GRANITE: GRANULAR-BALL AUTOENCODING WITH SIGNED TRIPLET EVIDENCE FOR UNCERTAINTY-AWARE DEEP CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Clustering is a fundamental task in data mining and plays an important role in complex data analysis. Traditional single-granularity instance-level clustering methods treat samples as isolated points, limiting their ability to learn multi-scale representations. Granular-ball (GB) clustering obtains multi-granularity representations through spherical topology evolution, making it more suitable for handling complex data. However, existing GB clustering methods often rely on fixed features and overlook uncertainty in ball–cluster assignments. To address these limitations, we propose the **GRAN**ular-ball autoencoding with **I**ntuitionistic **T**riplet **E**vidence (**GRANITE**) algorithm for uncertainty-aware deep clustering. A GB autoencoder learns latent representations while granular balls are reconstructed as sample representations evolve. In addition, to explicitly characterize the uncertainty in ball–cluster relationships, we introduce intuitionistic fuzzy triplet evidence comprising membership, non-membership, and hesitation degrees. Furthermore, GRANITE uses signed evidence and a covariance-aware GB distribution loss to regularize clustering. Experiments on multiple benchmark datasets demonstrate that GRANITE achieves superior clustering performance compared with existing methods. The code is available for anonymous review at *https://anonymous.4open.science/r/111111-3A89*.

1 INTRODUCTION

Clustering is a fundamental task in unsupervised learning that aims to discover latent group structures based on similarities among data Jain (2010), providing an important means for understanding the underlying patterns of complex data. According to their modeling mechanisms, clustering methods can be categorized into spectral clustering, hierarchical clustering, density-based clustering, and centroid-based clustering, among others Chen et al. (2025). Among them, centroid-based clustering is widely adopted for its simplicity and effectiveness Miklautz et al. (2025).

Traditional center-based clustering methods, as illustrated in Fig. 1(a), typically treat individual samples as the basic units of analysis, representing them as individual points in the feature space, and perform clustering based on pairwise similarity or distance. Representative methods include k-meansLloyd (1982) and fuzzy C-meansBezdek et al. (1984). Subsequent studies have refined clustering objectives through residual modeling, alternative distance measures, and additional prior information Jing et al. (2026); Yu et al. (2024); Wang et al. (2026). Although these advances have improved clustering performance, these methods still optimize at a single-granularity and make limited explicit use of local groups and higher-level structures.

Granular-ball (GB) clustering, takes GBs as basic units and constructs multi-granularity representations through spherical topology evolution, as illustrated in Fig. 1(b), providing a new perspective for capturing multi-scale structures in complex data. Building on this property, researchers have integrated GB computing with various clustering mechanisms to develop a range of methods. For example, Xia et al. (2022) incorporated GB computing into k-means to develop GB clustering; Xia et al. (2025a) introduced GBCT, which achieves efficient adaptive clustering through novel splitting and merging measures; Ni et al. (2026) proposed GB density peaks clustering to improve clustering performance. These studies demonstrate that GB representations can provide structural information

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

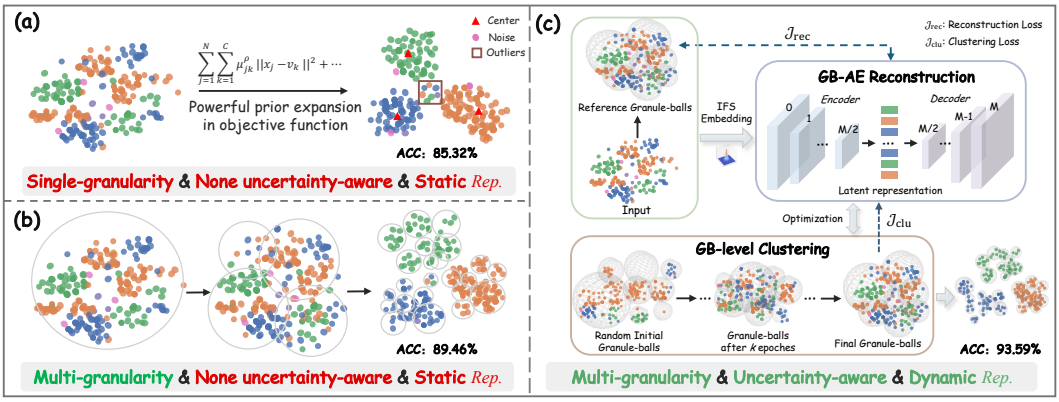


Figure 1: Illustration of the various clustering paradigms. (a) Traditional center-based clustering method, (b) GB clustering and (c) The proposed paradigm. *Rep*. denotes representations.

at different granularities and show potential for improving computational efficiency and robustness Xia et al. (2025b); Li et al. (2025).

However, existing GB clustering methods still have inevitable limitations. First, they often rely on static characterizations, which constrain the geometric relationships among GBs to the initial feature space and make it difficult to use feedback from the clustering objective. Second, ball–cluster assignments may be ambiguous, especially near cluster boundaries or in overlapping regions, while existing methods often lack explicit modeling of this uncertainty. With the advancement of computational power, deep clustering has attracted extensive attention. Among them, autoencodersBerahmand et al. (2024) can learn nonlinear latent representations through reconstruction, providing a feasible way to adaptively learn dynamic features at the ball level. In addition, intuitionistic fuzzy sets, as an extension of classical fuzzy sets Qian (2023); Ding et al. (2024); Yang et al. (2020), represent the relationship between a sample and a cluster as triplet evidence composed of membership, non-membership, and hesitation degrees, providing a new way to model uncertainty in assignments.

On this basis, we propose the **GRAN**ular-ball autoencoding with **I**ntuitionistic **T**riplet **E**vidence (**GRANITE**) algorithm for uncertainty-aware deep clustering. As illustrated in Fig. 1(c), we design a GB-level autoencoder and compute an intrinsic hesitation descriptor specific to each initial GB. The encoder maps the descriptors into a latent space, and the decoder reconstructs the center, radius, and intrinsic hesitation, encouraging the latent representations to preserve local geometry and within-ball heterogeneity. In the latent space, GRANITE introduces intuitionistic fuzzy triplet evidence consisting of membership, non-membership, and cluster-specific hesitation degrees for each ball–cluster relationship, and defines a new form of hesitation evidence to jointly account for the evidence supporting and opposing the assignment. In addition, we design a covariance-aware GB distribution loss to account for residual correlations, large deviations, and differences in the contributions of different GBs. GRANITE is a novel GB clustering paradigm that alleviates the limitations caused by static feature representations and insufficient use of uncertainty, providing a new perspective for uncertainty-aware GB clustering.

The main contributions of this work are summarized as follows:

- We design a GB-level autoencoder to learn dynamic deep feature in the latent space, alleviating the limitations imposed by fixed feature representations on existing GB algorithms.
- We introduce intuitionistic fuzzy triplet evidence to encode the supporting, opposing, and hesitation states arising in the clustering process, thereby explicitly characterizing uncertainty in ball–cluster assignments.
- We formulate a novel form of hesitation evidence and introduce a GB distribution measure to regularize the clustering process, thereby improving clustering robustness and efficiency.
- We propose a new GB clustering paradigm that jointly optimizes reconstruction and clustering objectives, thereby achieving uncertainty-aware clustering.

2 PRELIMINARIES

An intuitionistic fuzzy set (IFS) extends a fuzzy set by introducing membership and non-membership degrees, with the hesitation degree defined as the residual uncertainty Atanassov (1989); Chaira (2011); a detailed introduction is provided in Appendix A.1. GB computing represents locally related samples as hyperspherical granules characterized by centers and radii, providing a multi-granularity description of the data distribution Xia et al. (2024); see Appendix A.2 for details. An autoencoder (AE) learns nonlinear representations through an encoder–decoder reconstruction objective Zhang et al. (2019); Wei et al. (2025); Ren et al. (2025), and GRANITE extends this principle to structured IFGB representations; the detailed formulation is given in Appendix A.3.

3 METHODOLOGY

3.1 GRANULAR-BALL INITIALIZATION AND INTRINSIC HESITATION

GB methods are more effective at capturing multi-scale structures in complex data. Given the initial sample representations $\mathbf{X}_0 = \{\mathbf{x}_k\}_{k=1}^{N}$, where $\mathbf{x}_k \in \mathbb{R}^{d_0}$, we apply Ball k-means to initialize N_G GBs with nonempty sample index sets $\{\mathcal{I}_i\}_{i=1}^{N_G}$. For the i-th ball, let $n_i = |\mathcal{I}_i|$. Its center and radius are $\mathbf{c}_i = n_i^{-1} \sum_{k \in \mathcal{I}_i} \mathbf{x}_k$ and $r_i = \max_{k \in \mathcal{I}_i} \|\mathbf{x}_k - \mathbf{c}_i\|_2$, respectively. These initial statistics summarize the location and spatial extent of each local sample group.

To additionally describe the initial within-ball dispersion, we compute the mean scatter $\sigma_i = n_i^{-1} \sum_{k \in \mathcal{I}_i} \|\mathbf{x}_k - \mathbf{c}_i\|_2$ and map it to a bounded intrinsic hesitation descriptor. The descriptor is defined as follows:

$$\eta_i = \frac{\sigma_i}{\sigma_i + \tau}, \qquad \tau > 0. \tag{1}$$

Here, $\eta_i \in [0, 1)$ increases with intra-ball scatter, and τ sets the dispersion scale of the mapping.

Concatenating these quantities gives the initial structured descriptor $\mathbf{s}_i = [\mathbf{c}_i^\top, r_i, \eta_i]^\top \in \mathbb{R}^{d_0+2}$. These descriptors help preserve local geometric structure and within-ball heterogeneity during subsequent latent representation learning. The dynamic GB reconstruction procedure is detailed in Appendix B.

3.2 GRANULAR-BALL AUTOENCODING

To the best of our knowledge, we are the first to design a GB-level autoencoder for learning deep representations that can adapt to the uncertainty-aware clustering objective. Specifically, each structured descriptor is encoded by f_ϕ and reconstructed by g_ψ. Their parameters are denoted by $\Theta_{\mathrm{AE}} = \{\phi, \psi\}$, where $\phi = \{\mathbf{W}^{(m)}, \mathbf{b}^{(m)}\}_{m=1}^{M_b}$ contains the encoder weights and biases.

The encoder produces $\mathbf{z}_i = f_\phi(\mathbf{s}_i)$, with $\mathbf{z}_i = [(\mathbf{c}_i^{(b)})^\top, r_i^{(b)}, \eta_i^{(b)}]^\top$, where the superscript (b) denotes the bottleneck layer. The decoder returns $\widehat{\mathbf{s}}_i = g_\psi(\mathbf{z}_i)$, with $\widehat{\mathbf{s}}_i = [\widehat{\mathbf{c}}_i^\top, \widehat{r}_i, \widehat{\eta}_i]^\top$.

The embedding $\mathbf{c}_i^{(b)} \in \mathbb{R}^{d_b}$ is used for clustering. The auxiliary channels $r_i^{(b)}$ and $\eta_i^{(b)}$ are learned coordinates associated with scale and intrinsic hesitation. Their ranges, $r_i^{(b)} \geq 0$ and $0 \leq \eta_i^{(b)} \leq 1$, are enforced by the output parameterization. In particular, $r_i^{(b)}$ serves as a learned scale variable in the distribution loss introduced below; it is not a radius recomputed from individually encoded samples. The reconstruction objective is defined as:

$$\mathcal{J}_{\mathrm{rec}} = \frac{1}{2} \sum_{i=1}^{N_G} \|\widehat{\mathbf{c}}_i - \mathbf{c}_i\|_2^2 + \frac{1}{2} \sum_{i=1}^{N_G} \left[(\widehat{r}_i - r_i)^2 + (\widehat{\eta}_i - \eta_i)^2 \right]. \tag{2}$$

Reconstructing these descriptors encourages the latent representations to preserve information about the location, spatial extent, and within-ball dispersion of each GB.

3.3 INTUITIONISTIC TRIPLET EVIDENCE

GRANITE introduces an intuitionistic fuzzy evidence triplet for each ball–cluster pair, thereby explicitly characterizing uncertainty in ball–cluster assignments. Let $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_K] \in \mathbb{R}^{d_b \times K}$

contain the cluster prototypes. For ball i and cluster j, define $\mathbf{p}_{ij} = (\mu_{ij}, \nu_{ij}, \pi_{ij})^\top$, where the membership and non-membership degrees satisfy $\mu_{ij} \geq 0$, $\nu_{ij} \geq 0$, and $\mu_{ij} + \nu_{ij} \leq 1$. The remaining evidence is represented by $\pi_{ij} = 1 - \mu_{ij} - \nu_{ij}$.

Each triplet therefore belongs to a three-component simplex. Normalization applies within each ball–cluster pair, rather than across the K clusters. The descriptor η_i summarizes within-ball dispersion, whereas the trainable variable π_{ij} represents hesitation with respect to a particular cluster.

We combine membership and non-membership through the signed hesitation evidence coefficient $w_{ij} = \mu_{ij} - \nu_{ij}$. As illustrated in Fig. 2, positive and negative coefficients indicate dominant membership and non-membership evidence, respectively. A zero coefficient indicates equal membership and non-membership.

Let $\mathbf{w}_i = (w_{i1}, \ldots, w_{iK})^\top$. The coefficients define a signed prototype representation given by $\mathbf{H}\mathbf{w}_i = \sum_{j=1}^{K} w_{ij}\mathbf{h}_j$. The residual is then defined as:

$$\mathbf{v}_i = \mathbf{c}_i^{(b)} - \mathbf{H}\mathbf{w}_i. \tag{3}$$

This residual couples the embedding to all prototype coefficients jointly.

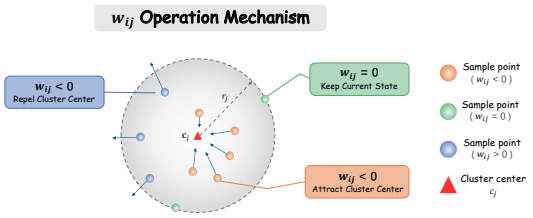


Figure 2: Signed hesitation evidence for a granular ball–cluster pair.

Because the residual depends on the signed evidence coefficient w_{ij}; we additionally regularize all three triplet components with a negative-entropy term, defined as follows:

$$\mathcal{R}_{\mathrm{NE}} = \sum_{i=1}^{N_G}\sum_{j=1}^{K}\sum_{\ell=1}^{3} p_{ij\ell}\log p_{ij\ell}, \tag{4}$$

where $p_{ij\ell}$ denotes the ℓ-th component of $\mathbf{p}_{ij}$ and $0\log 0 := 0$ by continuity. This intuitionistic fuzzy evidence triplet, comprising membership, non-membership, and cluster-specific hesitation degrees, explicitly represents support, rejection, and unresolved uncertainty in ball–cluster relationships.

3.4 Covariance-Aware Granular-Ball Distribution Regularization

To enable uncertainty-aware clustering, we evaluate the residuals using a shared covariance metric and a smooth robust penalty, and modulate the contribution of each GB through its learned scale channel. Let $\mathbf{\Sigma} \in \mathbb{R}^{d_b \times d_b}$ be a learned covariance metric satisfying $\mathbf{\Sigma} \succeq \delta \mathbf{I}_{d_b}$ for a fixed $\delta > 0$. Define the squared Mahalanobis residual as $q_i = \mathbf{v}_i^\top \mathbf{\Sigma}^{-1}\mathbf{v}_i$. The robust penalty is defined as:

$$\rho_{\varepsilon,\mathbf{\Sigma}}(\mathbf{v}_i) = (1+\varepsilon)\frac{q_i}{\sqrt{q_i + \varepsilon^2}}, \qquad \varepsilon > 0. \tag{5}$$

As ε approaches 0, the penalty approaches the Mahalanobis distance $\sqrt{q_i}$. For fixed ε, it is smooth near zero and asymptotically linear in $\sqrt{q_i}$. This growth moderates the contribution of large residuals relative to a squared-distance penalty.

To incorporate the learned scale channel, we use $\omega_i = \exp[-(r_i^{(b)})^2/\tau_r^2]$, where $\tau_r > 0$. Larger values of $r_i^{(b)}$ yield smaller weights. Since the scale channel is produced by the encoder, the weights adapt with the latent representations and retain their dependence on Θ_{AE} during optimization.

The resulting distribution loss is defined as:

$$\mathcal{L}_{\mathrm{dist}} = \sum_{i=1}^{N_G} \omega_i \rho_{\varepsilon,\mathbf{\Sigma}}(\mathbf{v}_i) + \frac{\gamma_\Sigma}{2}\log\det\mathbf{\Sigma}, \tag{6}$$

where $\gamma_\Sigma > 0$. The log-determinant term penalizes large covariance volume, while the eigenvalue floor prevents the metric from collapsing along individual directions. The adaptive distribution loss $\mathcal{L}_{\mathrm{dist}}$ improves both the robustness and efficiency of clustering.

3.5 Joint Learning Objective

We propose a new GB clustering paradigm that jointly optimizes reconstruction and clustering objectives. To control the scale of the cluster prototypes, we introduce a squared Frobenius regularizer. Accordingly, the clustering objective is formulated as:

$$\mathcal{J}_{\text{clu}} = \mathcal{L}_{\text{dist}} + \lambda_1 \mathcal{R}_{\text{NE}} + \frac{\lambda_H}{2}\|\mathbf{H}\|_F^2, \tag{7}$$

where $\lambda_1 > 0$ and $\lambda_H > 0$. The prototype regularizer penalizes unbounded rescaling of $\mathbf{H}$, including rescaling that leaves the reconstruction residuals unchanged.

Let $\Theta = \{\Theta_{\text{AE}}, \mathbf{H}, \boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\Sigma}\}$ denote the trainable variables. The joint optimization problem is formulated as follows:

$$\begin{aligned} \min_{\Theta} \quad & \mathcal{J}(\Theta) = \mathcal{J}_{\text{rec}} + \mathcal{J}_{\text{clu}} + \frac{\lambda_2}{2}\sum_{m=1}^{M_b}\left(\|\mathbf{W}^{(m)}\|_F^2 + \|\mathbf{b}^{(m)}\|_2^2\right), \\ & \text{s.t.} \quad \mu_{ij} \geq 0,\ \nu_{ij} \geq 0,\ \mu_{ij} + \nu_{ij} \leq 1,\ \boldsymbol{\Sigma} \succeq \delta \mathbf{I}_{d_b}. \end{aligned} \tag{8}$$

The final term applies weight decay to the encoder, with $\lambda_2 \geq 0$. The range constraints on the auxiliary latent channels are enforced by the network parameterization. Both the residuals and their scale-dependent weights depend on the encoder. Consequently, the encoder receives training signals from reconstruction and clustering, while the decoder is trained through reconstruction. All hyperparameters remain fixed during the optimization analyzed below. GRANITE alleviates the limitations caused by static feature representations and insufficient use of uncertainty, providing a new perspective for uncertainty-aware GB clustering.

4 Optimization and Convergence Analysis

4.1 Optimization

Theorem 1. *For $\varepsilon > 0$ and fixed nonnegative coefficients, minimizing the tight quadratic upper bound of the adaptive residual objective is equivalent to minimizing the unchanged remainder plus a weighted sum of squared Euclidean residual norms, with weights evaluated at the current iterate.*

Since covariance-aware GB distribution loss in equation 8 is difficult to optimize directly within the objective, we apply Theorem 1 to construct its tight iterative surrogate. The resulting surrogate for the distribution loss is defined as follows:

$$\widetilde{\mathcal{L}}_{\text{dist}}^{(t)}(\Theta) = \sum_{i=1}^{N_G} \omega_i(\Theta_{\text{AE}}) \left[d_i^{(t)} q_i(\Theta) + b_i^{(t)}\right] + \frac{\gamma_\Sigma}{2}\log\det\boldsymbol{\Sigma}. \tag{9}$$

The proof of Theorem 1 and its application to GRANITE are provided in Appendices C.1 and C.2.

Prototype update. To guiding the joint learning of GB representations and cluster assignments, let $\alpha_i = \omega_i d_i^{(t)}$, $\mathbf{M} = \sum_i \alpha_i \mathbf{w}_i \mathbf{w}_i^\top$, and $\mathbf{B} = \sum_i \alpha_i \mathbf{c}_i^{(b)} \mathbf{w}_i^\top$. The prototype update is as follows:

$$\mathbf{H}^+ = \mathbf{U}_\Sigma \mathbf{Y}^+ \mathbf{U}_M^\top, \qquad Y_{ab}^+ = \frac{(\mathbf{U}_\Sigma^\top \mathbf{B} \mathbf{U}_M)_{ab}}{m_b + \lambda_H \ell_a/2}. \tag{10}$$

The derivation is provided in Appendix C.3.

Assignment update. We update the triplets to achieve uncertainty-aware ball–cluster assignments. Starting from strictly positive triplets, let k index the inner assignment iterations and define $u_{ij}^{(k)} = 2\alpha_i \mathbf{h}_j^\top \boldsymbol{\Sigma}^{-1} \mathbf{v}_i^{(k)}$. The triplets are updated as follows:

$$z_{ijr}^{(k)} = \frac{\log p_{ijr}^{(k)} - \beta_i g_{ijr}^{(k)}}{1 + \beta_i \lambda_1}, \quad p_{ijr}^{(k+1)} = \frac{\exp z_{ijr}^{(k)}}{\sum_{s=1}^{3} \exp z_{ijs}^{(k)}}. \tag{11}$$

The derivation and descent property are given in Appendix C.4.

Covariance update. We update the covariance matrix to adapt the clustering metric to directional scale differences and correlations in the residuals. Let $\mathbf{S} = \sum_i \alpha_i \mathbf{v}_i \mathbf{v}_i^\top$. The covariance update is as follows:

$$\mathbf{\Sigma}^+ = \mathbf{U}\,\mathrm{diag}(\ell_1^+, \ldots, \ell_{d_b}^+)\mathbf{U}^\top, \qquad \ell_k^+ = \max\left\{\delta, \frac{2s_k}{\gamma_\Sigma}\right\}. \tag{12}$$

The derivation is provided in Appendix C.5.

Autoencoder update. Let F_t denote the network-dependent part of $\mathcal{Q}_t$, including the weighted intercepts $\sum_i \omega_i b_i^{(t)}$. The autoencoder parameters are updated as follows:

$$\Theta_{\mathrm{AE}}^+ = \Theta_{\mathrm{AE}} - \alpha_t \nabla_{\Theta_{\mathrm{AE}}} F_t(\Theta_{\mathrm{AE}}). \tag{13}$$

The gradient and step-size details are given in Appendix C.6.

4.2 Convergence Analysis

Theorem 2. *If the objective value sequence $\{\mathcal{J}^t\}_{t\geq 0}$ satisfies the following two conditions:*

1. $\mathcal{J}^{t+1} \leq \mathcal{J}^t, \quad \forall t \geq 0$;
2. $\exists M > 0$ *such that* $|\mathcal{J}^t| \leq M, \quad \forall t \geq 0$,

then it converges to a finite limit.

For our objective, the outer update stage satisfies the following inequality:

$$\mathcal{J}(\Theta^{(t+1)}) \leq \mathcal{J}(\Theta^{(t)}), \tag{14}$$

which implies monotonicity; the detailed proof is given in Appendix D.1. Moreover, under the constraints in equation 8, the stage-wise objective values satisfy the following lower bound:

$$\mathcal{J}(\Theta^{(0)}) \geq \mathcal{J}(\Theta^{(t)}) \geq \frac{\gamma_\Sigma d_b}{2}\log\delta - \lambda_1 N_G K \log 3. \tag{15}$$

with the detailed proof provided in Appendix D.2. Therefore, we conclude that $\mathcal{J}(\Theta)$ is amenable to iterative optimization.

5 Experimental Analysis and Results

Detailed experimental settings are provided in Appendix E.1. We conducted clustering comparisons, presented visualizations of the results and training dynamics, and performed ablation studies and parameter sensitivity analyses. To evaluate the generalization ability of the algorithm, we additionally conducted experiments on image segmentation. The time complexity analysis is provided in Appendix E.5.

5.1 Clustering Scenarios

Table 1 compares GRANITE with traditional, deep, and GB-based clustering methods on eight UCI datasets. GRANITE achieves the highest reported ACC on all eight datasets and leads in ACC, NMI, and ARI simultaneously on Iris, Seed, and WDBC. On WDBC, it improves ACC from 0.905 to 0.951 and ARI from 0.654 to 0.812 over Ball k-means, corresponding to relative gains of 5.1% and 24.2%, respectively. These results suggest that combining adaptive latent representations, multi-scale granular structure, and signed intuitionistic evidence can improve clustering robustness, particularly on datasets with complex or ambiguous distributions. Fig. 3(a) further shows that GRANITE learns more compact and better-separated latent clusters than representative baselines. The improved separation near cluster boundaries is consistent with the quantitative results in Table 1, supporting the effectiveness of the proposed uncertainty-aware granular representation. Additional t-SNE visualizations and clustering analyses are provided in Appendix E.2.

Table 1: Quantitative results on eight datasets for traditional clustering methods ($\mathcal{M}_\mathcal{T}$), deep clustering methods ($\mathcal{M}_\mathcal{D}$) and granular-ball clustering methods ($\mathcal{M}_\mathcal{G}$). **Best** and suboptimal results.

	Methods	Venue	Iris			Wine			Seed			Glass		
			ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
$\mathcal{M}_\mathcal{T}$	k-means	-	0.886	0.741	0.716	0.573	0.423	0.351	0.895	0.694	0.716	0.542	**0.426**	0.273
	DPHKM	*ICML'23*	0.660	0.465	0.410	0.534	0.212	0.137	0.586	0.209	0.211	0.383	0.206	0.114
	SBSC	*SIGMOD'25*	0.900	0.728	0.737	0.719	0.635	0.543	0.895	0.682	0.720	0.519	0.362	0.237
$\mathcal{M}_\mathcal{D}$	DeepCluster	*ECCV'18*	0.680	0.552	0.476	0.579	0.410	0.326	0.676	0.415	0.385	0.341	0.259	0.124
	TableDC	*SIGMOD'25*	0.547	0.143	0.089	0.388	0.013	0.004	0.533	0.251	0.185	0.472	0.286	0.177
	SpEx	*NeurIPS'25*	0.893	0.791	0.732	0.635	0.562	0.442	0.867	0.629	0.651	0.407	0.306	0.146
	MiniClustering	*ICLR'26*	0.833	0.704	0.631	0.635	0.625	0.473	0.886	0.675	0.701	0.481	0.357	0.188
$\mathcal{M}_\mathcal{G}$	Ball k-means	*TPAMI'22*	0.853	0.661	0.645	0.960	**0.861**	0.880	0.919	0.728	0.773	0.495	0.335	0.173
	MGNR	*TPAMI'24*	0.440	0.245	0.082	0.382	0.027	-0.005	0.471	0.346	0.193	0.463	0.235	0.107
	GBCT	*TNNLS'25*	0.740	0.704	0.564	0.893	0.740	0.703	0.505	0.306	0.161	0.411	0.294	0.133
	GBK-DPC	*PR'26*	0.960	0.870	0.885	0.719	0.412	0.405	0.919	0.731	0.771	0.584	0.394	0.267
	GRANITE	-	**0.973**	**0.898**	**0.922**	**0.961**	0.852	**0.882**	**0.929**	**0.752**	**0.799**	**0.589**	0.361	**0.283**

	Methods	Venue	WDBC			Segment			Landsat			Pendigits		
			ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
$\mathcal{M}_\mathcal{T}$	k-means	-	0.854	0.464	0.491	0.546	0.584	0.461	0.512	0.473	0.325	0.665	0.679	0.531
	DPHKM	*ICML'23*	0.677	0.101	0.123	0.479	0.466	0.367	0.501	0.332	0.244	0.337	0.321	0.185
	SBSC	*SIGMOD'25*	0.636	0.025	0.012	0.412	0.315	0.207	0.717	0.612	0.562	0.651	0.704	0.530
	SpEx	*NeurIPS'25*	0.631	0.010	0.005	0.143	0.001	0.000	0.242	0.012	0.002	0.453	0.552	0.340
$\mathcal{M}_\mathcal{D}$	DeepCluster	*ECCV'18*	0.671	0.230	0.108	0.635	0.602	0.521	0.614	0.474	0.446	0.696	0.682	0.561
	TableDC	*SIGMOD'25*	0.550	0.001	0.004	0.385	0.295	0.168	0.241	0.026	0.020	0.616	0.539	0.453
	MiniClustering	*ICLR'26*	0.852	0.461	0.486	0.433	0.487	0.289	0.480	0.430	0.311	0.661	0.681	0.528
$\mathcal{M}_\mathcal{G}$	Ball k-means	*TPAMI'22*	0.905	0.532	0.654	0.569	0.561	0.451	0.678	**0.613**	0.530	0.708	0.683	**0.588**
	MGNR	*TPAMI'24*	0.652	0.064	0.036	0.404	0.363	0.206	0.511	0.406	0.303	0.251	0.366	0.141
	GBCT	*TNNLS'25*	0.578	0.055	-0.030	0.535	0.573	0.375	0.260	0.125	0.029	0.138	0.075	0.001
	GBK-DPC	*PR'26*	0.887	0.500	0.595	0.672	**0.684**	0.544	0.710	0.592	0.542	0.369	0.543	0.318
	GRANITE	-	**0.951**	**0.706**	**0.812**	**0.673**	0.648	**0.552**	**0.744**	0.591	**0.567**	**0.709**	**0.705**	0.573

5.2 Process Analysis and Visualization

We further analyze the training process through loss curves and t-SNE visualizations. Fig. 3(b) presents the total, reconstruction, and clustering loss curves of GRANITE on Landsat and Pendigits, together with the learned representations at four training stages. During early training, the clustering loss decreases rapidly, while the reconstruction loss decreases more slowly. As training progresses, the reconstruction loss drops substantially, and all three losses gradually stabilize. These results illustrate the training behavior on both datasets, providing insight into the optimization process and the evolution of the learned representations. Training loss curves for the remaining datasets and further analysis are provided in Appendix E.3.

The t-SNE snapshots further illustrate how the latent representations evolve during training. At the initial stage, the embeddings are scattered and cluster boundaries are ambiguous. As training proceeds, the embeddings become progressively more compact and better separated, and the separation around ambiguous boundaries improves markedly at the converged stage. These observations suggest that the joint reconstruction and clustering objectives gradually shape a more discriminative latent representation while preserving the structural constraints of the intuitionistic GBs.

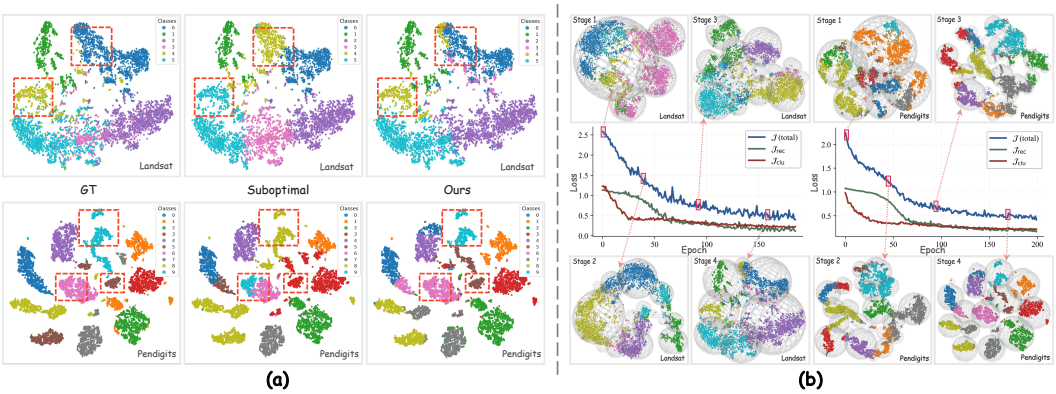


Figure 3: (a) Qualitative comparison between GRANITE and the best-performing baseline using t-SNE. (b) Training loss curves and t-SNE visualizations at different training stages.

5.3 Ablation Study

To quantify the contributions of the main components of GRANITE, we conduct ablation experiments. Table 2 reports the effects of the reconstruction objective, the clustering objective, and the intuitionistic fuzzy formulation on the Segment dataset. Removing either objective degrades performance: in the IFS configuration, removing reconstruction drops ACC/NMI from $67.27\%/64.77\%$ to $59.22\%/53.57\%$, and removing clustering reduces them to $62.85\%/55.14\%$. Using both clearly outperforms either alone, showing that reconstruction preserves multi-scale granular structure while clustering drives class-discriminative organization. Replacing IFS with conventional FS also lowers ACC and NMI, with a larger drop in NMI, indicating that non-membership and hesitation provide additional evidence for ambiguous assignments. Notably, when reconstruction is removed, the FS and IFS results become nearly identical, suggesting that the benefit of intuitionistic modeling is coupled with the structure-aware reconstruction objective.

TABLE 2. Ablation study on $\mathcal{J}_{\text{recon}}$ and $\mathcal{J}_{\text{clust}}$ under FS/IFS.

Methods	IFS/FS	$\mathcal{J}_{\text{recon}}$	$\mathcal{J}_{\text{clust}}$	*Segment* ACC(%)↑	*Segment* NMI(%)↑
Traditional Granular-balls					
w/o $\mathcal{J}_{\text{recon}}$ and $\mathcal{J}_{\text{clust}}$	-	✗	✗	53.78	57.34
w/o $\mathcal{J}_{\text{recon}}$	-	✗	✓	59.22	53.60
w/o $\mathcal{J}_{\text{clust}}$	-	✓	✗	60.24	53.82
w/ $\mathcal{J}_{\text{recon}}$ and $\mathcal{J}_{\text{clust}}$	-	✓	✓	<u>65.31</u>	<u>59.24</u>
Intuitionistic Fuzzy Granular-balls					
w/o $\mathcal{J}_{\text{recon}}$	IFS	✗	✓	59.22	53.57
w/o $\mathcal{J}_{\text{clust}}$	IFS	✓	✗	62.85	55.14
*IFS+w/ $\mathcal{J}_{\text{recon}}$ and $\mathcal{J}_{\text{clust}}$	IFS	✓	✓	**67.27**	**64.77**

Note: FS/IFS: traditional/intuitionistic fuzzy sets.

TABLE 3. Quantitative analysis results on image segmentation. ACC(%), mIoU(%), NMI(%).

Methods	BSD 500			DUTS		
	ACC	mIoU	NMI	ACC	mIoU	NMI
DFKM *TFS'19*	69.98	48.52	67.20	72.33	58.51	76.82
FLRSC *TFS'23*	75.58	62.55	76.63	77.69	65.08	76.87
MK-SPCM *TFS'24*	<u>78.86</u>	67.34	<u>77.14</u>	76.93	65.96	73.29
SpEx *NeurIPS'25*	71.34	62.67	51.96	78.41	<u>67.80</u>	55.38
GBCT *TNNLS'25*	76.66	67.46	63.61	73.59	53.76	61.97
AWRFCM *TFS'26*	75.39	<u>70.07</u>	69.75	72.14	57.82	75.08
SegGBC *CVPR'26*	77.32	69.95	71.84	<u>80.54</u>	65.77	<u>76.90</u>
GRANITE	**83.20**	**73.74**	**81.53**	**82.41**	**71.22**	**77.54**

Fig. 4 compares the proposed covariance-aware GB distribution loss with four alternative formulations (L_1, squared L_2, $L_{2,1}$, and $L_{1,2}$) on eight datasets in terms of NMI. Each variant replaces only the residual penalty in the distribution loss, while keeping all other model components unchanged. GRANITE achieves the highest NMI across all datasets, supporting the effectiveness of the residual penalty designed for the proposed covariance-aware GB distribution loss.

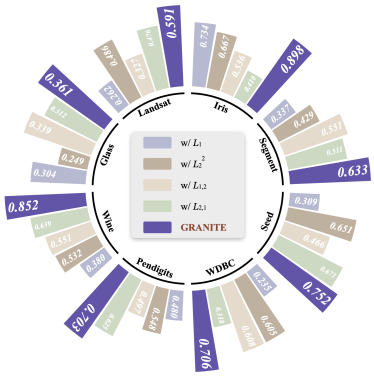


Figure 4: Ablation study across different paradigms.

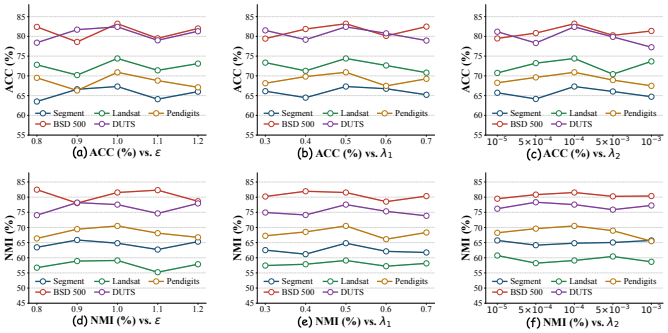


Figure 5: Sensitivity analysis of the parameters ε, λ_1, and λ_2.

5.4 Parameter Sensitivity Analysis

In addition, we further examine the sensitivity of GRANITE to the smoothing parameter ε and the regularization coefficients λ_1 and λ_2, as shown in Fig. 5. Here, ε controls the smoothness of the covariance-aware distribution loss and its transition from quadratic growth near zero to linear growth for large residuals. The coefficients λ_1 and λ_2 control the strengths of negative-entropy regularization and ℓ_2 weight regularization, respectively. Performance remains generally stable across the tested parameter ranges, with only minor fluctuations, indicating that GRANITE is relatively robust to these hyperparameters within the evaluated ranges.

5.5 Image Segmentation Performance

To demonstrate the generalization ability of GRANITE, we apply it to image segmentation and also achieve strong performance. Table 3 reports the results on BSD500 and DUTS, where GRANITE achieves the best overall performance on both datasets, with particularly consistent gains in mIoU. Compared with structural methods such as SegGBC, SpEx, and FLRSC, GRANITE better adapts the representation and jointly models granular geometry and intuitionistic fuzzy uncertainty.

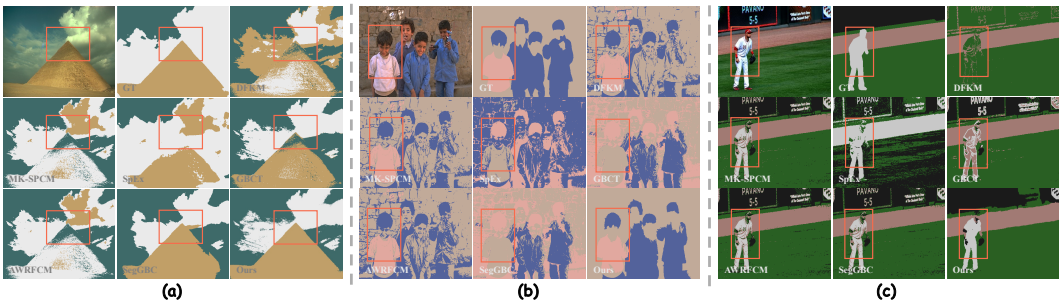


Figure 6: Qualitative comparison of representative segmentation results on BSD500 and DUTS.

Fig. 6 presents a qualitative comparison of segmentation results on representative images. Compared with competing methods, GRANITE produces boundaries that align more closely with the ground truth and better preserves fine structures in complex regions. Particularly around object edges and in regions with ambiguous boundaries, GRANITE preserves boundaries more effectively, reducing over-smoothing and boundary leakage. These observations highlight the benefits of uncertainty-aware modeling in handling ambiguous regions. The qualitative results are consistent with the quantitative performance reported in Table 3.

6 Failure Cases and Limitations



Figure 7: Failure cases for image segmentation.

Although GRANITE performs well in terms of efficiency and accuracy, it still has certain limitations. As shown in Fig. 7, part of the person highlighted by the yellow box is incorrectly assigned to the background, while the tree and grass regions marked by the red boxes contain scattered segmentation fragments. These observations suggest that, as a clustering framework, GRANITE does not fully capture the spatial relationships and boundary information required for image segmentation. When different regions share similar features, the model may confuse their assignments; when features vary substantially within the same object, it may generate many small granular balls, increasing the risk of fragmented segmentation.

7 Conclusion

In this paper, we propose GRANITE, an algorithm that integrates intuitionistic fuzzy modeling into a ball-level autoencoding framework. By jointly optimizing the reconstruction loss and an uncertainty-aware clustering objective, GRANITE learns latent representations that account for both GB structural information and assignment uncertainty, thereby establishing a new GB clustering paradigm. Compared with traditional GB clustering methods, this design alleviates the limitations imposed by fixed feature and insufficient modeling of assignment uncertainty, improving clustering robustness and generalization.

Future work will focus on efficient GB construction for high-resolution images, spatial and boundary-aware regularization, and extensions to medical, remote-sensing, multi-modal, and video segmentation. Automatic estimation of the number of regions and tighter integration with pixel-level segmentation objectives also remain promising directions.

AI USE STATEMENT

In this work, we used generative AI tools solely for language polishing and editing of the manuscript, including suggesting optimized phrasing and correcting grammatical and spelling errors.

We have not used generative AI tools for research ideation, hypothesis formulation, determination of research directions, methodological design, or experimental design, nor have we used them for decisions concerning data collection and analysis. The remaining disclosure items are not applicable to this work.

Furthermore, all AI-assisted content has been reviewed and revised by the authors to ensure technical accuracy and consistency with the authors' intended meaning. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we have released the source code in an anonymous repository at *https://anonymous.4open.science/r/111111-3A89*. The repository contains a complete and executable dynamic GB algorithm, as well as complete pseudocode from raw data to final labels, clearly specifying the inputs, outputs, parameter inheritance, and label generation formulas for each reconstruction. The appendix states the assumptions underlying our theoretical results and provides complete proofs of both theorems. It further presents detailed derivations of the parameter updates, algorithm pseudocode, and the acceptance and stopping criteria. To support experimental replication, the appendix documents the datasets, evaluation metrics, compared methods, and implementation settings. Additional clustering visualizations, training curves, image segmentation results, and computational complexity analysis complement the main-text results and facilitate further assessment of the proposed method.

REFERENCES

Tal Argov and Tal Wagner. SpEx: A spectral approach to explainable clustering. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*, pp. 22578–22605, 2025.

Krassimir T Atanassov. More on intuitionistic fuzzy sets. *Fuzzy Sets and Systems*, 33(1):37–45, 1989.

Kamal Berahmand, Sondos Bahadori, Maryam Nooraei Abadeh, Yuefeng Li, and Yue Xu. SDAC-DA: Semi-supervised deep attributed clustering using dual autoencoder. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):6989–7002, 2024.

James C Bezdek, Robert Ehrlich, and William Full. FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, 10(2):191–203, 1984.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 132–149, 2018.

Tamalika Chaira. A novel intuitionistic fuzzy c means clustering algorithm and its application to medical images. *Applied Soft Computing*, 11(2):1711–1717, 2011.

Jie Chen, Zhen Wu, Jian Cheng, Xin Xu, and Feiping Nie. Graph clustering with harmonic-maxmin cut guidance. *IEEE Transactions on Knowledge and Data Engineering*, 37(5):2600–2613, 2025.

Qianpeng Chong, Wenyi Zeng, Xiuxuan Shen, Jiajie Li, Qian Yin, and Xin Zheng. SegGBC: Justifiable coarse-to-fine granular-ball computing for enhancing clustering image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pp. 42104–42114, 2026.

Yuhang Ding, Liulei Li, Wenguan Wang, and Yi Yang. Clustering propagation for universal medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3357–3369, 2024.

Jiangxiong Fang, Huaxiang Liu, Liting Zhang, Jun Liu, and Hesheng Liu. Region-edge-based active contours driven by hybrid and local fuzzy region-based energy for image segmentation. *Information Sciences*, 546:397–419, 2021.

Jacob Imola, Alessandro Epasto, Mohammad Mahdian, Vincent Cohen-Addad, and Vahab Mirrokni. Differentially private hierarchical clustering with provable approximation guarantees. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 14353–14375, 2023.

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

Junfeng Jing, Siyuan Qu, Cong Wang, and Xuelong Li. Adaptively weighted residual-driven fuzzy c-means for image segmentation in presence of mixed or unknown noise. *IEEE Transactions on Fuzzy Systems*, 34(4):1307–1320, 2026.

Abdul Atif Khan, Rashmi Maheshwari, Mohammad Maksood Akhter, and Sraban Kumar Mohanty. SBSC: A fast self-tuned bipartite proximity graph-based spectral clustering. *Proceedings of the ACM on Management of Data*, 3(3):213:1–213:22, 2025.

Wentao Li, Lingwei Wei, Witold Pedrycz, Weiping Ding, Chao Zhang, Tao Zhan, and Shuyin Xia. Granular-ball regeneration clustering with principle of justifiable granularity. *IEEE Transactions on Neural Networks and Learning Systems*, 36(10):18173–18187, 2025.

Zhixin Li, Yuheng Jia, Guanliang Chen, Hui Liu, and Junhui Hou. Mini-cluster guided long-tailed deep clustering. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pp. 79896–79916, 2026.

Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28 (2):129–137, 1982.

Lukas Miklautz, Timo Klein, Kevin Sidak, Collin Leiber, Thomas Lang, Andrii Shkabrii, Sebastian Tschiatschek, and Claudia Plant. Breaking the reclustering barrier in centroid-based deep clustering. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pp. 100295–100335, 2025.

Yongting Ni, Jin Qian, Enhong Chen, Shaowei Yan, and Jiangang Ye. GBK-DPC: Density peak clustering based on granular ball with k-nearest neighbor. *Pattern Recognition*, 171:112243, 2026.

Feiping Nie, Hua Wang, Heng Huang, Chris HQ Ding, et al. Adaptive loss minimization for semi-supervised elastic embedding. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1565–1571, 2013.

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

Hafiz Tayyab Rauf, André Freitas, and Norman W. Paton. TableDC: Deep clustering for tabular data. *Proceedings of the ACM on Management of Data*, 3(3):229:1–229:28, 2025.

Yazhou Ren, Jingyu Pu, Zhimeng Yang, Jie Xu, Guofeng Li, Xiaorong Pu, Philip S. Yu, and Lifang He. Deep clustering: A comprehensive survey. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4):5858–5878, 2025.

Sensen Song, Zhenhong Jia, Jie Yang, and Nikola Kasabov. Image segmentation based on fuzzy low-rank structural clustering. *IEEE Transactions on Fuzzy Systems*, 31(7):2153–2166, 2023.

Cong Wang, Shengnan Jiang, Yuan Yuan, Junfeng Jing, MengChu Zhou, and Witold Pedrycz. PRFCM: Poisson-specific residual-driven fuzzy c-means clustering for image segmentation. *IEEE Transactions on Fuzzy Systems*, 34(6):1869–1883, 2026.

Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pp. 136–145, 2017.

Kongyuan Wei, Rongzhen Zhao, Haixia Kou, Pengfei Chen, Yongyong Cao, Yuqiao Zheng, and Linfeng Deng. Dimensionality reduction of rolling bearing fault data based on graph-embedded semi-supervised deep auto-encoders. *Engineering Applications of Artificial Intelligence*, 152: 110689, 2025.

Shuyin Xia, Daowan Peng, Deyu Meng, Changqing Zhang, Guoyin Wang, Elisabeth Giem, Wei Wei, and Zizhong Chen. Ball k-means: Fast adaptive clustering with no bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):87–99, 2022.

Shuyin Xia, Xiaochuan Dai, Guoyin Wang, Xinbo Gao, and Elisabeth Giem. An efficient and adaptive granular-ball generation method in classification problem. *IEEE Transactions on Neural Networks and Learning Systems*, 35(4):5319–5331, 2024.

Shuyin Xia, Bolun Shi, Yifan Wang, Jiang Xie, Guoyin Wang, and Xinbo Gao. GBCT: Efficient and adaptive clustering via granular-ball computing for complex data. *IEEE Transactions on Neural Networks and Learning Systems*, 36(7):12159–12172, 2025a.

Shuyin Xia, Yifan Wang, Lifeng Shen, and Guoyin Wang. Granular-ball-induced multiple kernel k-means. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 6633–6641, 2025b.

Jiang Xie, Xuexin Xiang, Shuyin Xia, Lian Jiang, Guoyin Wang, and Xinbo Gao. MGNR: A multi-granularity neighbor relationship and its application in knn classification and clustering methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):7956–7972, 2024.

Xu Yang, Cheng Deng, Kun Wei, Junchi Yan, and Wei Liu. Adversarial learning for robust deep clustering. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*, pp. 9098–9108, 2020.

Haiyan Yu, Shuang Xie, Jiulun Fan, Rong Lan, and Bo Lei. Mahalanobis-kernel distance-based suppressed possibilistic c-means clustering algorithm for imbalanced image segmentation. *IEEE Transactions on Fuzzy Systems*, 32(8):4595–4609, 2024.

Rui Zhang, Xuelong Li, Hongyuan Zhang, and Feiping Nie. Deep fuzzy k-means with adaptive loss and entropy regularization. *IEEE Transactions on Fuzzy Systems*, 28(11):2814–2824, 2019.

Appendix

A Detailed Preliminaries **14**

A.1 Intuitionistic Fuzzy Set 14

A.2 Granular-Ball Computing 14

A.3 Deep Autoencoder 15

B Dynamic Granular-Ball Reconstruction and Label Propagation **15**

B.1 Initialization 15

B.2 Single-Point Encoding and Membership Reassignment 15

B.3 Descriptor Recomputation 16

B.4 Triplet Inheritance via Membership Overlap 16

B.5 Variable Inheritance and Acceptance 16

B.6 Stopping Budget 16

B.7 Final Label Propagation 16

C Optimization Details **16**

C.1 Proof of Theorem 1 17

C.2 Application to GRANITE 17

C.3 Update of the Cluster Prototypes 18

C.4 Update of the Intuitionistic Assignments 18

C.5 Update of the Covariance Metric 19

C.6 Update of the Autoencoder Parameters 20

C.7 Complete Outer Iteration 20

C.8 Acceptance and Stopping Criteria 20

D Convergence Analysis **21**

D.1 Monotonic Objective Descent 21

D.2 Convergence of Objective Values 22

D.3 Role of the Covariance Eigenvalue Floor 23

D.4 Bounds on the Objective Sublevel Set 24

D.5 Training-Order Matching for Convergence 24

E Additional Experiments and Analysis **24**

E.1 Experimental Setting 24

E.2 Clustering Experiment 25

E.3 Analysis of the Convergence Process 25

E.4 Image Segmentation Experiment 27

E.5 Time Complexity Analysis 28

A Detailed Preliminaries

This appendix summarizes the concepts needed for GRANITE and clarifies their roles in the model.

A.1 Intuitionistic Fuzzy Set

An intuitionistic fuzzy set extends a traditional fuzzy set by assigning both membership and non-membership degrees, with the remaining evidence represented as hesitation Atanassov (1989); Chaira (2011). For an element x, the three degrees are constrained as follows:

$$\mu_A(x) \geq 0, \quad \nu_A(x) \geq 0, \quad \pi_A(x) = 1 - \mu_A(x) - \nu_A(x) \geq 0. \tag{16}$$

Membership, non-membership, and hesitation represent supporting, opposing, and unresolved evidence, respectively. Unlike traditional fuzzy sets, this formulation does not require non-membership to equal the complement of membership, allowing incomplete evidence to be represented explicitly. When hesitation vanishes, membership and non-membership become complementary.

In GRANITE, each ball–cluster pair has a triplet $\mathbf{p}_{ij} = (\mu_{ij}, \nu_{ij}, \pi_{ij})^\top \in \Delta_2$, where $\Delta_2 = \{\mathbf{p} \in \mathbb{R}^3_+ : \mathbf{1}^\top \mathbf{p} = 1\}$. The signed coefficient $w_{ij} = \mu_{ij} - \nu_{ij}$ summarizes the balance between supporting and opposing evidence and satisfies $|w_{ij}| \leq 1 - \pi_{ij}$. Normalization is within each triplet, rather than across clusters, so the signed coefficients are not cluster probabilities. Moreover, the signed coefficient alone does not determine the triplet: both balanced support and opposition and complete hesitation can yield a zero coefficient.

The negative-entropy term favors less concentrated triplets and satisfies the following bounds:

$$-\log 3 \leq \sum_{r=1}^{3} p_{ijr} \log p_{ijr} \leq 0, \qquad 0 \log 0 := 0. \tag{17}$$

The lower bound is attained by the uniform triplet, while the upper bound is attained at a simplex vertex. Entropy differs from hesitation: $(0, 0, 1)$ has maximal hesitation but zero entropy. Thus the regularizer balances residual fitting by discouraging concentrated evidence, rather than directly maximizing hesitation.

Distinction between two hesitation-related quantities. It is critical to separate $\boldsymbol{\eta_i}$ (introduced in the main text Sec. 3.1) and $\boldsymbol{\pi_{ij}}$ defined above. $\boldsymbol{\eta_i}$ is a hand-crafted descriptive feature computed from raw input data, which quantifies intra-granular-ball scatter and is not a learned uncertainty variable. In contrast, $\boldsymbol{\pi_{ij}}$ is a trainable intuitionistic-fuzzy hesitation variable for each ball–cluster pair, which is jointly optimized alongside other model parameters.

A.2 Granular-Ball Computing

GB computing represents local sample groups by hyperspherical granules Xia et al. (2024). Given a nonempty partition $\{\mathcal{I}_i\}_{i=1}^{N_G}$ and $n_i = |\mathcal{I}_i|$, the center and radius of each ball are defined as follows:

$$\mathbf{c}_i = \frac{1}{n_i} \sum_{k \in \mathcal{I}_i} \mathbf{x}_k, \qquad r_i = \max_{k \in \mathcal{I}_i} \|\mathbf{x}_k - \mathbf{c}_i\|_2. \tag{18}$$

The center summarizes the location of the assigned samples, while the radius describes their spatial extent. The index sets specify sample membership; their enclosing balls may overlap geometrically. Different ball sizes provide summaries at different local scales. Coarser balls reduce the number of representation units, whereas finer balls retain more local variation.

GRANITE supplements the radius with the mean scatter $\sigma_i = n_i^{-1} \sum_{k \in \mathcal{I}_i} \|\mathbf{x}_k - \mathbf{c}_i\|_2$ and the bounded descriptor $\eta_i = \sigma_i / (\sigma_i + \tau)$, where $\tau > 0$. Thus $0 \leq \sigma_i \leq r_i$ and $0 \leq \eta_i < 1$. While the radius measures the largest deviation from the center, the mean scatter captures average dispersion. The parameter τ controls the scale of the bounded mapping.

The input is $\mathbf{s}_i = [\mathbf{c}_i^\top, r_i, \eta_i]^\top$. Here η_i describes within-ball dispersion, whereas π_{ij} describes uncertainty about a particular cluster. The partition and descriptors are fixed within each analyzed optimization stage and updated when the balls are reconstructed. Ball labels are propagated to samples through the current index sets.

A.3 Deep Autoencoder

An autoencoder learns nonlinear representations through an encoder–decoder reconstruction objective Zhang et al. (2019); Wei et al. (2025); Ren et al. (2025). The encoder maps an input to a latent representation, and the decoder reconstructs the input from that representation. For the structured GB descriptors, their mappings and reconstruction loss are defined as follows:

$$\mathbf{z}_i = f_\phi(\mathbf{s}_i), \quad \widehat{\mathbf{s}}_i = g_\psi(\mathbf{z}_i), \quad \mathcal{J}_{\mathrm{rec}} = \frac{1}{2}\sum_i \|\widehat{\mathbf{s}}_i - \mathbf{s}_i\|_2^2. \tag{19}$$

By composing affine transformations and nonlinear activations, the network learns representations that retain information needed to reconstruct the supplied descriptors.

GRANITE partitions the latent vector as $\mathbf{z}_i = [(\mathbf{c}_i^{(b)})^\top, r_i^{(b)}, \eta_i^{(b)}]^\top$. The center channel is used for clustering, while the scale and intrinsic-hesitation channels satisfy $r_i^{(b)} \geq 0$ and $0 \leq \eta_i^{(b)} \leq 1$ through their output parameterization. The learned scale enters the distribution weight; it is not a radius recomputed from encoded samples.

Reconstruction encourages the latent representation to preserve the descriptors' geometric and dispersion information, but does not by itself ensure cluster separation. GRANITE therefore combines reconstruction with the clustering objective to learn representations suited to both purposes. The encoder receives both signals, while the decoder receives the reconstruction signal. The corresponding parameter update is derived in Appendix C.6.

B Dynamic Granular-Ball Reconstruction and Label Propagation

This appendix specifies the dynamic GB reconstruction used by GRANITE. The number of GBs is fixed at N_G, while sample memberships are updated periodically during training. The procedure below also defines the final sample-to-ball mapping used for label propagation.

B.1 Initialization

Let $\mathbf{X}_0 = \{\mathbf{x}_k\}_{k=1}^N$ be the standardized input. Ball k-means partitions the samples into N_G nonempty index sets $\{\mathcal{I}_i^{(0)}\}_{i=1}^{N_G}$. For each ball i, we compute its center $\mathbf{c}_i^{(0)}$, radius $r_i^{(0)}$, mean scatter $\sigma_i^{(0)}$, and intrinsic hesitation $\eta_i^{(0)} = \sigma_i^{(0)}/(\sigma_i^{(0)} + \tau)$, where $\tau > 0$. These quantities form the initial structured descriptor

$$\mathbf{s}_i^{(0)} = [(\mathbf{c}_i^{(0)})^\top,\, r_i^{(0)},\, \eta_i^{(0)}]^\top \in \mathbb{R}^{d_0+2}. \tag{20}$$

B.2 Single-Point Encoding and Membership Reassignment

At a reconstruction step, each original sample is represented as a single-point GB descriptor:

$$\mathbf{s}_k^{\mathrm{single}} = [\mathbf{x}_k^\top,\, 0,\, 0]^\top \in \mathbb{R}^{d_0+2}. \tag{21}$$

Let $f_\phi^c(\cdot)$ denote the center channel of the encoder $f_\phi(\cdot)$. The frozen reassignment coordinate is

$$\mathbf{y}_k = \frac{f_\phi^c(\mathbf{s}_k^{\mathrm{single}}) - \bar{\mathbf{c}}}{\mathbf{s}_c}, \tag{22}$$

where $\bar{\mathbf{c}}$ and $\mathbf{s}_c$ are the componentwise mean and standard deviation of $\{f_\phi^c(\mathbf{s}_k^{\mathrm{single}})\}_{k=1}^N$, with a small positive constant added to $\mathbf{s}_c$ for numerical stability. The vector $\mathbf{y}_k$ is treated as frozen during membership reassignment.

For each current ball i, its reassignment center is the mean of $\mathbf{y}_k$ over $k \in \mathcal{I}_i$. Each sample is preliminarily assigned to the ball whose reassignment center is closest in Euclidean distance. Ties are broken by the smallest ball index. If a ball becomes empty, we repair it by moving the sample farthest from its center in the currently largest nonempty ball to the empty ball, and repeat until all N_G balls are nonempty. Ties in this repair step are broken by the smallest sample index.

B.3 Descriptor Recomputation

After empty-ball repair, the GB descriptors are recomputed on the original standardized data. For each ball i, let $\mathcal{I}'_i$ be its new nonempty index set. We recompute its center, radius, mean scatter, and intrinsic hesitation in the same way as in the initialization step, and form the new structured descriptor

$$\mathbf{s}'_i = [(\mathbf{c}'_i)^\top,\, r'_i,\, \eta'_i]^\top. \tag{23}$$

B.4 Triplet Inheritance via Membership Overlap

The intuitionistic triplets are transferred from the old partition to the new partition according to membership overlap:

$$\mathbf{p}'_{\ell j} = \sum_{i=1}^{N_G} \frac{|\mathcal{I}'_\ell \cap \mathcal{I}_i|}{|\mathcal{I}'_\ell|} \mathbf{p}_{ij}, \qquad j = 1, \ldots, K. \tag{24}$$

Since the overlap fractions sum to one for each new ball ℓ, the transferred triplet remains normalized: $\mathbf{p}'_{\ell j} \in \Delta_2$. Moreover, if the inherited triplets are strictly positive and the overlap is nonempty, then $\mathbf{p}'_{\ell j}$ is strictly positive. This preserves the simplex constraint and the strict-positivity required by the assignment update.

B.5 Variable Inheritance and Acceptance

The encoder and decoder parameters, cluster prototypes $\mathbf{H}$, covariance metric $\mathbf{\Sigma}$, and optimizer state are inherited from the previous stage. The ball-level descriptors are replaced by $\{\mathbf{s}'_i\}_{i=1}^{N_G}$, and the triplets are replaced by $\{\mathbf{p}'_{\ell j}\}$. The auxiliary latent channels $r_i^{(b)}$ and $\eta_i^{(b)}$ are recomputed by encoding the new descriptors $\mathbf{s}'_i$.

A reconstruction is accepted only if it does not increase the current objective $\mathcal{J}$. If the condition fails, the previous partition and variables are retained. This acceptance test prevents a reconstruction from degrading the joint reconstruction–clustering objective.

B.6 Stopping Budget

Reconstruction is performed at most R_G times. It stops early if the relative objective decrease is below a threshold $\eta_{\mathrm{GB}} > 0$ for s_{GB} consecutive reconstructions, or if the acceptance condition fails for s_{GB} consecutive reconstructions. The values of R_G, η_{GB}, and s_{GB} are fixed during training and reported in the implementation details.

B.7 Final Label Propagation

Let $b(k)$ denote the sample-to-ball mapping after the last accepted reconstruction. The final ball label is

$$\widehat{y}_i = \min \arg\max_{j \in \{1,\ldots,K\}} w_{ij}, \qquad w_{ij} = \mu_{ij} - \nu_{ij}, \tag{25}$$

where ties are broken by the smallest cluster index. The final sample label is

$$\widehat{y}_k = \widehat{y}_{b(k)}. \tag{26}$$

This propagation uses only the last accepted membership map, so the final labels are consistent with the final GB partition used by the model.

C Optimization Details

We derive the four block updates for the stage-wise objective in equation 8. The current GB descriptors remain fixed throughout each analyzed stage. Following the reweighting approach of Zhang et al. (2019); Nie et al. (2013), we replace the robust residual term by a quadratic surrogate and retain the other terms. Assume $\varepsilon, \lambda_1, \lambda_H, \gamma_\Sigma, \delta, \tau_r > 0$ and $\lambda_2 \geq 0$.

C.1 Proof of Theorem 1

Proof. Consider $\mathcal{F}(\mathbf{x}) = f(\mathbf{x}) + \sum_i s_i \varphi_\varepsilon(q_i(\mathbf{x}))$, where $q_i(\mathbf{x}) = \|\mathbf{g}_i(\mathbf{x})\|_2^2$, $s_i \geq 0$ are fixed, and $\varphi_\varepsilon(q) = (1+\varepsilon)q/\sqrt{q+\varepsilon^2}$ with $\varepsilon > 0$. The scalar penalty is strictly concave in the squared residual:

$$\varphi_\varepsilon''(q) = -\frac{(1+\varepsilon)(q+4\varepsilon^2)}{4(q+\varepsilon^2)^{5/2}} < 0, \qquad q \geq 0. \tag{27}$$

Consequently, its tangent is bounded as follows:

$$\varphi_\varepsilon(q) \leq \varphi_\varepsilon(q_0) + \varphi_\varepsilon'(q_0)(q - q_0) = d(q_0)q + b(q_0), \tag{28}$$

with equality at $q = q_0$, where $d(q_0) = \varphi_\varepsilon'(q_0)$ and $b(q_0) = \varphi_\varepsilon(q_0) - d(q_0)q_0$.

At iteration t, set $q_i^{(t)} = q_i(\mathbf{x}^{(t)})$, $d_i^{(t)} = d(q_i^{(t)})$, and $b_i^{(t)} = b(q_i^{(t)})$. Multiplying the tangent inequalities by s_i and adding $f(\mathbf{x})$ gives the tight upper bound is given as follows:

$$\begin{aligned} \widetilde{\mathcal{F}}_t(\mathbf{x}) &= f(\mathbf{x}) + \sum_i s_i \left[d_i^{(t)} \|\mathbf{g}_i(\mathbf{x})\|_2^2 + b_i^{(t)} \right] \\ &= f(\mathbf{x}) + \sum_i s_i d_i^{(t)} \|\mathbf{g}_i(\mathbf{x})\|_2^2 + C_t. \end{aligned} \tag{29}$$

Here $C_t = \sum_i s_i b_i^{(t)}$ is independent of $\mathbf{x}$. Thus minimizing this upper bound has exactly the same minimizers as minimizing $f(\mathbf{x}) + \sum_i s_i d_i^{(t)} \|\mathbf{g}_i(\mathbf{x})\|_2^2$, proving the theorem. The weights $s_i d_i^{(t)}$ are nonnegative because $\varphi_\varepsilon'(q) > 0$ for $q \geq 0$. □

C.2 Application to GRANITE

Write $\Theta = (\theta, \mathbf{H}, \mathbf{P}, \mathbf{\Sigma})$, where $\theta = \Theta_{\mathrm{AE}}$ and $\mathbf{P}$ collects the triplets. Let $\mathcal{C}$ denote the feasible set, including $\mathbf{p}_{ij} \in \Delta_2$ and $\mathbf{\Sigma} \succeq \delta \mathbf{I}$. Using the residual $\mathbf{v}_i = \mathbf{c}_i^{(b)} - \mathbf{H}\mathbf{w}_i$, define $q_i = \mathbf{v}_i^\top \mathbf{\Sigma}^{-1} \mathbf{v}_i$ and $\varphi_\varepsilon(q) = (1+\varepsilon)q/\sqrt{q+\varepsilon^2}$. The objective is expressed as follows:

$$\mathcal{J}(\Theta) = \mathcal{R}_0(\Theta) + \sum_i \omega_i(\theta) \varphi_\varepsilon(q_i(\Theta)), \qquad \omega_i = \exp[-(r_i^{(b)})^2/\tau_r^2], \tag{30}$$

where $\mathcal{R}_0$ collects the reconstruction, negative-entropy, log-determinant, and parameter-regularization terms from equation 8.

At outer iteration t, put $q_i^{(t)} = q_i(\Theta^{(t)})$. The tangent coefficients of φ_ε are defined as follows:

$$d_i^{(t)} = \varphi_\varepsilon'(q_i^{(t)}) = \frac{(1+\varepsilon)(q_i^{(t)} + 2\varepsilon^2)}{2(q_i^{(t)} + \varepsilon^2)^{3/2}}, \tag{31}$$

$$b_i^{(t)} = \varphi_\varepsilon(q_i^{(t)}) - d_i^{(t)} q_i^{(t)} = \frac{(1+\varepsilon)(q_i^{(t)})^2}{2(q_i^{(t)} + \varepsilon^2)^{3/2}}. \tag{32}$$

When the network parameters are fixed, Theorem 1 applies with $s_i = \omega_i$ and $\mathbf{g}_i = \mathbf{\Sigma}^{-1/2}\mathbf{v}_i$. For the network update, $\omega_i(\theta)$ varies with the candidate state. Applying the same scalar tangent inequality pointwise and retaining its weighted intercept, the full surrogate is defined as follows:

$$\mathcal{Q}_t(\Theta) = \mathcal{R}_0(\Theta) + \sum_i \omega_i(\theta) \left[d_i^{(t)} q_i(\Theta) + b_i^{(t)} \right]. \tag{33}$$

The coefficients $d_i^{(t)}$ and $b_i^{(t)}$ remain fixed during the sweep, while each block uses the latest values of the other variables. Since ω_i depends on the network, the weighted intercepts can be omitted from the prototype, assignment, and covariance blocks, but must be retained in the network block.

For the joint variables, define $e_i^{(t)}(\Theta) = d_i^{(t)} q_i(\Theta) + b_i^{(t)} - \varphi_\varepsilon(q_i(\Theta))$. Multiplying the nonnegative scalar gaps by the weights at the candidate state yields the following:

$$\mathcal{E}_t(\Theta) := \mathcal{Q}_t(\Theta) - \mathcal{J}(\Theta) = \sum_i \omega_i(\theta) e_i^{(t)}(\Theta) \geq 0, \quad \mathcal{Q}_t(\Theta^{(t)}) = \mathcal{J}(\Theta^{(t)}). \tag{34}$$

At the anchor, both $e_i^{(t)}$ and its derivative with respect to q_i vanish. The product rule then gives first-order agreement wherever the gradients exist, even for network-dependent weights. Together with the preceding bound, this yields the following:

$$\begin{aligned} \nabla_\Theta \mathcal{Q}_t(\Theta^{(t)}) &= \nabla_\Theta \mathcal{J}(\Theta^{(t)}), \\ \mathcal{Q}_t(\Theta^+) \le \mathcal{Q}_t(\Theta^{(t)}) &\implies \mathcal{J}(\Theta^+) \le \mathcal{J}(\Theta^{(t)}). \end{aligned} \tag{35}$$

Thus feasible surrogate descent is sufficient; exact joint minimization is unnecessary. Algorithm 1 summarizes the iterative reweighting procedure, alternating between surrogate coefficient updates and updates of the model parameters.

Algorithm 1 Reweighting Procedure for the GB Distribution Loss

Input: GB descriptors, feasible initial parameters $\Theta^{(0)}$, and $\varepsilon > 0$.

1: Set $t = 0$.
2: **repeat**
3: Evaluate $q_i^{(t)}$ and $\omega_i^{(t)}$ at $\Theta^{(t)}$ for each GB.
4: Update $d_i^{(t)} = \dfrac{(1+\varepsilon)(q_i^{(t)} + 2\varepsilon^2)}{2(q_i^{(t)} + \varepsilon^2)^{3/2}}$.
5: Update $b_i^{(t)} = \dfrac{(1+\varepsilon)(q_i^{(t)})^2}{2(q_i^{(t)} + \varepsilon^2)^{3/2}}$.
6: Set $\alpha_i^{(t)} = \omega_i^{(t)} d_i^{(t)}$.
7: Keep $d_i^{(t)}$ and $b_i^{(t)}$ fixed.
8: Update $\mathbf{H}$ and $\mathbf{P}$ using equation 38 and equation 43.
9: Update $\mathbf{\Sigma}$ and Θ_{AE} using equation 49 and equation 52.
10: Collect the updated parameters as $\Theta^{(t+1)}$ and set $t \leftarrow t + 1$.
11: **until** The stopping criterion is satisfied.

Output: Updated parameters $\Theta^{(t)}$.

C.3 Update of the Cluster Prototypes

With the other blocks fixed, let $\alpha_i = \omega_i d_i^{(t)}$, $\mathbf{C} = [\mathbf{c}_1^{(b)}, \dots, \mathbf{c}_{N_G}^{(b)}]$, $\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_{N_G}]$, and $\mathbf{A} = \mathrm{diag}(\alpha_1, \dots, \alpha_{N_G})$. The prototype objective is defined as follows:

$$F_H(\mathbf{H}) = \mathrm{tr}\big[\mathbf{\Sigma}^{-1}(\mathbf{C} - \mathbf{H}\mathbf{W})\mathbf{A}(\mathbf{C} - \mathbf{H}\mathbf{W})^\top\big] + \frac{\lambda_H}{2}\|\mathbf{H}\|_F^2. \tag{36}$$

Set $\mathbf{M} = \mathbf{W}\mathbf{A}\mathbf{W}^\top$ and $\mathbf{B} = \mathbf{C}\mathbf{A}\mathbf{W}^\top$. Setting the gradient to zero, the stationarity condition is as follows:

$$2\mathbf{\Sigma}^{-1}(\mathbf{H}\mathbf{M} - \mathbf{B}) + \lambda_H \mathbf{H} = \mathbf{0} \quad \Longleftrightarrow \quad \mathbf{H}\mathbf{M} + \frac{\lambda_H}{2}\mathbf{\Sigma}\mathbf{H} = \mathbf{B}. \tag{37}$$

The ridge term makes F_H strictly convex. Diagonalize $\mathbf{\Sigma} = \mathbf{U}_\Sigma \,\mathrm{diag}(\ell_1, \dots, \ell_{d_b})\mathbf{U}_\Sigma^\top$ and $\mathbf{M} = \mathbf{U}_M \,\mathrm{diag}(m_1, \dots, m_K)\mathbf{U}_M^\top$. In these bases, the equation separates entrywise. The prototype update is as follows:

$$Y_{ab}^+ = \frac{(\mathbf{U}_\Sigma^\top \mathbf{B}\mathbf{U}_M)_{ab}}{m_b + \lambda_H \ell_a/2}, \qquad \mathbf{H}^+ = \mathbf{U}_\Sigma \mathbf{Y}^+ \mathbf{U}_M^\top. \tag{38}$$

The denominator is positive since $m_b \ge 0$ and $\ell_a \ge \delta$. Expanding the quadratic objective around its minimizer, the objective decrease is expressed as follows:

$$F_H(\mathbf{H}) - F_H(\mathbf{H}^+) = \|\mathbf{\Sigma}^{-1/2}(\mathbf{H} - \mathbf{H}^+)\mathbf{W}\mathbf{A}^{1/2}\|_F^2 + \frac{\lambda_H}{2}\|\mathbf{H} - \mathbf{H}^+\|_F^2 \ge 0. \tag{39}$$

C.4 Update of the Intuitionistic Assignments

Stack the triplets of ball i into $\mathbf{p}_i \in \Delta_2^K$, and put $\mathbf{D} = \mathbf{I}_K \otimes [1, -1, 0]$, so that $\mathbf{D}\mathbf{p}_i = \mathbf{w}_i$. With $h_i(\mathbf{p}_i) = \sum_{j,r} p_{ijr} \log p_{ijr}$, the assignment objective is defined as follows:

$$F_i(\mathbf{p}_i) = G_i(\mathbf{p}_i) + \lambda_1 h_i(\mathbf{p}_i), \qquad G_i(\mathbf{p}_i) = \alpha_i \|\mathbf{c}_i^{(b)} - \mathbf{H}\mathbf{D}\mathbf{p}_i\|_{\mathbf{\Sigma}^{-1}}^2. \tag{40}$$

The quadratic term is convex, and the negative entropy makes the minimizer unique and interior. Since the residual couples all cluster coefficients, we use feasible inner iterations rather than treating the stationarity relation as a one-step solution. The gradient and a Lipschitz bound are expressed as follows:

$$\mathbf{a}_i := \nabla G_i(\mathbf{p}_i) = -2\alpha_i \mathbf{D}^\top \mathbf{H}^\top \mathbf{\Sigma}^{-1} \mathbf{v}_i, \qquad L_i = 4\alpha_i \lambda_{\max}(\mathbf{H}^\top \mathbf{\Sigma}^{-1} \mathbf{H}). \tag{41}$$

For strictly positive $\mathbf{p}_i^{(\ell)}$, choose $\beta_i > 0$ with $\beta_i L_i \le 1$. Let $D_i(\mathbf{p}\|\mathbf{q}) = \sum_j D_{\mathrm{KL}}(\mathbf{p}_j\|\mathbf{q}_j)$. After linearizing G_i and adding a KL proximity term, the inner surrogate is defined as follows:

$$\mathcal{U}_i(\mathbf{p}_i \mid \mathbf{p}_i^{(\ell)}) = G_i(\mathbf{p}_i^{(\ell)}) + \langle \mathbf{a}_i^{(\ell)}, \mathbf{p}_i - \mathbf{p}_i^{(\ell)} \rangle + \lambda_1 h_i(\mathbf{p}_i) + \beta_i^{-1} D_i(\mathbf{p}_i\|\mathbf{p}_i^{(\ell)}). \tag{42}$$

After differentiation under each triplet's sum-to-one constraint and elimination of its multiplier, the explicit update is as follows:

$$z_{ijr}^{(\ell)} = \frac{\log p_{ijr}^{(\ell)} - \beta_i a_{ijr}^{(\ell)}}{1 + \beta_i \lambda_1}, \qquad p_{ijr}^{(\ell+1)} = \frac{\exp z_{ijr}^{(\ell)}}{\sum_{s=1}^{3} \exp z_{ijs}^{(\ell)}}. \tag{43}$$

Here $(a_{ij1}^{(\ell)}, a_{ij2}^{(\ell)}, a_{ij3}^{(\ell)}) = (-u_{ij}^{(\ell)}, u_{ij}^{(\ell)}, 0)$, where $u_{ij}^{(\ell)} = 2\alpha_i \mathbf{h}_j^\top \mathbf{\Sigma}^{-1} \mathbf{v}_i^{(\ell)}$. These components are denoted by $g_{ijr}^{(k)}$ in the main-text update, with k corresponding to the inner index ℓ here.

For strictly positive $\mathbf{p}, \mathbf{q} \in \Delta_2^K$, let $\mathbf{d} = \mathbf{p} - \mathbf{q}$. The Hessians satisfy $\nabla^2 h_i(\mathbf{p}) = \mathrm{diag}(1/p_{ijr}) \succeq \mathbf{I}$ and $\nabla^2 G_i = 2\alpha_i \mathbf{D}^\top \mathbf{H}^\top \mathbf{\Sigma}^{-1} \mathbf{H} \mathbf{D} \preceq L_i \mathbf{I}$. Together with $\beta_i L_i \le 1$, these bounds establish the inner-surrogate majorization and descent as follows:

$$\begin{aligned} D_i(\mathbf{p}\|\mathbf{q}) &= h_i(\mathbf{p}) - h_i(\mathbf{q}) - \langle \nabla h_i(\mathbf{q}), \mathbf{d} \rangle \\ &= \int_0^1 (1-u) \mathbf{d}^\top \nabla^2 h_i(\mathbf{q} + u\mathbf{d}) \mathbf{d} \, \mathrm{d}u \\ &\ge \int_0^1 (1-u) \|\mathbf{d}\|_2^2 \, \mathrm{d}u = \frac{1}{2} \|\mathbf{d}\|_2^2. \end{aligned} \tag{44}$$

$$\begin{aligned} G_i(\mathbf{p}) &= G_i(\mathbf{q}) + \langle \nabla G_i(\mathbf{q}), \mathbf{d} \rangle + \frac{1}{2} \mathbf{d}^\top \nabla^2 G_i \mathbf{d} \\ &\le G_i(\mathbf{q}) + \langle \nabla G_i(\mathbf{q}), \mathbf{d} \rangle + \frac{L_i}{2} \|\mathbf{d}\|_2^2, \end{aligned} \tag{45}$$

$$\mathcal{U}_i(\mathbf{p} \mid \mathbf{q}) - F_i(\mathbf{p}) = \beta_i^{-1} D_i(\mathbf{p}\|\mathbf{q}) - \frac{1}{2} \mathbf{d}^\top \nabla^2 G_i \mathbf{d} \ge \frac{1}{2} (\beta_i^{-1} - L_i) \|\mathbf{d}\|_2^2 \ge 0, \tag{46}$$

$$\begin{aligned} F_i(\mathbf{p}_i^{(\ell+1)}) &\le \mathcal{U}_i(\mathbf{p}_i^{(\ell+1)} \mid \mathbf{p}_i^{(\ell)}) - \frac{\beta_i^{-1} - L_i}{2} \|\mathbf{p}_i^{(\ell+1)} - \mathbf{p}_i^{(\ell)}\|_2^2 \\ &\le \mathcal{U}_i(\mathbf{p}_i^{(\ell)} \mid \mathbf{p}_i^{(\ell)}) - \frac{\beta_i^{-1} - L_i}{2} \|\mathbf{p}_i^{(\ell+1)} - \mathbf{p}_i^{(\ell)}\|_2^2 \\ &= F_i(\mathbf{p}_i^{(\ell)}) - \frac{\beta_i^{-1} - L_i}{2} \|\mathbf{p}_i^{(\ell+1)} - \mathbf{p}_i^{(\ell)}\|_2^2 \\ &\le F_i(\mathbf{p}_i^{(\ell)}). \end{aligned} \tag{47}$$

The descent chain follows because the explicit update minimizes $\mathcal{U}_i$ and $\mathcal{U}_i(\mathbf{q} \mid \mathbf{q}) = F_i(\mathbf{q})$. Each step preserves positive, normalized triplets and does not increase the same outer surrogate. Recompute the residual after each step and use a log-domain softmax for numerical stability.

C.5 Update of the Covariance Metric

Let $\mathbf{S} = \sum_i \alpha_i \mathbf{v}_i \mathbf{v}_i^\top$, using the latest prototypes and assignments. The covariance objective and its precision-variable form are defined as follows:

$$\begin{aligned} F_\Sigma(\mathbf{\Sigma}) &= \mathrm{tr}(\mathbf{\Sigma}^{-1} \mathbf{S}) + \frac{\gamma_\Sigma}{2} \log\det \mathbf{\Sigma}, \quad \mathbf{\Sigma} \succeq \delta \mathbf{I}, \\ F_T(\mathbf{T}) &= \mathrm{tr}(\mathbf{T}\mathbf{S}) - \frac{\gamma_\Sigma}{2} \log\det \mathbf{T}, \qquad \mathbf{0} \prec \mathbf{T} \preceq \delta^{-1} \mathbf{I}, \end{aligned} \tag{48}$$

where $\mathbf{T} = \mathbf{\Sigma}^{-1}$. The precision objective is strictly convex. Let s_k be the eigenvalues of $\mathbf{S}$ and let $\mathbf{U}$ contain its orthonormal eigenvectors. In this basis, replacing the precision matrix by its diagonal preserves the trace and feasibility and cannot increase the negative log-determinant, by Hadamard's inequality. Each precision eigenvalue therefore minimizes $s_k t - (\gamma_\Sigma/2)\log t$ on $0 < t \le \delta^{-1}$. The interior solution is $\gamma_\Sigma/(2s_k)$ when $s_k > \gamma_\Sigma\delta/2$; otherwise the solution is δ^{-1}. After inverting these eigenvalues, the covariance update is as follows:

$$\mathbf{\Sigma}^+ = \mathbf{U}\,\mathrm{diag}(\ell_1^+, \ldots, \ell_{d_b}^+)\mathbf{U}^\top, \qquad \ell_k^+ = \max\{\delta, 2s_k/\gamma_\Sigma\}. \tag{49}$$

This is the unique constrained minimizer, including when $\mathbf{S}$ is singular, and hence decreases the covariance block.

C.6 Update of the Autoencoder Parameters

Let $\mathbf{P}_\phi$ select the encoder weights and biases from $\theta = \Theta_{\mathrm{AE}}$. With the other blocks fixed, the network-dependent objective is defined as follows:

$$F_t(\theta) = \frac{1}{2}\sum_i \|\widehat{\mathbf{s}}_i(\theta) - \mathbf{s}_i\|_2^2 + \sum_i \omega_i(\theta)[d_i^{(t)} q_i(\theta) + b_i^{(t)}] + \frac{\lambda_2}{2}\|\mathbf{P}_\phi\theta\|_2^2. \tag{50}$$

Write $\mathbf{e}_i = \widehat{\mathbf{s}}_i - \mathbf{s}_i$. Let $\mathbf{J}_{\widehat{s},i}$ and $\mathbf{J}_{c,i}$ be the Jacobians of the reconstruction and latent center with respect to θ, and let $\mathbf{j}_{r,i} = \nabla_\theta r_i^{(b)}$. Using the product rule, the complete gradient is expressed as follows:

$$\begin{aligned}\mathbf{g} := \nabla_\theta F_t = &\sum_i \left[\mathbf{J}_{\widehat{s},i}^\top \mathbf{e}_i + 2\omega_i d_i^{(t)} \mathbf{J}_{c,i}^\top \mathbf{\Sigma}^{-1}\mathbf{v}_i\right] \\ &- \sum_i \frac{2r_i^{(b)}}{\tau_r^2}\omega_i[d_i^{(t)} q_i + b_i^{(t)}]\mathbf{j}_{r,i} + \lambda_2 \mathbf{P}_\phi^\top \mathbf{P}_\phi \theta.\end{aligned} \tag{51}$$

The weighted intercept contributes to the scale-channel derivative and must not be dropped. Standard backpropagation evaluates this gradient: reconstruction affects both encoder and decoder, while the residual and scale terms act through the encoder.

For $\mathbf{g} \neq \mathbf{0}$, backtracking tests $\alpha_n = \bar{\alpha}\beta_{\mathrm{AE}}^n$, where $\bar{\alpha} > 0$ and $0 < \beta_{\mathrm{AE}} < 1$, and accepts the first n_* for which the following condition holds:

$$\theta^+ = \theta - \alpha_{n_*}\mathbf{g}, \qquad F_t(\theta^+) \le F_t(\theta) - c\alpha_{n_*}\|\mathbf{g}\|_2^2, \quad 0 < c < 1. \tag{52}$$

If the gradient is locally L-Lipschitz, any sufficiently short step with $\alpha \le 2(1-c)/L$ satisfies this inequality, so backtracking terminates. For a zero gradient, retain the current parameters. Other optimizers, including SGD, satisfy the analyzed descent condition only if their candidate steps pass the same full-surrogate nonincrease test.

C.7 Complete Outer Iteration

Algorithm 2 summarizes one outer iteration, where $\theta^t = \Theta_{\mathrm{AE}}^{(t)}$ and $\mathbf{P}^t$ collects the intuitionistic triplets. The intermediate states $\Theta^{(t,k)}$ record the variables after each block update and are used in the convergence proof. The complete outer iteration is as follows:

The reweighting coefficients are refreshed only at the next outer iteration. Within the sweep, residuals, scatter matrices, and gradients use the latest block values. Additional inner updates are permitted provided that they decrease the same surrogate $\mathcal{Q}_t$.

C.8 Acceptance and Stopping Criteria

For the update schedule in Appendix C.7, every accepted block update must preserve feasibility and decrease the same surrogate:

$$\begin{aligned}\Theta^{(t,k)} \in \mathcal{C}, \qquad \mathcal{Q}_t(\Theta^{(t,k)}) \le \mathcal{Q}_t(\Theta^{(t,k-1)}), \quad k = 1, \ldots, 4, \\ \mathcal{C} = \{\Theta : \mathbf{p}_{ij} \in \Delta_2 \text{ for all } i, j,\ \mathbf{\Sigma} \succeq \delta\mathbf{I}\}.\end{aligned} \tag{53}$$

Algorithm 2 One Descent-Controlled Outer Sweep of GRANITE

Require: Feasible state $\Theta^{(t)} = (\theta^t, \mathbf{H}^t, \mathbf{P}^t, \mathbf{\Sigma}^t)$.
Ensure: Updated state $\Theta^{(t+1)}$.
1: Set $\Theta^{(t,0)} \leftarrow \Theta^{(t)}$.
2: Compute $q_i^{(t)}, d_i^{(t)}, b_i^{(t)}$ using equation 31; keep $d_i^{(t)}, b_i^{(t)}$ fixed throughout this sweep.
3: Compute $\mathbf{H}^{t+1}$ by equation 38.
$\Theta^{(t,1)} \leftarrow (\theta^t, \mathbf{H}^{t+1}, \mathbf{P}^t, \mathbf{\Sigma}^t)$.
4: Apply the feasible inner updates equation 43 to obtain $\mathbf{P}^{t+1}$.
$\Theta^{(t,2)} \leftarrow (\theta^t, \mathbf{H}^{t+1}, \mathbf{P}^{t+1}, \mathbf{\Sigma}^t)$.
5: Recompute $\mathbf{S}$ and obtain $\mathbf{\Sigma}^{t+1}$ from equation 49.
$\Theta^{(t,3)} \leftarrow (\theta^t, \mathbf{H}^{t+1}, \mathbf{P}^{t+1}, \mathbf{\Sigma}^{t+1})$.
6: Evaluate $\mathbf{g}_t = \nabla_\theta \mathcal{Q}_t(\Theta^{(t,3)})$ using equation 51.
7: Select α_t by equation 52 and set $\theta^{t+1} \leftarrow \theta^t - \alpha_t \mathbf{g}_t$; retain θ^t if $\mathbf{g}_t = \mathbf{0}$.
$\Theta^{(t,4)} \leftarrow (\theta^{t+1}, \mathbf{H}^{t+1}, \mathbf{P}^{t+1}, \mathbf{\Sigma}^{t+1})$.
8: **return** $\Theta^{(t+1)} \leftarrow \Theta^{(t,4)}$.

The network parameterization additionally enforces the latent-channel ranges. The spectral covariance update and triplet-normalized assignment update preserve their constraints by construction. If a proposed network step fails the full-surrogate test, backtracking reduces the step; retaining the previous feasible state is also a valid nonincreasing update.

A practical outer stopping test monitors the original objective after the complete sweep. With a tolerance $\eta_{\text{stop}} > 0$, the relative decrease is defined as follows:

$$r_t = \frac{|\mathcal{J}(\Theta^{(t+1)}) - \mathcal{J}(\Theta^{(t)})|}{\max\{1, |\mathcal{J}(\Theta^{(t)})|\}}, \tag{54}$$

and the outer loop is terminated when $r_t \le \eta_{\text{stop}}$. A prescribed iteration budget may be used as an additional stopping criterion. The relative-change test measures objective stabilization; it does not evaluate the first-order optimality residual.

D Convergence Analysis

This appendix proves stage-wise monotonicity and boundedness using the surrogate properties and block updates established in Appendix C. The parameters satisfy the assumptions in Appendix C.2, and the initial state is feasible with finite objective value. Each four-block sweep compares the same surrogate $\mathcal{Q}_t$, whose tangent coefficients are refreshed only after the sweep.

D.1 Monotonic Objective Descent

Proof. Use the intermediate states $\Theta^{(t,k)}$ defined in Algorithm 2. For each of the four blocks, we define its surrogate decrease as follows:

$$D_{t,k} = \mathcal{Q}_t(\Theta^{(t,k-1)}) - \mathcal{Q}_t(\Theta^{(t,k)}), \qquad k = 1, \dots, 4. \tag{55}$$

For K_i accepted assignment steps, let $\Delta\mathbf{p}_i^{(\ell)} = \mathbf{p}_i^{(\ell+1)} - \mathbf{p}_i^{(\ell)}$. The purpose of introducing these intermediate states is to ensure that every comparison concerns the same function $\mathcal{Q}_t$. Each block decrease can then be identified with the decrease of its own subproblem, because all other terms cancel. The first chain verifies these four decreases; the second uses the majorization and contact properties to transfer them to the original objective.

$$\begin{aligned} D_{t,1} &= F_H(\mathbf{H}^t) - F_H(\mathbf{H}^{t+1}) \\ &= \|(\mathbf{\Sigma}^t)^{-1/2}(\mathbf{H}^t - \mathbf{H}^{t+1})\mathbf{W}^t(\mathbf{A}^t)^{1/2}\|_F^2 + \frac{\lambda_H}{2}\|\mathbf{H}^t - \mathbf{H}^{t+1}\|_F^2 \\ &\ge 0, \end{aligned} \tag{56}$$

$$D_{t,2} = \sum_i [F_i(\mathbf{p}_i^{(0)}) - F_i(\mathbf{p}_i^{(K_i)})] = \sum_i \sum_{\ell=0}^{K_i-1} [F_i(\mathbf{p}_i^{(\ell)}) - F_i(\mathbf{p}_i^{(\ell+1)})]$$

$$\geq \frac{1}{2}\sum_i (\beta_i^{-1} - L_i) \sum_{\ell=0}^{K_i-1} \|\Delta \mathbf{p}_i^{(\ell)}\|_2^2 \geq 0, \tag{57}$$

$$\begin{aligned} D_{t,3} &= F_\Sigma(\mathbf{\Sigma}^t) - F_\Sigma(\mathbf{\Sigma}^{t+1}) \\ &= \mathrm{tr}\big[((\mathbf{\Sigma}^t)^{-1} - (\mathbf{\Sigma}^{t+1})^{-1})\mathbf{S}\big] + \frac{\gamma_\Sigma}{2}\log\frac{\det \mathbf{\Sigma}^t}{\det \mathbf{\Sigma}^{t+1}} \geq 0, \end{aligned} \tag{58}$$

$$D_{t,4} = F_t(\theta^t) - F_t(\theta^{t+1}) \geq c\alpha_t \|\nabla_\theta F_t(\theta^t)\|_2^2 \geq 0, \tag{59}$$

$$\mathcal{Q}_t(\Theta^{(t,k)}) = \mathcal{Q}_t(\Theta^{(t,k-1)}) - D_{t,k} \leq \mathcal{Q}_t(\Theta^{(t,k-1)}), \qquad k = 1,2,3,4, \tag{60}$$

$$\mathcal{Q}_t(\Theta^{(t+1)}) = \mathcal{Q}_t(\Theta^{(t)}) - \sum_{k=1}^{4} D_{t,k} \leq \mathcal{Q}_t(\Theta^{(t)}). \tag{61}$$

$$\begin{aligned} \mathcal{J}(\Theta^{(t+1)}) &= \mathcal{R}_0(\Theta^{(t+1)}) + \sum_i \omega_i(\theta^{t+1}) \frac{(1+\varepsilon) q_i(\Theta^{(t+1)})}{\sqrt{q_i(\Theta^{(t+1)}) + \varepsilon^2}} \\ &\leq \mathcal{R}_0(\Theta^{(t+1)}) + \sum_i \omega_i(\theta^{t+1})[d_i^{(t)} q_i(\Theta^{(t+1)}) + b_i^{(t)}] \\ &= \mathcal{Q}_t(\Theta^{(t,4)}) \leq \mathcal{Q}_t(\Theta^{(t,3)}) \leq \mathcal{Q}_t(\Theta^{(t,2)}) \\ &\leq \mathcal{Q}_t(\Theta^{(t,1)}) \leq \mathcal{Q}_t(\Theta^{(t,0)}) \\ &= \mathcal{R}_0(\Theta^{(t)}) + \sum_i \omega_i(\theta^t)[d_i^{(t)} q_i^{(t)} + b_i^{(t)}] \\ &= \mathcal{R}_0(\Theta^{(t)}) + \sum_i \omega_i(\theta^t)\varphi_\varepsilon(q_i^{(t)}) = \mathcal{J}(\Theta^{(t)}), \end{aligned} \tag{62}$$

$$\begin{aligned} \mathcal{J}(\Theta^{(t)}) - \mathcal{J}(\Theta^{(t+1)}) &= \mathcal{Q}_t(\Theta^{(t)}) - \mathcal{Q}_t(\Theta^{(t+1)}) + \mathcal{Q}_t(\Theta^{(t+1)}) - \mathcal{J}(\Theta^{(t+1)}) \\ &= \sum_{k=1}^{4} D_{t,k} + \sum_i \omega_i(\theta^{t+1}) e_i^{(t)}(\Theta^{(t+1)}) \geq \sum_{k=1}^{4} D_{t,k} \geq 0. \end{aligned} \tag{63}$$

The four block inequalities follow respectively from the prototype quadratic identity, the assignment inner majorization, the constrained covariance minimizer, and the network acceptance rule. Exact solution of every block is unnecessary: any feasible update with $D_{t,k} \geq 0$ preserves the conclusion. The weights at the new state multiply the majorization gap, so encoder-dependent weights do not invalidate the chain. Descent of the original objective is guaranteed after each complete outer sweep; an individual inner block need not decrease the original objective when the anchor is held fixed. □

D.2 Convergence of Objective Values

Proof. Let $\mathbf{u} = (1/3, 1/3, 1/3)^\top$, let $\ell_k(\Theta)$ be the eigenvalues of $\mathbf{\Sigma}$, and put $J_t = \mathcal{J}(\Theta^{(t)})$. Monotonicity alone is insufficient: the objective could otherwise decrease without a finite limit. The potentially negative contributions are the triplet entropy and the covariance log-determinant. We bound them separately using the simplex normalization and the covariance eigenvalue floor; all remaining terms are nonnegative. Once a common finite lower bound is available, an explicit infimum argument proves convergence without assuming convergence of the parameter iterates. The complete chain is as follows:

$$p_{ijr} \geq 0, \qquad \sum_{r=1}^{3} p_{ijr} = 1, \qquad 0\log 0 = 0,$$

$$D_{\mathrm{KL}}(\mathbf{p}_{ij}\|\mathbf{u}) = \sum_{r=1}^{3} p_{ijr}\log\frac{p_{ijr}}{1/3} = \sum_{r=1}^{3} p_{ijr}\log p_{ijr} + \log 3 \geq 0. \tag{64}$$

$$\sum_{r=1}^{3} p_{ijr}\log p_{ijr} \geq -\log 3, \qquad \mathcal{R}_{\mathrm{NE}} = \sum_{i=1}^{N_G}\sum_{j=1}^{K}\sum_{r=1}^{3} p_{ijr}\log p_{ijr} \geq -N_G K \log 3. \tag{65}$$

$$\boldsymbol{\Sigma} \succeq \delta \mathbf{I} \implies \ell_k(\Theta) \geq \delta > 0 \implies \det \boldsymbol{\Sigma} = \prod_{k=1}^{d_b} \ell_k(\Theta) \geq \delta^{d_b},$$

$$\log \det \boldsymbol{\Sigma} = \sum_{k=1}^{d_b} \log \ell_k(\Theta) \geq d_b \log \delta. \tag{66}$$

$$q_i(\Theta) = \|\boldsymbol{\Sigma}^{-1/2}\mathbf{v}_i\|_2^2 \geq 0, \qquad \omega_i \varphi_\varepsilon(q_i) = \exp[-(r_i^{(b)})^2/\tau_r^2] \, \frac{(1+\varepsilon)q_i}{\sqrt{q_i+\varepsilon^2}} \geq 0. \tag{67}$$

$$\begin{aligned} \mathcal{J}(\Theta) = \frac{1}{2}\sum_i \|\widehat{\mathbf{s}}_i - \mathbf{s}_i\|_2^2 + \sum_i \omega_i \varphi_\varepsilon(q_i) + \frac{\lambda_H}{2}\|\mathbf{H}\|_F^2 + \frac{\lambda_2}{2}\|\mathbf{P}_\phi \theta\|_2^2 \\ + \lambda_1 \sum_{i,j,r} p_{ijr} \log p_{ijr} + \frac{\gamma_\Sigma}{2}\sum_{k=1}^{d_b} \log \ell_k(\Theta) \\ \geq -\lambda_1 N_G K \log 3 + \frac{\gamma_\Sigma d_b}{2} \log \delta =: J_{\mathrm{lb}} > -\infty. \end{aligned} \tag{68}$$

$$J_{\mathrm{lb}} \leq J_{t+1} \leq J_t \leq \cdots \leq J_1 \leq J_0 < +\infty,$$

$$J_* := \inf_{t \geq 0} J_t \in [J_{\mathrm{lb}}, J_0].$$

$$\forall \eta > 0, \ \exists T_\eta \geq 0 : \quad J_* \leq J_{T_\eta} < J_* + \eta,$$

$$t \geq T_\eta \implies J_* \leq J_t \leq J_{T_\eta} < J_* + \eta \implies |J_t - J_*| < \eta.$$

$$\lim_{t\to\infty} \mathcal{J}(\Theta^{(t)}) = J_*, \qquad J_* \geq \frac{\gamma_\Sigma d_b}{2} \log \delta - \lambda_1 N_G K \log 3. \tag{69}$$

$$\Delta_t := J_t - J_{t+1} \geq 0, \qquad \sum_{t=0}^{\infty} \Delta_t = J_0 - J_* < +\infty, \qquad \lim_{t\to\infty} \Delta_t = 0, \tag{70}$$

$$\begin{aligned} \sum_{t=0}^{T-1} \Delta_t &= (J_0 - J_1) + (J_1 - J_2) + \cdots + (J_{T-1} - J_T) \\ &= J_0 - J_T \leq J_0 - J_{\mathrm{lb}}, \end{aligned} \tag{71}$$

$$0 \leq \sum_{t=0}^{T-1} \left[\sum_{k=1}^{4} D_{t,k} + \mathcal{E}_t(\Theta^{(t+1)}) \right] = J_0 - J_T \leq J_0 - J_{\mathrm{lb}},$$

$$\lim_{t\to\infty} D_{t,k} = 0 \quad (k = 1, \ldots, 4), \qquad \lim_{t\to\infty} \mathcal{E}_t(\Theta^{(t+1)}) = 0. \tag{72}$$

The entropy bound is per three-component evidence triplet, hence $\log 3$ rather than $\log K$; boundary triplets follow by continuity. The finite lower bound need not be nonnegative when $\delta < 1$. The infimum argument establishes objective-value convergence if the stage-wise sequence is continued indefinitely, and telescoping shows that the nonnegative decreases then vanish. □

D.3 Role of the Covariance Eigenvalue Floor

Positive definiteness alone does not provide the lower bound used above. Consider a constant encoder with $\mathbf{c}_i^{(b)} = \mathbf{c}_0$ for all i. For the first prototype, choose each triplet as $(2/3, 1/6, 1/6)^\top$; for the remaining prototypes, choose $(1/3, 1/3, 1/3)^\top$. These are feasible and give $\mathbf{w}_i = (1/2, 0, \ldots, 0)^\top$. Taking $\mathbf{H} = [2\mathbf{c}_0, \mathbf{0}, \ldots, \mathbf{0}]$ therefore makes every residual zero. A constant encoder and decoder can have finite reconstruction loss and finite parameter regularization for finite descriptors. Holding these variables fixed and setting $\boldsymbol{\Sigma} = \xi \mathbf{I}_{d_b}$, the objective behaves as follows:

$$\mathcal{J} = C + \frac{\gamma_\Sigma d_b}{2} \log \xi \longrightarrow -\infty \qquad (\xi \downarrow 0), \tag{73}$$

where C is finite and independent of ξ. Thus $\mathbf{\Sigma} \succ 0$ by itself is insufficient for the general boundedness claim. GRANITE enforces the stronger constraint $\mathbf{\Sigma} \succeq \delta \mathbf{I}$ at every update; a one-time initialization perturbation would not provide this guarantee.

D.4 Bounds on the Objective Sublevel Set

Every iterate of the analyzed sequence lies in $\{\Theta : \mathcal{J}(\Theta) \leq J_0\}$. Using the bounds in equation 65 and equation 66, the nonnegative regularization terms are bounded as follows:

$$\frac{\lambda_H}{2}\|\mathbf{H}\|_F^2 + \frac{\lambda_2}{2}\|\mathbf{P}_\phi \theta\|_2^2 \leq \mathcal{J}(\Theta) - J_{\mathrm{lb}} \leq J_0 - J_{\mathrm{lb}}, \tag{74}$$

$$\|\mathbf{H}\|_F \leq \sqrt{\frac{2(J_0 - J_{\mathrm{lb}})}{\lambda_H}}, \tag{75}$$

$$\|\mathbf{P}_\phi \theta\|_2 \leq \sqrt{\frac{2(J_0 - J_{\mathrm{lb}})}{\lambda_2}} \quad (\lambda_2 > 0). \tag{76}$$

For the covariance, bounding the entropy below while keeping the log-determinant term gives a determinant upper bound. Combining it with the eigenvalue floor yields the following bounds:

$$\log\det \mathbf{\Sigma} \leq \frac{2(J_0 + \lambda_1 N_G K \log 3)}{\gamma_\Sigma}, \tag{77}$$

$$\begin{aligned}\lambda_{\max}(\mathbf{\Sigma}) &\leq \frac{\det \mathbf{\Sigma}}{\delta^{d_b - 1}} \\ &\leq \delta^{1-d_b} \exp\left[\frac{2(J_0 + \lambda_1 N_G K \log 3)}{\gamma_\Sigma}\right].\end{aligned} \tag{78}$$

Together with $\mathbf{\Sigma} \succeq \delta \mathbf{I}$, these bounds control the prototype and covariance blocks. The encoder bound requires $\lambda_2 > 0$. Because the stated regularizer does not control decoder parameters, the bounds do not imply compactness of the complete parameter sequence.

D.5 Training-Order Matching for Convergence

GRANITE is optimized according to an alternating training pipeline: it first performs GB reconstruction, then performs the reconstruction loss update, next performs clustering updates, and finally performs the clustering loss update. Our convergence proof is derived specifically to align with this practical training schedule. Within each stage where the GB descriptors are held fixed, the four block updates guarantee that the stage-wise surrogate objective does not increase. Dynamic GB reconstruction is guarded by an acceptance criterion: only new partitions that do not increase the current full objective are accepted and retained. Consequently, the sequence of objective values is monotonically non-increasing and bounded below, so the overall algorithm is amenable to optimization.

E Additional Experiments and Analysis

E.1 Experimental Setting

Datasets and Evaluation Metrics. We evaluate GRANITE on both general clustering and unsupervised image segmentation tasks. For general clustering, we use eight widely adopted UCI benchmark datasets, including Iris, Wine, Seed, Glass, WDBC, Segment, Pendigits, and Landsat, which cover diverse sample sizes, feature dimensions, and numbers of clusters. Accuracy (ACC), normalized mutual information (NMI) Fang et al. (2021), and adjusted Rand index (ARI) are used as the evaluation metrics. For image segmentation, we conduct experiments on BSD500 Wang et al. (2017) and DUTS, both of which provide pixel-level ground-truth annotations. Segmentation accuracy (ACC), mean intersection over union (mIoU), and NMI are reported for quantitative evaluation.

Table 4: Statistics of the benchmark datasets.

Dataset	Samples	Features	Classes
Iris	150	4	3
Wine	178	13	3
Seed	210	7	3
Glass	214	9	6
WDBC	569	30	2
Segment	2310	19	7
Landsat	6435	36	6
Pendigits	10992	16	10

Compared Methods. For clustering tasks, we compare GRANITE with three groups of representative methods: (1) *traditional clustering*, including k-means Lloyd (1982), DPHKM Imola et al. (2023), and SBSC Khan et al. (2025); (2) *deep clustering*, including DeepCluster Caron et al. (2018), TableDC Rauf et al. (2025), SpEx Argov & Wagner (2025), and MiniClustering Li et al. (2026); and (3) *granular-ball-based methods*, including Ball k-means Xia et al. (2022), MGNR Xie et al. (2024), GBCT Xia et al. (2025a), and GBK-DPC Ni et al. (2026). In particular, SpEx was originally proposed as a spectral explainable clustering method; in this work, we enhance it with deep learning and therefore categorize the enhanced variant as a deep clustering method. For image segmentation tasks, we compare against DFKM Zhang et al. (2019), FLRSC Song et al. (2023), MK-SPCM Yu et al. (2024), SpEx Argov & Wagner (2025), GBCT Xia et al. (2025a), AWRFCM Jing et al. (2026), and SegGBC Chong et al. (2026).

Implementation Details. GRANITE is implemented in PyTorch and trained on an NVIDIA RTX 3060 GPU. Network parameters of GB-level autoencoder are updated using SGD; cluster prototypes, intuitionistic assignments, and the covariance metric are updated in alternating blocks. The main hyperparameters are set to $\varepsilon = 1$, $\lambda_1 = 0.5$, and $\lambda_2 = 0.0001$, where ε is the smoothing parameter of the covariance-aware distribution loss, λ_1 controls the negative-entropy regularization, and λ_2 controls the L_2 regularization of the network parameters. To ensure reproducibility, a fixed random seed of 42 is used, and all reported results are averaged over five independent runs.

E.2 Clustering Experiment

Fig. 8 presents additional t-SNE visualizations on the eight UCI datasets, complementing the qualitative comparisons in the main text. These datasets span different sample sizes, feature dimensions, and cluster structures, providing diverse settings for examining the learned representations. The visualizations highlight within-class compactness, inter-class overlap, and the arrangement of samples near ambiguous boundaries.

The qualitative results show that GRANITE achieves the highest ACC, NMI, and ARI among the compared methods on Iris, Seed, and WDBC. Its performance on the other datasets varies across metrics, suggesting that improvements in clustering accuracy do not always coincide with the strongest agreement under every evaluation criterion. The t-SNE projections provide a complementary perspective for examining class overlap and local organization that aggregate scores alone cannot fully describe.

The learned representations reflect the joint effects of reconstruction and clustering. The reconstruction objective encourages the embeddings to preserve information from the GB descriptors, while the clustering objective organizes them through prototype fitting and intuitionistic triplet evidence. This interpretation is also supported by the ablation results on Segment, where retaining both objectives yields higher ACC and NMI than removing either one.

E.3 Analysis of the Convergence Process

Fig. 9 presents the training curves of GRANITE on the eight UCI datasets, complementing the process analysis in the main text. These curves illustrate how the objective evolves from the initial optimization stage to later training and allow comparison of the rate of decrease, the approach to a plateau, and fluctuations across datasets with different sizes and structures.

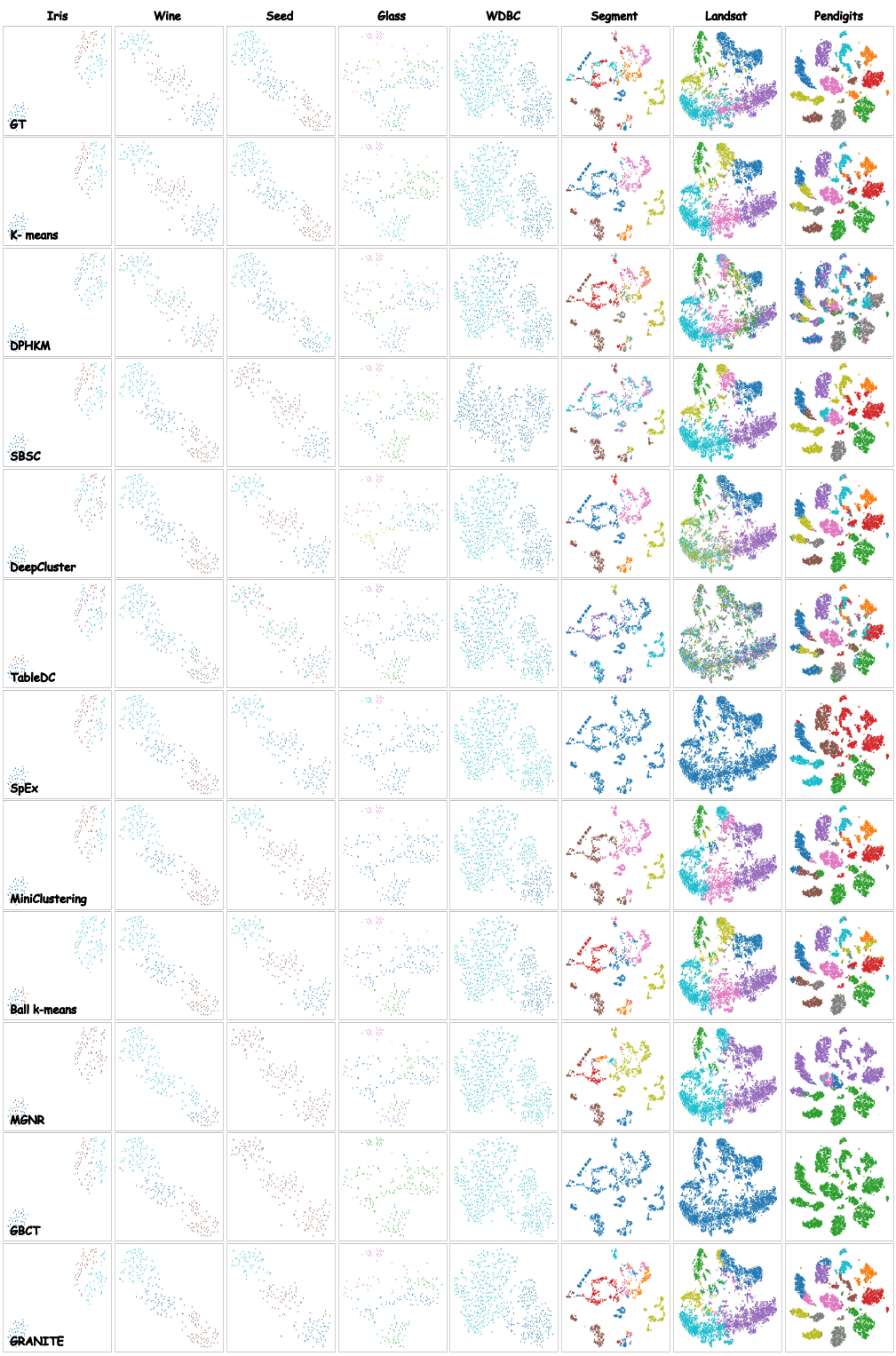


Figure 8: Qualitative comparison on eight UCI datasets using t-SNE.

The observed trajectories reflect the joint effects of reconstruction, robust residual fitting, and regularization. A sustained decrease followed by a plateau suggests that the monitored objective has stabilized, while fluctuations may arise from stochastic updates and changes in GB descriptors during dynamic reconstruction. These empirical observations complement the stage-wise analysis: the

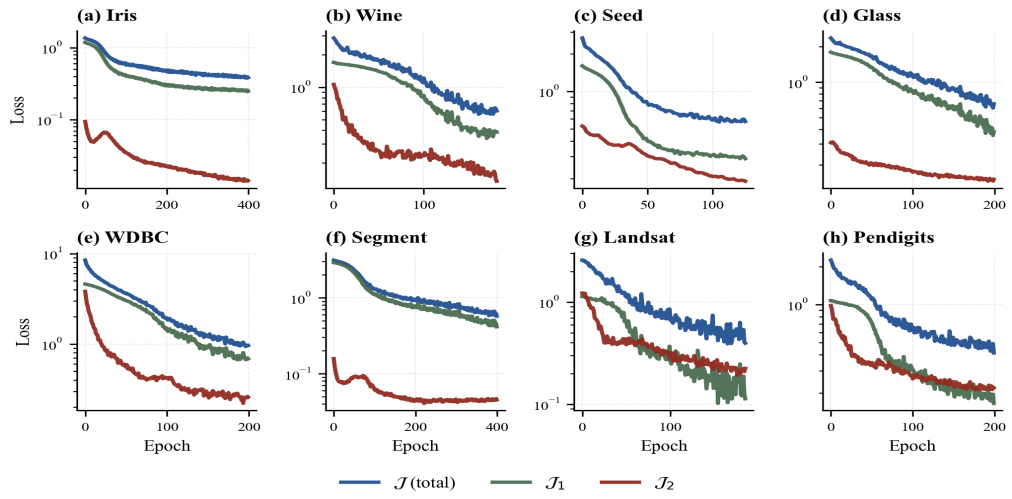


Figure 9: Training objective curves of GRANITE on eight UCI datasets.

theoretical descent guarantee applies when accepted updates decrease the same surrogate within a stage, whereas the curves characterize the behavior of the complete training procedure.

E.4 Image Segmentation Experiment

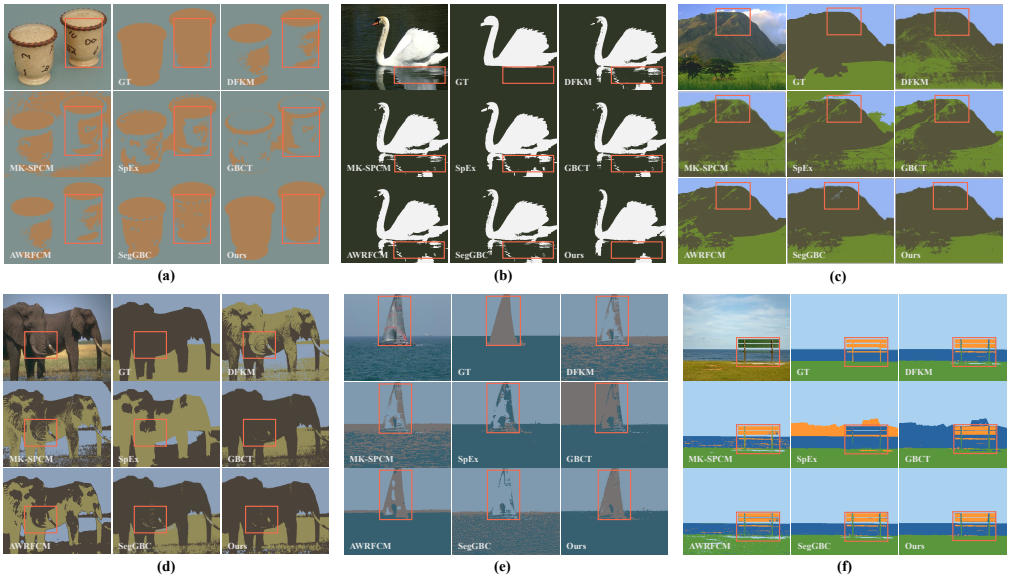


Figure 10: Additional qualitative comparisons of segmentation results on BSD500 and DUTS.

Fig. 10 provides additional qualitative comparisons on BSD500 and DUTS, complementing the aggregate segmentation results reported in the main text. The comparisons focus on agreement with the annotated regions, preservation of fine structures, boundary delineation, and consistency within object regions. These aspects help reveal local segmentation behavior that may not be fully reflected by image-level or dataset-level average scores.

For GRANITE, segmentation is obtained by assigning cluster labels to GBs and propagating those labels to their constituent pixels. The final segmentation is obtained by propagating the learned ball-level labels to pixels according to the sample-to-ball mapping after the last reconstruction. Regions containing small structures, similar foreground and background appearances, or uneven illumination are particularly informative for evaluating this formulation. Since pixels within the same ball receive a common label, mixed-content balls can limit boundary accuracy even when the ball-level objective

is effectively optimized. These comparisons consequently assess both the applicability of GRANITE to segmentation and the remaining need for explicit spatial and boundary-aware modeling.

E.5 Time Complexity Analysis

We analyze the computational complexity of GRANITE and compare it with representative clustering baselines. Let N denote the number of samples, d_0 the input dimension, N_G the number of generated GBs, d_b the latent dimension of GB centers, and K the number of target clusters. For GRANITE, let T denote the number of outer optimization iterations, $R_G \leq T$ the number of reconstruction steps during training (at most one per sweep), J the maximum number of inner assignment updates per outer iteration, and R the maximum number of additional objective evaluations required by backtracking. We use $\mathcal{F}_G$ to denote the cost of a forward and backward pass through the GB autoencoder for one GB. For a fully connected architecture with layer widths $h_0, \ldots, h_L$, this cost satisfies

$$\mathcal{F}_G = \mathcal{O}\left(\sum_{\ell=1}^{L} h_{\ell-1} h_\ell\right). \tag{79}$$

The following analysis accounts for reconstruction during training and assumes conventional dense matrix operations.

Granular-ball construction. Let $\mathcal{C}_{\mathrm{GB}}$ denote an upper bound on the cost of one ball construction or reconstruction and descriptor update. This cost depends on the construction algorithm, its stopping criterion, and the resulting partition. Computing centers and radii for an already available partition requires $\mathcal{O}(Nd_0)$ operations. We retain $\mathcal{C}_{\mathrm{GB}}$ explicitly rather than assuming a universal construction complexity. In particular, compression to N_G GBs does not eliminate the initial cost of processing the original samples.

Alternating optimization. Each outer iteration updates the surrogate coefficients, cluster prototypes, intuitionistic assignments, covariance matrix, and autoencoder parameters. Evaluating the residuals $\mathbf{v}_i = \mathbf{c}_i^{(b)} - \mathbf{H}\mathbf{w}_i$ requires $\mathcal{O}(N_G d_b K)$ operations, while evaluating their quadratic forms under a dense covariance metric requires $\mathcal{O}(N_G d_b^2)$ operations once the metric factorization is available.

For the prototype update, forming $\mathbf{M} = \mathbf{W}\mathbf{A}\mathbf{W}^\top$ and $\mathbf{B} = \mathbf{C}\mathbf{A}\mathbf{W}^\top$ costs $\mathcal{O}(N_G K^2)$ and $\mathcal{O}(N_G d_b K)$, respectively. The spectral solution of the Sylvester equation involves eigendecompositions of matrices of sizes $d_b \times d_b$ and $K \times K$. For the assignment update, a direct gradient evaluation for one GB costs $\mathcal{O}(d_b K + d_b^2)$; the normalization of its K intuitionistic triplets adds only $\mathcal{O}(K)$ operations. The covariance update forms a weighted residual scatter matrix and performs its eigendecomposition. Consequently, the principal block costs are

$$\mathcal{C}_H = \mathcal{O}\Big(N_G K^2 + N_G d_b K + d_b^3 + K^3 + d_b^2 K + d_b K^2\Big), \tag{80}$$

$$\mathcal{C}_P = \mathcal{O}\Big(J N_G (d_b K + d_b^2)\Big), \tag{81}$$

$$\mathcal{C}_\Sigma = \mathcal{O}\Big(N_G d_b K + N_G d_b^2 + d_b^3\Big), \tag{82}$$

$$\mathcal{C}_{\mathrm{AE}} = \mathcal{O}\Big((R+1) N_G \big(\mathcal{F}_G + d_b K + d_b^2\big)\Big). \tag{83}$$

Here, the assignment bound corresponds to direct residual-based gradient evaluation; implementations that cache suitable Gram matrices may have different operation counts. The autoencoder bound includes both the gradient computation and the additional forward evaluations used by backtracking. Matrix factorizations are reused whenever the corresponding covariance matrix remains fixed.

Since $d_b^2 K + d_b K^2 = \mathcal{O}(d_b^3 + K^3)$, the cost of one complete outer iteration is $\mathcal{O}(\mathcal{C}_{\mathrm{step}})$, where

$$\mathcal{C}_{\mathrm{step}} = (R+1) N_G \mathcal{F}_G + N_G K^2 + (J+R+1) N_G (d_b K + d_b^2) + d_b^3 + K^3. \tag{84}$$

Assigning a final cluster label to each GB costs $\mathcal{O}(N_G K)$, and propagating these labels through the stored sample-to-ball mapping costs $\mathcal{O}(N)$. The former term is absorbed by the optimization bound for $T \geq 1$. Therefore, the overall complexity is

$$\mathcal{T}_{\mathrm{GRANITE}} = \mathcal{O}\left((R_G+1)\mathcal{C}_{\mathrm{GB}} + T\mathcal{C}_{\mathrm{step}} + N\right). \tag{85}$$

Table 5: Computational complexity of GRANITE and representative baselines. The entries describe the core algorithms under the stated conditions. Separate pretraining, external feature extraction, and additional initialization costs must be added when used. Reported bounds marked by † retain the assumptions of the corresponding original analyses.

Method	Time complexity	Basis or condition
k-means Lloyd (1982)	$\mathcal{O}(T_{\mathrm{KM}} N K d_0)$	Lloyd updates
Ball k-means Xia et al. (2022)	$\mathcal{O}\big(T_{\mathrm{BKM}}[NKd_0 + K^2 d_0 + K^2 \log K]\big)$	Worst-case upper bound
SBSC Khan et al. (2025)	$\mathcal{O}\big(N(K^2 + \log N)\big)^{\dagger}$	Reported bound; dimension suppressed
SpEx-Clique Argov & Wagner (2025)	$\mathcal{O}(\mathcal{C}_{\mathrm{ref}} + KNd_0 \log N)$	Reference clustering and tree construction
SpEx-kNN Argov & Wagner (2025)	$\mathcal{O}(\mathcal{C}_{k\mathrm{NN}} + KNd_0(\log N + \kappa))$	κ neighbors per sample
DeepCluster Caron et al. (2018)	$\mathcal{O}(T_{\mathrm{DC}}[N\mathcal{F}_{\mathrm{DC}} + I_{\mathrm{KM}} N K d_{\mathrm{DC}} + \mathcal{C}_{\mathrm{proj}}])$	Alternating feature learning and clustering
TableDC Rauf et al. (2025)	$\mathcal{O}(T_{\mathrm{Tab}} N[\mathcal{F}_{\mathrm{Tab}} + K d_{\mathrm{Tab}}])$	Efficient scaled-identity metric evaluation
MiniClustering Li et al. (2026)	$\mathcal{O}(T_{\mathrm{Mini}}[N\mathcal{F}_{\mathrm{Mini}} + NB(K+M) + \lceil N/B \rceil KM])$	Two prediction heads and batchwise alignment
GBCT Xia et al. (2025a)	$\mathcal{O}\big(N^{3/2} + m_{\mathrm{GBCT}}^2 \log m_{\mathrm{GBCT}}\big)^{\dagger}$	Reported average-case analysis
GRANITE	$\mathcal{O}((R_G+1)\mathcal{C}_{\mathrm{GB}} + T\mathcal{C}_{\mathrm{step}} + N)$	Includes dynamic reconstruction

When J and R are bounded constants, this becomes

$$\mathcal{T}_{\mathrm{GRANITE}} = \mathcal{O}\Big((R_G+1)\mathcal{C}_{\mathrm{GB}} + N + T\big[N_G(\mathcal{F}_G + K^2 + d_b K + d_b^2) + d_b^3 + K^3\big]\Big). \tag{86}$$

Comparison with representative methods. Table 5 summarizes the main computational costs of representative baselines. For each iterative baseline, T_a denotes its own number of iterations or training epochs; these counts are not assumed to be equal across methods. For neural methods, $\mathcal{F}_a$ denotes the per-sample forward and backward cost of the corresponding network, including its prediction heads, and d_a denotes its clustering-space dimension. The bounds describe arithmetic costs rather than hardware-dependent execution times.

For SpEx, $\mathcal{C}_{\mathrm{ref}}$ is the cost of obtaining the reference clustering, and $\mathcal{C}_{k\mathrm{NN}}$ is the cost of constructing the nearest-neighbor graph. The two variants are listed separately because their preprocessing requirements differ. For DeepCluster, I_{KM} is the number of k-means iterations per clustering round, and $\mathcal{C}_{\mathrm{proj}}$ accounts for feature projection or whitening when performed in that round. For TableDC, the listed bound exploits its scaled-identity covariance; a generic dense implementation can incur additional costs. For MiniClustering, B is the batch size and M is the number of mini-clusters. Its batchwise similarity alignment requires $\mathcal{O}(B^2(K+M))$ operations per batch, while constructing the target-to-mini-cluster association statistics can require an additional $\mathcal{O}(KM)$ per batch. The quantity m_{GBCT} is the number of GBs generated by GBCT and need not equal N_G. The reported SBSC and GBCT expressions suppress feature dimension and other quantities treated as constants in their original analyses; the GBCT expression additionally relies on average-case splitting and merging assumptions.

Computational Advantages of GRANITE. GRANITE performs representation learning and clustering at the GB level, reducing the number of units involved in these operations from N samples to N_G descriptors. When $N_G \ll N$, this compression reduces the cost of network evaluation and ball–cluster assignment updates. For fixed network architecture, latent dimension, cluster count, and inner iteration budgets, the per-sweep optimization cost is linear in N_G. Dynamic reconstruction further allows the granular representation to adapt during training, with its cost explicitly accounted for by $(R_G+1)\mathcal{C}_{\mathrm{GB}}$. GRANITE therefore combines compact ball-level optimization with adaptive representation updates, offering computational advantages when substantial compression is achieved and reconstruction costs remain moderate.