global coordination and assignment geometry. We use three complementary diagnostics to examine whether the learned internal structure follows the intended Local–Global design.

Table 21: Component ablations across the seven flat benchmarks without autoencoder pretraining at 5k constraints (five-run means, %). Each cell reports test ACC/NMI/ARI; best/second-best values within each dataset–condition–metric comparison are **bold**/underlined.

Variant	Binary	Diff. $b = 0.3$	Diff. $b = 0.5$	Diff. $b = 0.7$	Mix $\tau = 0.2$	Mix identity	Mix $\zeta = +3$
CIFAR100-20							
GDP (Ours)	**69.8 / 66.1 / 55.2**	**65.7 / 63.3 / 50.6**	**64.7 / 61.8 / 48.2**	**59.2 / 58.0 / 43.3**	57.4 / 53.7 / 39.0	59.8 / 55.6 / 40.9	**63.9 / 59.3 / 46.5**
Global BCE	65.3 / 63.8 / 51.6	62.8 / 58.9 / 44.9	56.8 / 54.8 / 34.5	52.0 / 51.0 / 27.6	54.9 / 54.6 / 38.7	**63.0 / 61.3 / 47.8**	55.8 / 55.8 / 40.5
w/o LGRC	25.8 / 25.4 / 11.2	45.8 / 42.5 / 31.9	30.9 / 27.5 / 16.9	33.9 / 32.8 / 20.5	41.4 / 40.2 / 27.4	30.6 / 27.9 / 15.5	25.6 / 21.7 / 10.4
w/o $\mathcal{L}_{\text{vertex}}$	65.9 / 59.9 / 47.7	58.1 / 54.2 / 41.2	49.7 / 45.9 / 31.9	48.3 / 41.9 / 29.0	55.9 / 49.3 / 36.1	54.9 / 47.8 / 34.6	50.8 / 42.3 / 29.9
w/o H_{marg}	**69.8 / 67.2** / 54.3	52.4 / 57.0 / 22.4	41.4 / 48.8 / 10.8	33.9 / 40.4 / 5.6	**59.3 / 54.9 / 42.2**	62.0 / 58.0 / 45.3	59.4 / 56.7 / 43.2
CIFAR10							
GDP (Ours)	94.8 / **89.2** / 89.2	**95.1 / 89.4 / 89.7**	**95.0 / 89.3 / 89.5**	**94.9 / 89.1 / 89.4**	94.3 / 87.9 / 88.1	**94.9 / 89.2 / 89.3**	**94.6 / 88.7 / 88.8**
Global BCE	94.8 / 89.1 / 89.1	94.9 / 89.0 / 89.3	94.8 / 88.6 / 89.0	94.6 / 88.2 / 88.6	**94.5 / 88.5 / 88.6**	94.6 / 88.8 / 88.7	93.1 / 87.6 / 86.9
w/o LGRC	61.1 / 49.4 / 41.3	69.1 / 63.4 / 58.2	36.6 / 28.1 / 19.3	53.7 / 43.0 / 37.6	66.1 / 61.5 / 54.0	70.3 / 65.5 / 59.7	70.0 / 61.5 / 56.4
w/o $\mathcal{L}_{\text{vertex}}$	93.8 / 86.8 / 87.1	94.5 / 88.4 / 88.5	94.3 / 87.9 / 88.0	93.9 / 87.1 / 87.2	94.1 / 87.7 / 87.7	94.6 / 88.6 / 88.7	93.7 / 87.0 / 87.0
w/o H_{marg}	**94.9 / 89.2 / 89.3**	94.9 / 89.2 / 89.2	93.3 / 89.0 / 87.9	94.8 / 89.0 / 89.1	94.0 / 87.5 / 87.6	94.1 / 88.8 / 88.2	92.9 / 88.0 / 86.7
FMNIST							
GDP (Ours)	**80.2 / 71.4 / 65.4**	**80.3** / 71.7 / **65.3**	**81.0 / 71.6 / 65.6**	**81.5 / 71.8 / 66.7**	76.7 / 68.2 / 61.8	76.9 / 69.3 / 62.7	**77.6 / 68.8 / 62.8**
Global BCE	80.0 / 71.1 / 65.2	79.0 / 70.2 / 64.5	78.4 / 68.2 / 62.1	72.4 / 65.4 / 52.3	**77.4 / 68.3 / 62.0**	**77.9 / 69.8 / 63.6**	76.3 / 67.9 / 61.6
w/o LGRC	58.6 / 43.0 / 36.8	74.5 / 69.7 / 61.7	55.7 / 43.1 / 36.9	41.4 / 29.4 / 23.1	48.3 / 35.6 / 27.9	54.5 / 45.6 / 37.0	62.8 / 52.5 / 44.6
w/o $\mathcal{L}_{\text{vertex}}$	77.3 / 67.8 / 61.3	77.9 / 69.4 / 61.4	73.1 / 64.8 / 53.8	72.0 / 63.3 / 49.6	74.5 / 64.7 / 58.6	76.4 / 67.3 / 61.2	74.6 / 65.1 / 58.8
w/o H_{marg}	79.5 / 71.3 / 64.7	77.9 / **72.2** / 63.6	72.9 / 70.9 / 52.9	64.9 / 64.7 / 36.2	74.5 / 67.0 / 59.1	75.8 / 69.0 / 61.9	72.4 / 67.3 / 58.1
ImageNet10							
GDP (Ours)	**95.9 / 90.4 / 91.0**	**95.6 / 90.1 / 90.5**	**95.2 / 89.5 / 89.7**	**95.8 / 90.1 / 90.8**	94.2 / 87.2 / 87.7	95.8 / 90.2 / 90.9	94.6 / 87.5 / 88.4
Global BCE	95.2 / 89.2 / 89.7	92.0 / 82.8 / 83.1	86.8 / 74.3 / 73.8	82.6 / 69.4 / 65.1	94.3 / 87.1 / 87.8	95.8 / 90.4 / 91.0	89.2 / 80.4 / 79.6
w/o LGRC	62.8 / 50.0 / 43.7	41.8 / 29.7 / 24.6	56.5 / 48.0 / 42.7	51.8 / 41.2 / 39.2	87.0 / 80.8 / 78.4	71.1 / 62.8 / 55.9	88.4 / 82.1 / 79.9
w/o $\mathcal{L}_{\text{vertex}}$	94.3 / 87.1 / 87.7	90.7 / 82.6 / 80.3	83.1 / 75.0 / 67.5	83.0 / 75.4 / 69.4	93.7 / 86.1 / 86.7	95.9 / 90.4 / 91.2	92.8 / 84.0 / 84.8
w/o H_{marg}	95.6 / 89.9 / 90.5	91.9 / 87.1 / 85.7	82.1 / 79.0 / 68.4	68.5 / 66.0 / 43.0	**94.8 / 87.9 / 88.8**	**96.2 / 91.0 / 91.7**	**94.7 / 88.2 / 88.8**
MNIST							
GDP (Ours)	**95.7 / 89.4 / 90.7**	**94.8 / 87.6 / 89.0**	**94.7 / 87.2 / 88.6**	**94.5 / 87.0 / 88.3**	**91.5 / 81.8 / 82.6**	93.3 / 84.9 / 86.0	**91.6 / 83.3 / 83.7**
Global BCE	95.5 / 89.1 / 90.4	94.0 / 86.0 / 87.3	93.3 / 84.8 / 86.1	93.0 / 84.0 / 85.2	90.0 / 81.3 / 81.1	**93.6 / 85.5 / 86.7**	87.5 / 78.5 / 77.5
w/o LGRC	63.9 / 65.5 / 59.5	63.8 / 56.2 / 51.9	56.4 / 46.3 / 42.7	44.9 / 33.3 / 28.0	47.5 / 35.8 / 28.0	72.1 / 64.2 / 61.1	68.1 / 58.3 / 53.9
w/o $\mathcal{L}_{\text{vertex}}$	92.1 / 82.7 / 83.7	90.5 / 80.6 / 80.9	90.4 / 80.5 / 80.0	88.2 / 78.0 / 75.1	89.6 / 78.4 / 79.0	91.7 / 81.9 / 82.8	89.7 / 78.3 / 79.1
w/o H_{marg}	95.4 / 89.1 / 90.3	92.7 / 85.9 / 86.1	94.3 / 86.8 / 88.0	82.7 / 81.5 / 74.8	90.8 / 80.6 / 81.2	91.0 / 82.9 / 83.1	83.2 / 75.9 / 72.1
Reuters							
GDP (Ours)	**95.1 / 81.9 / 88.2**	92.9 / 77.2 / 84.4	92.8 / **76.3** / 83.8	**92.9 / 76.1 / 83.9**	93.2 / **77.3 / 84.6**	**94.7 / 80.8 / 87.7**	**92.6 / 74.7 / 82.7**
Global BCE	95.0 / 81.6 / 87.9	93.3 / 77.7 / 85.1	**93.0** / 76.1 / **84.1**	90.8 / 71.2 / 79.7	91.5 / 73.3 / 81.3	94.5 / 80.2 / 87.0	90.7 / 69.3 / 77.2
w/o LGRC	85.9 / 72.5 / 75.5	59.9 / 31.5 / 31.5	60.5 / 35.8 / 34.7	67.4 / 43.3 / 43.9	78.2 / 57.2 / 61.7	90.2 / 73.3 / 78.1	89.2 / 70.7 / 77.9
w/o $\mathcal{L}_{\text{vertex}}$	92.3 / 75.7 / 82.9	**93.5 / 77.9 / 85.6**	92.5 / 75.4 / 83.0	91.2 / 72.8 / 80.2	86.9 / 65.8 / 75.7	91.2 / 72.9 / 81.7	88.4 / 65.5 / 73.4
w/o H_{marg}	94.5 / 80.4 / 86.9	91.4 / 74.9 / 81.5	88.5 / 70.7 / 77.0	88.3 / 70.1 / 76.0	**93.3** / 77.1 / 84.1	94.5 / 80.3 / 86.9	90.3 / 70.8 / 77.1
STL10							
GDP (Ours)	**89.4 / 80.4 / 78.7**	**89.0 / 79.9 / 77.9**	**87.5 / 78.6 / 75.9**	**87.6 / 79.0 / 76.4**	89.1 / 79.7 / 78.0	89.3 / 80.2 / 78.5	**89.1 / 79.9 / 78.2**
Global BCE	**89.4** / 80.1 / **78.7**	86.0 / 77.2 / 74.0	85.7 / 73.8 / 71.8	82.4 / 68.7 / 65.6	89.0 / 79.6 / 78.1	**89.4 / 80.3 / 78.7**	87.4 / 77.8 / 75.7
w/o LGRC	52.1 / 37.5 / 30.2	60.4 / 50.1 / 43.9	70.7 / 65.0 / 56.8	53.7 / 42.3 / 37.3	72.9 / 62.5 / 56.1	71.3 / 63.6 / 57.1	71.3 / 62.9 / 56.4
w/o $\mathcal{L}_{\text{vertex}}$	88.5 / 78.6 / 77.1	83.6 / 75.2 / 65.7	82.7 / 74.2 / 63.7	77.4 / 70.9 / 54.5	88.6 / 78.9 / 77.2	89.1 / 80.0 / 78.3	87.2 / 76.9 / 74.9
w/o H_{marg}	89.3 / 80.3 / 78.6	81.8 / 77.7 / 69.5	79.7 / 76.2 / 60.8	71.1 / 70.1 / 42.8	**89.3 / 79.9 / 78.5**	88.1 / 79.6 / 77.6	86.3 / 78.4 / 75.1

Table 22: Marginal-entropy ablation on T200 at 5k constraints. Each resolution reports test ACC/NMI/ARI (%); the final column gives the nonempty-child ratios for parent→leaf, ancestor2→parent, and ancestor3→ancestor2.

Variant	HVR ↓	leaf	parent	ancestor2	ancestor3	Nonempty child (%)
Full GDP-Hier	**4.8**	**52.9 / 77.3 / 40.5**	**50.7 / 75.0 / 40.7**	**45.9 / 68.9 / 34.7**	**39.5 / 61.3 / 26.6**	**95.7 / 94.9 / 99.3**
w/o H_{marg}	11.6	6.4 / 49.2 / 6.2	18.8 / 50.9 / 12.3	29.2 / 53.6 / 21.8	35.8 / 52.3 / 25.3	6.5 / 10.6 / 20.0

(i) *Local–Global coordination.* Table 23 compares the final partitions produced by the two branches. The Local partition is obtained by K-means on the Local representation, while the Global partition is given by the maximum-probability Global assignment. ACC is computed after optimal permutation of the two cluster labelings, whereas NMI and ARI are permutation-invariant. Local–Global ACC exceeds 88% in every dataset–source case and is near-perfect throughout Binary supervision, showing that the two branches converge to closely coordinated partitions. Figure 13 provides a complementary relation-level view by tracking the Spearman correlation between the supplied grades g_i and the Local or Global relation scores. Under Difficulty and Mix, both branches develop positive grade–relation correlations across all seven datasets, indicating that their coordination extends beyond the final hard partition to the ordering of graded relations.

Table 23: Final Local–Global partition agreement at 5k constraints without autoencoder pretraining. ACC compares the two partitions after optimal label permutation; NMI and ARI are permutation-invariant. Values are five-run means (%); higher is better.

Source	CIFAR100-20	CIFAR10	FMNIST	ImageNet10	MNIST	Reuters	STL10
Binary	98.3/97.0/96.4	99.9/99.6/99.7	98.7/97.0/97.3	99.9/99.7/99.8	99.8/99.4/99.6	99.6/98.1/99.1	99.5/98.7/98.9
Difficulty	95.1/93.0/87.7	99.5/98.7/98.9	92.7/88.3/82.1	97.9/95.7/95.2	99.2/97.7/98.3	97.8/92.8/95.3	97.7/95.2/94.7
Mix	88.2/86.5/72.2	99.8/99.4/99.6	98.1/95.3/96.0	99.4/98.4/98.6	99.2/97.7/98.3	99.5/97.5/98.9	99.1/97.7/98.1

Table 24: Source-specific semantic alignment in the Global branch at 5k constraints without autoencoder pretraining. Difficulty reports decisiveness alignment and Mix reports ambiguity–entropy alignment. Values are Spearman correlations (five-run mean±std); larger positive values indicate stronger alignment with the corresponding graded semantics.

Global semantic alignment (ρ_s)	CIFAR100-20	CIFAR10	FMNIST	ImageNet10	MNIST	Reuters	STL10
Difficulty: decisiveness	0.448 ± 0.06	0.599 ± 0.04	0.582 ± 0.07	0.721 ± 0.03	0.645 ± 0.02	0.554 ± 0.03	0.726 ± 0.03
Mix: ambiguity–entropy	0.194 ± 0.05	0.370 ± 0.06	0.389 ± 0.02	0.476 ± 0.04	0.467 ± 0.01	0.527 ± 0.02	0.391 ± 0.03

(ii) *Assignment geometry.* Figure 14 examines the assignment concentration $c(q)$ from Eq. (6), with test samples stratified post hoc according to whether the Global argmax prediction is correct. Correct predictions have higher terminal concentration in every dataset–source case, consistent with stronger prototype calibration being applied to more concentrated assignments. Figure 15 further tracks the standard deviation of Global relation scores over diagnostic pairs. Its terminal ordering is Binary $>$ Difficulty $>$ Mix on all seven datasets, showing that the learned Global relations become progressively less polarized as the supervision moves from hard binary relations toward more graded and mixed-membership structure.

(iii) *Source-specific semantic alignment.* Table 24 tests whether the Global geometry preserves the semantics specific to each graded source. Under Difficulty, we correlate source decisiveness $|g_i - b|$ with Global decisiveness $|r_i^{\mathrm{glob}} - b|$ over diagnostic pairs. Under Mix, we correlate held-out mixture ambiguity $1 - |2 \max_c p_c - 1|$ with Global assignment entropy $-\sum_k q_k \log q_k$. Both correlations are positive on all seven datasets: relations farther from the Difficulty boundary are represented more decisively in the Global branch, while more ambiguous mixed samples receive less concentrated Global assignments. These diagnostics show that GDP coordinates the two branches while preserving source-relevant graded structure in the resulting assignment geometry.

D.7 Hyperparameter sensitivity

We evaluate sensitivity around the shared GDP configuration by varying one factor at a time under Binary supervision on the seven flat benchmarks, using 1k, 5k, and 10k constraints. The shared defaults are $\alpha = 0.2$, $\gamma = 1$, $\eta = 2$, $\lambda_{\mathrm{rec}} = 0.02$, and a synchronized Global update/feedback interval of 5. Figures 16–20 report test ACC, NMI, and ARI across the three supervision budgets. Overall, the defaults lie within broad stable regions, while the effect of parameter choice is more visible when pairwise supervision is sparse.

Loss weights. The four loss weights exhibit different degrees of sensitivity, but none requires a narrow operating point around the selected default. For the marginal-entropy weight α (Figure 16), the range 0.1–0.5 is broadly stable across budgets, whereas the larger value $\alpha = 1$ can degrade performance, most visibly on Reuters and under sparse supervision on CIFAR100-20. The vertex-calibration weight γ (Figure 17) shows the clearest budget dependence. With 1k constraints, very weak calibration at $\gamma = 0.1$ is notably less effective on CIFAR100-20, while the largest value $\gamma = 5$ can also be unstable, most clearly on ImageNet10; these differences contract substantially at 5k and

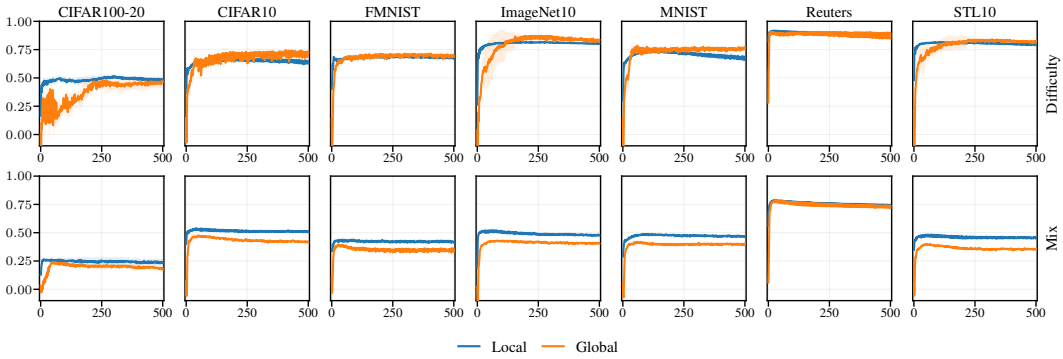

Figure 13: Local and Global grade–relation Spearman correlation over training under Difficulty and Mix supervision (5k constraints without autoencoder pretraining; five-run mean±std).

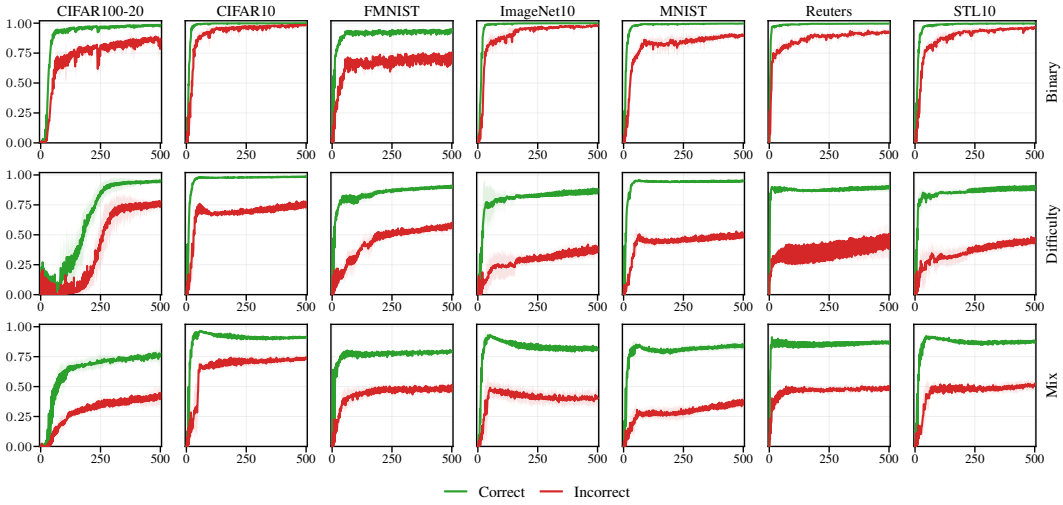

Figure 14: Assignment concentration $c(q)$ for post-hoc correct and incorrect Global predictions over training under Binary, Difficulty, and Mix supervision (5k constraints without autoencoder pretraining; five-run mean±std).

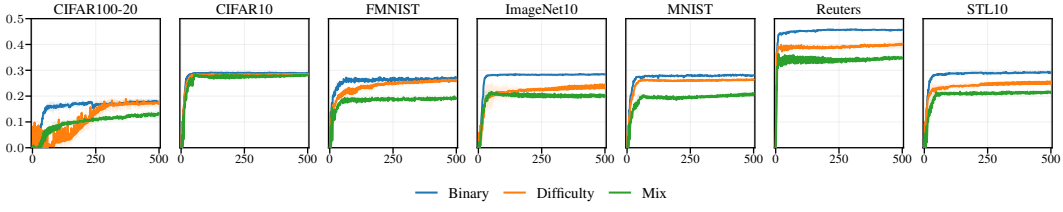

Figure 15: Dispersion of Global relation scores over training under Binary, Difficulty, and Mix supervision (5k constraints without autoencoder pretraining; five-run mean±std).

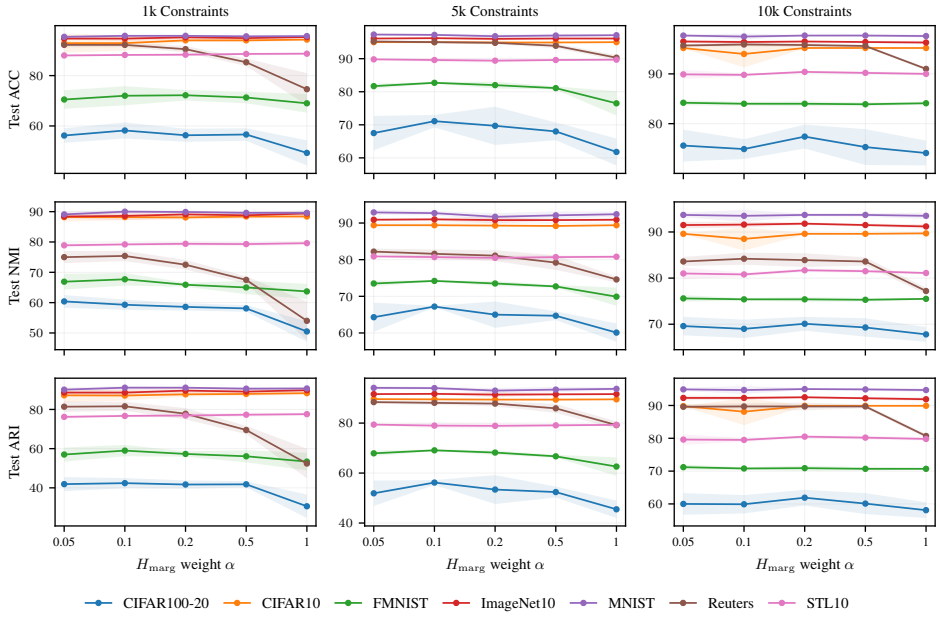

Figure 16: Sensitivity to the marginal-entropy weight α (default 0.2). Columns correspond to 1k, 5k, and 10k constraints, and rows report test ACC, NMI, and ARI for pretrained GDP under Binary supervision. Curves show five-run means with standard-deviation bands.

10k, where the region around $\gamma = 0.5$–2 is consistently strong. The LGRC weight η (Figure 18) is similarly stable over a moderate range, with $\eta = 1$–2 providing strong performance across the three budgets and larger values becoming less reliable on datasets such as FMNIST and Reuters. Reconstruction is the least sensitive of the four terms (Figure 19): $\lambda_{\text{rec}} = 0$–0.05 produces only modest changes across the tested settings, while the largest value 0.1 introduces noticeable instability for CIFAR100-20 at 10k. These results place the shared loss weights in stable interior regions rather than at isolated dataset-specific optima.

Update and feedback interval. The synchronized Global update/feedback interval controls how frequently the Global branch is updated and its feedback is propagated to the Local branch. Figure 20 shows that very frequent updates are less favorable in several settings, especially under sparse supervision. An interval of 1 is consistently weaker on FMNIST and is also clearly suboptimal for MNIST at 1k and 5k constraints, whereas intervals of 3 and 5 yield stronger and more stable performance. Increasing the interval further to 10 provides no consistent additional benefit. The default interval of 5 therefore offers a robust operating point across supervision budgets while avoiding unnecessarily frequent Global updates.

Taken together, the sensitivity sweeps support the shared configuration used throughout the main experiments. The selected defaults remain within stable performance regions across constraint budgets, while the increased variation under 1k supervision is concentrated mainly in the vertex-calibration strength and update frequency rather than reflecting broad hyperparameter fragility.

E Learning efficiency

Appendix B.2 analyzes the computational complexity of GDP and GDP-Hier. Under fixed model settings, GDP scales linearly with the dataset and constraint-set sizes, while GDP-Hier shares a single Local backbone across hierarchy resolutions and adds only resolution-specific Global computation. We complement this theoretical analysis with empirical wall-clock measurements in Tables 25 and 26. The reported times cover the complete training pipeline on a single NVIDIA H200 141GB GPU, including autoencoder pretraining where applicable. For VolMaxDCC, they additionally include the dataset-specific validation search and therefore represent end-to-end model-selection cost rather than

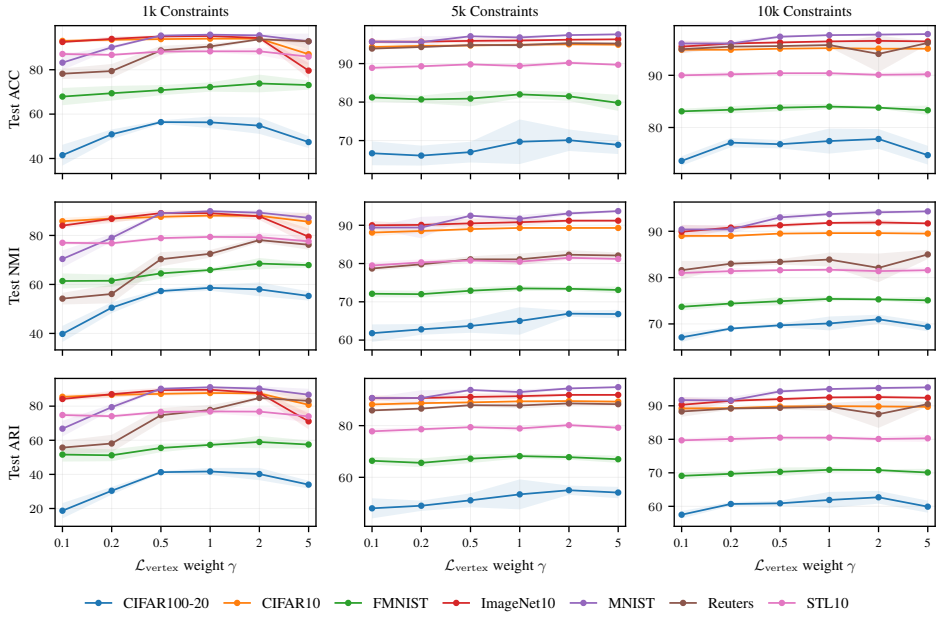

Figure 17: Sensitivity to the vertex-calibration weight γ (default 1). Columns correspond to 1k, 5k, and 10k constraints, and rows report test ACC, NMI, and ARI for pretrained GDP under Binary supervision. Curves show five-run means with standard-deviation bands.

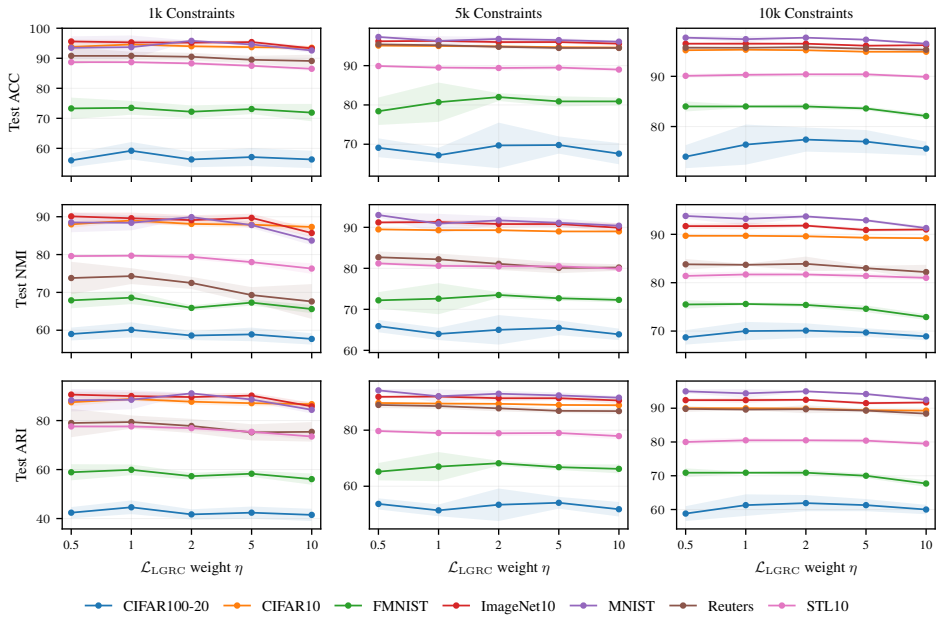

Figure 18: Sensitivity to the LGRC weight η (default 2). Columns correspond to 1k, 5k, and 10k constraints, and rows report test ACC, NMI, and ARI for pretrained GDP under Binary supervision. Curves show five-run means with standard-deviation bands.

the optimization time of a single fixed configuration; GDP and GDP-Hier use the shared default configuration adopted throughout our experiments. For hierarchy benchmarks, the measurements report the total time required to obtain all prescribed resolutions.

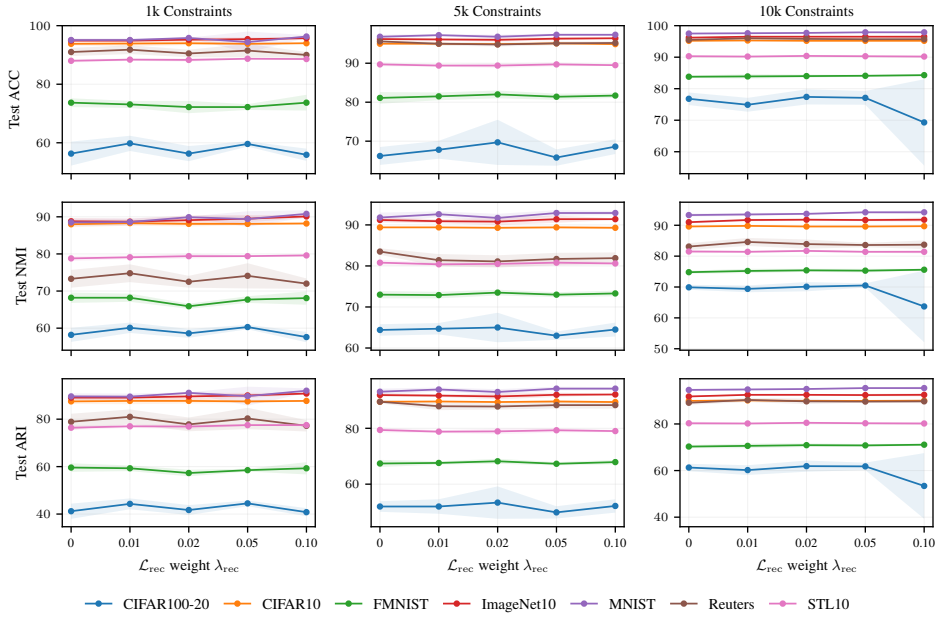

Figure 19: Sensitivity to the reconstruction weight λ_{rec} (default 0.02). Columns correspond to 1k, 5k, and 10k constraints, and rows report test ACC, NMI, and ARI for pretrained GDP under Binary supervision. Curves show five-run means with standard-deviation bands.

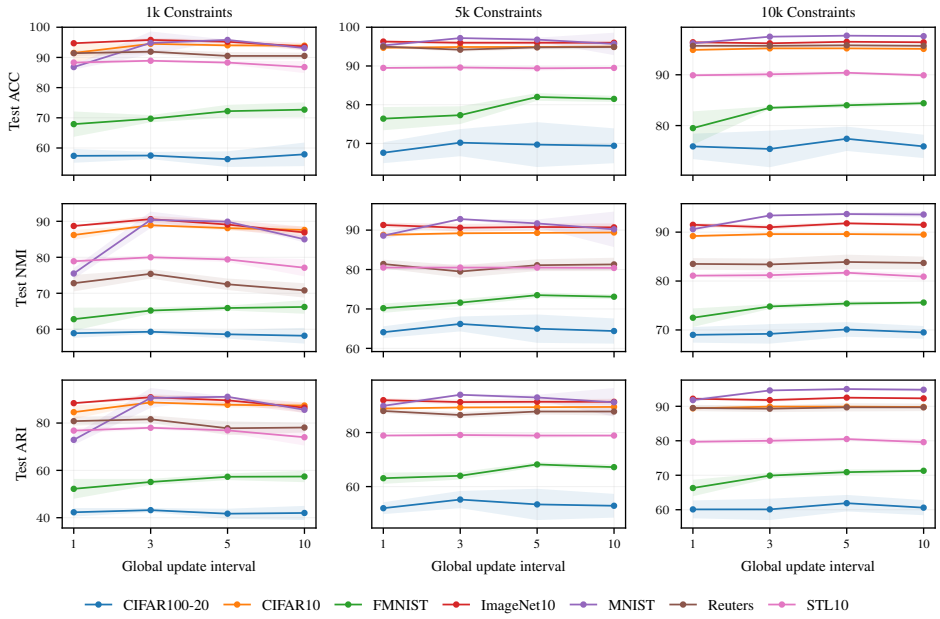

Figure 20: Sensitivity to the synchronized Global update/feedback interval (default 5). Columns correspond to 1k, 5k, and 10k constraints, and rows report test ACC, NMI, and ARI for pretrained GDP under Binary supervision. Curves show five-run means with standard-deviation bands.

Flat clustering. Table 25 shows that GDP remains computationally comparable to the established pretrained pairwise methods while delivering the stronger clustering performance reported in the main experiments. In particular, the additional Global assignment and Local–Global coordination objectives do not introduce a substantial wall-clock overhead relative to SpherePair, ProbPair, and

Table 25: Overall wall-clock training time for different methods on seven flat datasets using 5k Mix constraints, measured on a single NVIDIA H200 141GB GPU. Times requiring hyperparameter search are underlined.

	CIFAR100-20	CIFAR10	FMNIST	ImageNet10	MNIST	Reuters	STL10
VanillaDCC	14s	14s	14s	14s	14s	14s	14s
VolMaxDCC	16m38s	16m30s	18m58s	16m20s	18m47s	3m35s	16m13s
CIDEC	22m17s	22m04s	27m11s	5m06s	27m49s	5m42s	5m07s
SpherePair	17m31s	17m03s	21m39s	4m13s	24m33s	4m48s	4m09s
ProbPair	17m31s	17m03s	21m37s	4m13s	24m31s	4m47s	4m09s
Weighted ProbPair	17m33s	17m05s	21m23s	4m16s	24m50s	4m49s	4m11s
ECI-PP	22m54s	22m19s	27m34s	8m48s	29m57s	9m45s	8m50s
GDP (Ours)	17m05s	16m40s	20m59s	3m57s	24m04s	4m31s	3m52s

Table 26: Overall wall-clock time to obtain all prescribed hierarchy resolutions on five datasets using 5k constraints, measured on a single NVIDIA H200 141GB GPU. Methods marked with * are trained independently at each hierarchy resolution.

	CIFAR100-HIER	RCV1-HIER	Aircraft	T200	T608
VanillaDCC*	26s	27s	39s	52s	56s
CIDEC*	37m35s	15m45s	12m18s	4h57m51s	17h33m32s
SpherePair	17m24s	7m27s	3m58s	1h05m30s	3h02m17s
ProbPair	17m23s	7m21s	3m59s	1h06m14s	3h03m06s
Weighted ProbPair	17m25s	7m28s	4m04s	1h05m37s	3h06m38s
FlatGDP*	47m36s	19m22s	25m29s	5h18m42s	14h06m50s
GDP-Hier (Ours)	33m42s	11m47s	20m37s	1h36m51s	4h09m39s

Weighted ProbPair, and GDP is consistently more efficient than CIDEC and ECI-PP across the seven benchmarks. VolMaxDCC additionally incurs dataset-specific hyperparameter search, whereas GDP uses the same shared configuration throughout. Overall, the flat-clustering results indicate that the gains of GDP are achieved without materially increasing the practical training cost.

Multi-resolution hierarchy recovery. The benefit of joint optimization becomes more apparent as multiple hierarchy resolutions are required. Methods such as CIDEC and FlatGDP must repeat the end-to-end training procedure for each target K, so their total cost grows with the number of prescribed resolutions. GDP-Hier instead shares the Local backbone and jointly optimizes the resolution-specific Global heads, as anticipated by the complexity analysis in Appendix B.2. On T200 and T608, repeated resolution-wise training is considerably more expensive. Importantly, this efficiency is obtained together with the stronger levelwise recovery and markedly lower HVR reported in Table 2. GDP-Hier therefore combines joint multi-resolution prediction with a substantially lower cost than repeatedly training an end-to-end clustering model at each resolution.