AM-GMAE: ADAPTIVE MANIFOLD GRAPH MASKED AUTOENCODER FOR MULTI-VIEW REMOTE SENSING DATA CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-view clustering has emerged as a mainstream unsupervised paradigm for analyzing remote sensing data by jointly exploiting complementary multi-sensor observations. Graph Masked Autoencoders (GMAEs), which learn informative graph representations through self-supervised reconstruction, offer a promising solution for these multi-view scenarios. Yet, effectively extending GMAEs to heterogeneous multi-view remote sensing data remains challenging for three main reasons: i) existing masking strategies ignore node importance, risking trivial cross-view reconstruction shortcuts; ii) initial similarity-based graphs suffer from noise and class ambiguity, distorting the underlying manifold during message passing; and iii) forced alignment of heterogeneous latent spaces often induces negative transfer. To address these issues, we propose the Adaptive Manifold Graph Masked Autoencoder (AM-GMAE). Specifically, AM-GMAE introduces a Clustering-Saliency Adversarial Mask module to identify and adaptively mask nodes with high clustering-uncertainty sensitivity across correlated views, thereby increasing reconstruction difficulty and reducing reliance on trivial cross-view cues. Furthermore, we develop a Dynamic Graph Purification and Topology Optimization module to prune misleading edges and refine the graph manifold using structural information-theoretic constraints. Finally, a Multi-marginal Optimal Transport module with Wasserstein barycenter estimation is employed to align heterogeneous representations through a shared Wasserstein semantic geometry without imposing rigid pairwise coupling. Extensive experiments on four widely used remote sensing benchmarks demonstrate that AM-GMAE outperforms SOTA methods.

1 INTRODUCTION

Multi-view remote sensing data clustering aims to partition unlabeled land-cover samples by jointly exploiting complementary observations of the same geographical scene, and has become an important unsupervised paradigm for Earth observation analysis (Zhang et al., 2026c; Guan et al., 2024a; Cai et al., 2023). Modern platforms commonly acquire optical, hyperspectral, SAR, and LiDAR observations over the same region (Zhang et al., 2026b; Guan et al., 2025d), providing complementary yet heterogeneous descriptions of distinct physical properties (Liu, 2026; Liu et al., 2025b). Meanwhile, remote sensing scenes exhibit strong spatial regularity, with neighboring regions often sharing similar semantics. Therefore, effective multi-view clustering must jointly exploit cross-view complementarity while preserving spatial and semantic structures.

Graph-based modeling provides a natural way to capture such spatial and relational structures (Zhang et al., 2026c). By representing pixels or homogeneous regions as nodes and their relationships as edges, graph-based multi-view clustering (MVGC) can model both intra-view dependencies and cross-view complementarity. Existing MVGC methods have evolved from shallow graph learning to deep graph representation learning. Shallow methods typically construct view-specific similarity graphs and integrate them through graph fusion, consensus graph learning, or spectral optimization (Tang et al., 2026; Zhang et al., 2025c), but their representation capacity is often limited by the original feature space. More recently, deep MVGC methods employ graph neural networks (GNNs) to propagate neighborhood information and learn nonlinear representations

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.

for complex remote sensing scenes (Peng et al., 2025; Guan et al., 2025d; 2024b). However, their performance still depends on reliable input graphs and effective unsupervised learning objectives.

To learn discriminative representations without labels, self-supervised learning has been widely explored (Moujahid & Dornaika, 2025; Gui et al., 2024). Graph contrastive learning is a common choice, but reliable positive and negative pairs are difficult to define for ambiguous samples in unsupervised clustering (Yao et al., 2025). Masked reconstruction avoids explicit pair construction by learning to recover missing information from graph context. Graph Masked Autoencoders (GMAEs) follow this paradigm by masking node attributes and reconstructing them from neighboring information (Hou et al., 2022; Tan et al., 2023; Tian et al., 2023; Li et al., 2023), thereby capturing semantic and structural dependencies. This property is particularly attractive for multi-view remote sensing data, where complementary views provide additional contextual cues for reconstruction (Guan et al., 2025c; Wang et al., 2024). However, existing GMAEs are mainly designed for single-view graphs (Fu et al., 2024) and, without clustering supervision, cannot explicitly identify information that is most relevant to cluster discovery (Gupta et al., 2025; Guan et al., 2026; Zhu et al., 2026).

Extending GMAEs to multi-view remote sensing clustering therefore faces three key challenges. First, existing masking strategies largely ignore clustering relevance, so sensitive information may remain visible across correlated views and enable trivial reconstruction shortcuts (Guan et al., 2026). Second, similarity-based graphs are easily corrupted by noisy or ambiguous samples, and message passing over spurious cross-class edges can distort the underlying manifold (Zhuang et al., 2026; Ju et al., 2025). Third, heterogeneous sensing modalities induce different latent distributions, making direct fusion or rigid pairwise alignment prone to semantic mismatch and negative transfer (Maiorca et al., 2023). These challenges call for a unified framework that jointly considers clustering-aware masking, reliable graph refinement, and geometry-preserving cross-view interaction.

To address these challenges, we propose an **Adaptive Manifold Graph Masked AutoEncoder** (AM-GMAE) for multi-view remote sensing clustering. First, a Clustering-Saliency Adversarial Mask (CSAM) identifies clustering-sensitive nodes and coordinates their masking across correlated views, strengthening masked reconstruction while suppressing trivial cross-view shortcuts. Second, Dynamic Graph Purification and Topology Optimization (DGPO) refines unreliable graph connectivity and optimizes the manifold through structure- and topology-consistent constraints. Finally, Multi-marginal Optimal Transport (MOT) with Wasserstein barycenter estimation aligns heterogeneous representations through a shared Wasserstein semantic geometry without imposing rigid pairwise correspondence, thereby reducing negative transfer while preserving cross-view complementarity. Extensive experiments and visualization results on multiple remote sensing datasets demonstrate the effectiveness of AM-GMAE. The main contributions are summarized as follows:

1. **Clustering-aware adaptive masking:** We propose CSAM to adaptively coordinate clustering-sensitive masking across correlated views, suppressing cross-view shortcuts.
2. **Dynamic manifold refinement:** We develop DGPO to purify graph connectivity and optimize topology, yielding a reliable manifold for masked reconstruction.
3. **Geometry-aware cross-view alignment:** We introduce MOT with Wasserstein barycenter estimation to align heterogeneous views through shared Wasserstein semantic geometry while preserving view-specific geometric structures, thereby reducing negative transfer.

2 Related Work

Multi-View Graph Clustering for Remote Sensing Data. Multi-view graph clustering (MVGC) efficiently handles remote sensing tasks by integrating multi-modal data and complex spatial structures (Guan et al., 2025a;b; Zhang et al., 2025c). Recent advancements have transitioned from limited traditional methods (Moujahid & Dornaika, 2025; Liu et al., 2025a; Feng et al., 2025) to GNN-based models (Guan et al., 2025d;c). These deep models excel by adaptively optimizing graph topologies, cluster structures, and underlying relations (Guan et al., 2025d;a; Peng et al., 2025). However, current GNN approaches in remote sensing are bottlenecked by contrastive class conflicts and unsupervised reconstruction difficulties (Moujahid & Dornaika, 2025; Wu et al., 2025; Guan et al., 2024b). Graph Masked Autoencoders (GMAE) offer a promising alternative by replacing contrastive sample pairs with masked reconstruction, thereby avoiding class conflicts while preserving the ability to learn graph structures Guan et al. (2026). Nevertheless, how to extend GMAE to multi-view remote sensing data and exploit information across views remains largely unexplored.

Graph Masked Autoencoders. GMAEs extend Graph Autoencoders with masking for robust representation learning (Hou et al., 2022; Shi et al., 2023). Existing methods adopt random masking (Hou et al., 2022; 2023; Tan et al., 2023; 2022), structure-guided masking (Zhao et al., 2024; Li et al., 2023; Liu et al., 2024), adaptive masking (Tian et al., 2023; Li et al., 2025; Shi & Pu, 2025), adversarial masking (Wang et al., 2024; Tian et al., 2024; Shen et al., 2025), and knowledge-guided masking (Zhang et al., 2025b; 2026a). Random masking is efficient but ignores differences among nodes, while structure-aware and adaptive methods require additional structural scoring, mask generation, or difficulty scheduling (Liu et al., 2024; Tian et al., 2023). Discrete structure masking may disrupt message propagation and informative neighborhood relationships (Zhao et al., 2024). Adversarial methods construct harder training instances, whereas knowledge-guided methods introduce structural, semantic, or domain-specific priors (Wang et al., 2024; Zhang et al., 2025b). Recent studies have also explored reconstruction difficulty, dynamic masking, and clustering-guided masking (Li et al., 2025; Zhang et al., 2026a; Guan et al., 2026). Nevertheless, existing methods do not coordinate masking across views to jointly model uncertainty and suppress cross-view redundancy, which may cause redundant masking or insufficient perturbation of ambiguous regions, limiting complementary information utilization while potentially damaging local manifold structure.

3 METHODS

In this section, we present the detailed methodology of the proposed AM-GMAE, whose overall architectural framework is illustrated in Figure 1.

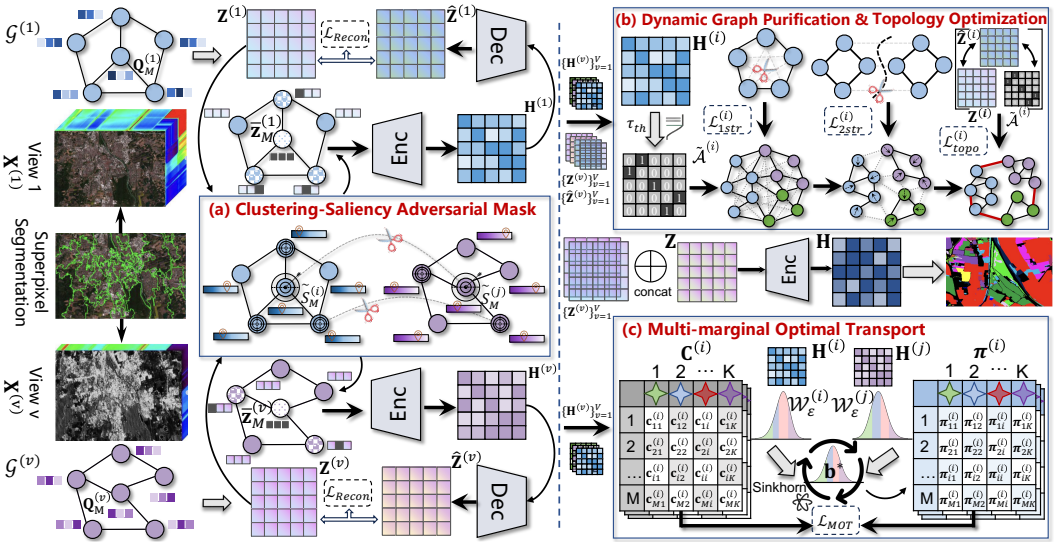


Figure 1: The overview of the proposed AM-GMAE. Multi-view remote sensing data is processed by three modules: (a) Clustering-Saliency Adversarial Mask for node feature reconstruction ($\mathcal{L}_{Recon}$); (b) Dynamic Graph Purification & Topology Optimization for refining adjacency matrices ($\mathcal{L}_{1str}$, $\mathcal{L}_{2str}$, $\mathcal{L}_{topo}$); and (c) Multi-marginal Optimal Transport for cross-view alignment ($\mathcal{L}_{MOT}$).

3.1 PRELIMINARIES

Consider a multiview remote sensing dataset $\mathcal{X} = \{\mathbf{X}^{(v)}\}_{v=1}^{V}$, where $\mathbf{X}^{(v)} \in \mathbb{R}^{W\times H\times D^{(v)}}$ contains $N = W \times H$ pixels and $D^{(v)}$ spectral channels. The spatial domain is partitioned by SLIC on the channel-wise concatenated views into $M \ll N$ disjoint superpixel regions $\mathcal{R} = \{\mathcal{R}_i\}_{i=1}^{M}$, satisfying $\mathcal{R}_i \cap \mathcal{R}_j = \emptyset$ for $i \neq j$ and $\sum_{i=1}^{M} |\mathcal{R}_i| = N$. Treating the regions as nodes, we obtain $\mathbf{Z}^{(v)} = [\mathbf{z}_1^{(v)}, \ldots, \mathbf{z}_M^{(v)}]^\top \in \mathbb{R}^{M\times D^{(v)}}$ via region-wise average pooling, with $\mathbf{z}_i^{(v)} = \frac{1}{|\mathcal{R}_i|}\sum_{\mathbf{x}\in\mathcal{R}_i}\mathbf{x}^{(v)}$. We construct an undirected k-nearest neighbor (k-NN) graph $\mathcal{G}^{(v)} = (\mathcal{V}^{(v)}, \mathcal{E}^{(v)})$, where $\mathcal{V}^{(v)} =$

$\{\mathbf{v}_i^{(v)}\}_{i=1}^M$. The pairwise Gaussian affinity is defined as $\mathbf{S}_{ij}^{(v)} = \exp\left(-\frac{\|\mathbf{z}_i^{(v)}-\mathbf{z}_j^{(v)}\|_2^2}{2\sigma^2}\right)$, where σ is the kernel bandwidth. Based on $\mathbf{S}^{(v)}$, $\mathbf{A}_{ij}^{(v)} = 1$ if $\mathbf{v}_j^{(v)} \in \mathrm{kNN}(\mathbf{v}_i^{(v)})$ or $\mathbf{v}_i^{(v)} \in \mathrm{kNN}(\mathbf{v}_j^{(v)})$, and 0 otherwise. For message passing, $\tilde{\mathbf{A}}^{(v)} = (\mathbf{D}^{(v)})^{-\frac{1}{2}}(\mathbf{A}^{(v)} + \mathbf{I})(\mathbf{D}^{(v)})^{-\frac{1}{2}}$, where $\mathbf{I}$ is the identity matrix and $\mathbf{D}^{(v)}$ is diagonal with $\mathbf{D}_{ii}^{(v)} = \sum_{j=1}^M (\mathbf{A}_{ij}^{(v)} + \mathbf{I}_{ij})$. The GMAE learns the mapping $f : \{\mathbf{Z}^{(v)}, \tilde{\mathbf{A}}^{(v)}\}_{v=1}^V \mapsto \{\mathbf{H}^{(v)}\}_{v=1}^V$ to obtain discriminative $\mathbf{H}^{(v)} \in \mathbb{R}^{M\times d}$ for K-way node clustering. Before masking, an L-layer GCN provides structural and semantic priors, producing $\tilde{\mathbf{H}}^{(v)} = \tilde{\mathbf{H}}_{(L)}^{(v)} \in \mathbb{R}^{M\times d}$ via $\tilde{\mathbf{H}}_{(l)}^{(v)} = \mathrm{ReLU}(\tilde{\mathbf{A}}^{(v)}\tilde{\mathbf{H}}_{(l-1)}^{(v)}\mathbf{W}_{(l)}^{(v)})$ with $\tilde{\mathbf{H}}_{(0)}^{(v)} = \mathbf{Z}^{(v)}$. The resulting $\mathbf{Q}^{(v)} \in [0,1]^{M\times K}$ is obtained by temperature-scaled softmax over cosine similarities between $\tilde{\mathbf{h}}_i^{(v)}$ and $\boldsymbol{\mu}_k^{(v)}$:

$$\mathbf{Q}_{ik}^{(v)} = \frac{\exp\left(\frac{1}{\tau_c}\frac{(\tilde{\mathbf{h}}_i^{(v)})^\top \boldsymbol{\mu}_k^{(v)}}{\|\tilde{\mathbf{h}}_i^{(v)}\|_2\|\boldsymbol{\mu}_k^{(v)}\|_2}\right)}{\sum_{j=1}^K \exp\left(\frac{1}{\tau_c}\frac{(\tilde{\mathbf{h}}_i^{(v)})^\top \boldsymbol{\mu}_j^{(v)}}{\|\tilde{\mathbf{h}}_i^{(v)}\|_2\|\boldsymbol{\mu}_j^{(v)}\|_2}\right)}, \tag{1}$$

where $\Delta^{K-1} = \{\mathbf{q} \in \mathbb{R}^K \mid \sum_{k=1}^K q_k = 1, q_k \geq 0\}$ is the standard $(K-1)$-dimensional probability simplex, $\tau_c > 0$ is the temperature, and $\boldsymbol{\mu}^{(v)} = [\boldsymbol{\mu}_1^{(v)}, \ldots, \boldsymbol{\mu}_K^{(v)}]^\top \in \mathbb{R}^{K\times d}$ denotes the view-specific cluster prototypes jointly optimized with the network.

3.2 Clustering-Saliency Adversarial Mask

Existing GMAEs for multi-view suffer from suboptimal masking and alignment. Random masking ignores node importance and may permit trivial reconstruction shortcuts, whereas excessively aggressive adversarial masking may disrupt informative local neighborhoods; forced cross-view alignment under semantic conflicts may also induce negative transfer. We therefore propose the Clustering-Saliency Adversarial Mask (CSAM), which quantifies the local sensitivity of clustering uncertainty and dynamically masks clustering-sensitive nodes across correlated views, thereby increasing reconstruction difficulty while preserving the graph structure. Soft clustering assignments obtained from the unmasked representations are first used to characterize the perturbation sensitivity of node-wise clustering uncertainty.

Definition 1. *Given $\mathbf{Q}_i^{(v)} \in \Delta^{K-1}$ for node i, and a differentiable mapping $\tilde{\mathbf{h}}_i^{(v)} \mapsto \mathbf{Q}_i^{(v)}$, the Shannon entropy $\mathcal{H}_i^{(v)} : \Delta^{K-1} \to \mathbb{R}_{\geq 0}$ and clustering-saliency score $S_i^{(v)} : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ are:*

$$\mathcal{H}_i^{(v)} = -\sum_{k=1}^K \mathbf{Q}_{ik}^{(v)} \log \mathbf{Q}_{ik}^{(v)}, \qquad S_i^{(v)} \triangleq \left\|\nabla_{\tilde{\mathbf{h}}_i^{(v)}} \mathcal{H}_i^{(v)}\right\|_2 = \left\|\mathcal{J}_{\tilde{\mathbf{h}}_i^{(v)}}^\top(\mathbf{Q}_i^{(v)}) \log \mathbf{Q}_i^{(v)}\right\|_2. \tag{2}$$

Here $\mathcal{H}_i^{(v)} \in [0, \log K]$ is concave on the simplex interior, while $\mathcal{J}_{\tilde{\mathbf{h}}_i^{(v)}}(\mathbf{Q}_i^{(v)}) \in \mathbb{R}^{K\times d}$ is the Jacobian at $\tilde{\mathbf{h}}_i^{(v)}$. Under bounded representation norms and smooth prototype-based assignments, $S_i^{(v)}$ is bounded and vanishes as the assignment approaches either a one-hot or a perfectly uniform distribution. Therefore, $S_i^{(v)}$ quantifies the local sensitivity of clustering uncertainty to perturbations of the latent representation: to first order, its magnitude determines the maximum entropy variation induced by a norm-bounded perturbation. Since clustering-sensitive information visible in correlated views may facilitate cross-view reconstruction, we further calibrate the saliency using inter-view representation correlation:

$$\tilde{S}_i^{(v)} = S_i^{(v)} + \eta \sum_{u\neq v} \max\left(0, \cos(\tilde{\mathbf{h}}_i^{(v)}, \tilde{\mathbf{h}}_i^{(u)})\right) \cdot S_i^{(u)}, \tag{3}$$

where η controls the penalty strength and emphasizes features that are salient across views. Boltzmann sampling with temperature τ_m then determines the node-level selection probability for each masking draw. We repeatedly sample nodes according to these probabilities to form $\tilde{\mathcal{V}}^{(v)}$, with $\mathbf{m}_i^{(v)} = 0$ for $\mathbf{v}_i^{(v)} \in \tilde{\mathcal{V}}^{(v)}$ and 1 otherwise, so that the masked input follows the sampling event:

$$p_i^{(v)} = \frac{\exp(\tilde{S}_i^{(v)}/\tau_m)}{\sum_{j=1}^M \exp(\tilde{S}_j^{(v)}/\tau_m)}, \qquad \overline{\mathbf{z}}_i^{(v)} = \mathbf{m}_i^{(v)}\mathbf{z}_i^{(v)} + (1-\mathbf{m}_i^{(v)})\mathbf{z}_{[\mathrm{MASK}]}^{(v)}, \tag{4}$$

where $\mathbf{z}_{[\text{MASK}]}^{(v)} \in \mathbb{R}^{D^{(v)}}$ is a shared, view-specific learnable token, and $p_i^{(v)}$ denotes the probability of selecting node $\mathbf{v}_i^{(v)}$ in a single masking draw. Repeated sampling without replacement using these Boltzmann probabilities produces the final masked set $\tilde{\mathcal{V}}^{(v)}$. Given $\overline{\mathbf{Z}}^{(v)} = [\overline{\mathbf{z}}_1^{(v)}, \ldots, \overline{\mathbf{z}}_M^{(v)}]^\top$, the graph encoder reconstructs masked attributes via layer-wise propagation $\mathbf{H}_{(l)}^{(v)} = \text{ReLU}(\tilde{\mathbf{A}}^{(v)}\mathbf{H}_{(l-1)}^{(v)}\mathbf{\Theta}_{(l)}^{(v)})$, with $\mathbf{H}_{(0)}^{(v)} = \overline{\mathbf{Z}}^{(v)}$. A symmetric decoder then maps $\mathbf{H}^{(v)} = \mathbf{H}_{(L)}^{(v)}$ to $\hat{\mathbf{Z}}^{(v)}$. To suppress trivial restoration and promote semantically robust embeddings, we introduce $\gamma \geq 1$ to control the sensitivity of the reconstruction loss, which is optimized only on masked nodes as:

$$\mathcal{L}_{Recon} = \sum_{v=1}^{V} \frac{1}{|\tilde{\mathcal{V}}^{(v)}|} \sum_{\mathbf{v}_i^{(v)} \in \tilde{\mathcal{V}}^{(v)}} \left(1 - \frac{(\mathbf{z}_i^{(v)})^\top \hat{\mathbf{z}}_i^{(v)}}{\|\mathbf{z}_i^{(v)}\|_2 \|\hat{\mathbf{z}}_i^{(v)}\|_2}\right)^\gamma, \tag{5}$$

3.3 Dynamic Graph Purification and Topology Optimization

While CSAM masks salient nodes, the initial static k-NN graph $\mathbf{A}^{(v)}$ remains vulnerable to over-smoothing and spurious cross-cluster connections, making message-passing recovery ill-posed. We therefore propose a pseudo-label-free Dynamic Graph Purification and Topology Optimization (DGPO) module to prune misleading edges and construct a cohesive, loosely coupled manifold that guides reconstruction of CSAM-induced semantic gaps while reducing geometric noise.

Dynamic Graph Construction. We reconstruct a dynamic adjacency matrix $\tilde{\mathcal{A}}^{(v)} \in \mathbb{R}^{M \times M}$ from $\mathbf{H}^{(v)}$. With a view-shared learnable threshold $\tau_{th} \in [0, 1)$, the purified edge weights are defined as

$$\tilde{\mathcal{A}}_{ij}^{(v)} = \mathbb{I}(i \neq j) \cdot \max\left(0, \frac{(\mathbf{h}_i^{(v)})^\top \mathbf{h}_j^{(v)}}{\|\mathbf{h}_i^{(v)}\|_2 \|\mathbf{h}_j^{(v)}\|_2} - \tau_{th}\right), \tag{6}$$

where $\mathbb{I}(\cdot)$ removes self-loops to focus on inter-node dependencies. If $d_i^{(v)} = \sum_j \tilde{\mathcal{A}}_{ij}^{(v)} = 0$, we connect node i to its most similar neighbor $j_i^{*(v)} = \arg\max_{j \neq i} \cos(\mathbf{h}_i^{(v)}, \mathbf{h}_j^{(v)})$ by setting $\tilde{\mathcal{A}}_{ij_i^{*(v)}}^{(v)} = \tilde{\mathcal{A}}_{j_i^{*(v)} i}^{(v)} = \epsilon_g$, where $\epsilon_g > 0$ is a small constant used to prevent isolated nodes.

Structural Entropy Minimization. To optimize $\tilde{\mathcal{A}}^{(v)}$ without pseudo-label confirmation bias, we impose structural information-theoretic constraints. Let $d_i^{(v)} = \sum_{j=1}^{M} \tilde{\mathcal{A}}_{ij}^{(v)}$ be the dynamic degree and $\Gamma_G^{(v)} = \sum_{i=1}^{M} d_i^{(v)}$ the topological volume. The 1D structural entropy $\mathcal{L}_{1str}^{(v)}$ minimizes node-level random-walk uncertainty and promotes a low-entropy degree organization, while the 2D structural entropy $\mathcal{L}_{2str}^{(v)}$ further enforces semantic boundaries by minimizing inter-cluster cut cost. Reusing the pre-CSAM soft assignment $\mathbf{Q}^{(v)}$, define $\Gamma_k^{(v)} = \sum_{i=1}^{M} \mathbf{Q}_{ik}^{(v)} d_i^{(v)}$ as the semantic volume of cluster k and $m_k^{(v)} = \sum_{i,j} \mathbf{Q}_{ik}^{(v)} \mathbf{Q}_{jk}^{(v)} \tilde{\mathcal{A}}_{ij}^{(v)}$ as its expected intra-cluster edge weight:

$$\mathcal{L}_{1str}^{(v)} = -\sum_{i=1}^{M} \frac{d_i^{(v)}}{\Gamma_G^{(v)}} \log_2\left(\frac{d_i^{(v)}}{\Gamma_G^{(v)}}\right), \qquad \mathcal{L}_{2str}^{(v)} = -\sum_{k=1}^{K} \frac{\Gamma_k^{(v)} - m_k^{(v)}}{\Gamma_G^{(v)}} \log_2\left(\frac{\Gamma_k^{(v)}}{\Gamma_G^{(v)}}\right). \tag{7}$$

The unified structural objective is $\mathcal{L}_{str} = \sum_{v=1}^{V} \left(\mathcal{L}_{1str}^{(v)} + \mathcal{L}_{2str}^{(v)}\right)$. By regularizing degree uncertainty and semantic boundary cuts, $\mathcal{L}_{str}$ encourages $\tilde{\mathcal{A}}^{(v)}$ to form a compact modular structure consistent with the semantic assignments $\mathbf{Q}^{(v)}$.

Topology-Consistent Reconstruction. Although $\mathcal{L}_{Recon}$ preserves feature fidelity on masked nodes, it does not explicitly enforce topological coherence. We therefore constrain $\hat{\mathbf{Z}}^{(v)}$ to approximate the topologically smoothed pristine features under the purified graph. With $\tilde{\mathbf{D}}^{(v)} = \text{diag}(d_1^{(v)}, \ldots, d_M^{(v)})$, the topology-consistent reconstruction is

$$\mathcal{L}_{topo} = \sum_{v=1}^{V} \left\| \hat{\mathbf{Z}}^{(v)} - (\tilde{\mathbf{D}}^{(v)})^{-1} \tilde{\mathcal{A}}^{(v)} \mathbf{Z}^{(v)} \right\|_F^2. \tag{8}$$

Thus, DGPO guides masked attribute reconstruction from purified neighborhoods while enforcing a topology-consistent latent manifold for global semantic alignment.

3.4 MULTI-MARGINAL OPTIMAL TRANSPORT

CSAM and DGPO primarily improve the quality of view-specific representations by increasing reconstruction difficulty and refining unreliable intra-view graph structures, respectively. However, these improvements do not eliminate the inherent distributional and geometric discrepancies among heterogeneous views, whose latent representations may still occupy distinct manifolds. Directly enforcing point-wise or pairwise consistency across such manifolds may distort view-specific geometry and consequently increase the risk of negative transfer. To achieve cross-view integration, we formulate semantic alignment as Multi-marginal Optimal Transport (MOT), which couples multiple view distributions through a shared semantic marginal rather than enforcing direct coordinate-level correspondence, while enabling efficient optimization through entropy-regularized transport.

Geometric Cost Matrix and Barycentric Alignment. To handle modal asymmetry without incurring the $\mathcal{O}(M^2)$ cost of pairwise comparison, we define a sample-to-prototype transport cost. Given K latent cluster prototypes $\boldsymbol{\mu}^{(v)}$, the cost vector $\mathbf{c}_i^{(v)} \in \mathbb{R}^K$ of node $\mathbf{v}_i^{(v)}$ is defined by $\mathbf{c}_{ik}^{(v)} = \left\| \mathbf{h}_i^{(v)} - \boldsymbol{\mu}_k^{(v)} \right\|_2^2$. Exact MOT seeks a joint probability tensor $\boldsymbol{\pi}$ coupling all V views: $\min_{\boldsymbol{\pi}\in\Pi} \sum_{k_1,\ldots,k_V} \boldsymbol{\pi}(k_1,\ldots,k_V)\mathcal{C}(k_1,\ldots,k_V)$, but requires a joint coupling tensor with $\mathcal{O}(K^V)$ entries, becoming computationally prohibitive as V increases. We therefore adopt a Wasserstein barycentric relaxation that couples all views through a shared semantic marginal while preserving their view-specific prototype geometry. Specifically, let $\mathcal{W}_\varepsilon^{(v)}(\mathbf{b}) \triangleq \mathcal{W}_{2,\varepsilon}^2\left(\nu^{(v)}, \zeta_{\mathbf{b}}^{(v)}\right)$ denote the entropy-regularized Wasserstein transport discrepancy of view v. The empirical node measure, prototype-supported semantic measure, and resulting consensus marginal are defined as

$$\nu^{(v)} = \frac{1}{M}\sum_{i=1}^{M} \delta_{\mathbf{h}_i^{(v)}}, \qquad \zeta_{\mathbf{b}}^{(v)} = \sum_{k=1}^{K} b_k \delta_{\boldsymbol{\mu}_k^{(v)}}, \qquad \mathbf{b}^* = \arg\min_{\mathbf{b}\in\Delta^{K-1}} \sum_{v=1}^{V} \rho_v \mathcal{W}_\varepsilon^{(v)}(\mathbf{b}), \tag{9}$$

where $\delta_{\mathbf{x}}$ denotes the Dirac measure at $\mathbf{x}$, $\mathbf{b} \in \Delta^{K-1}$ denotes the shared semantic mass to optimize, and $\rho_v = 1/V$. Here, $\mathcal{W}_\varepsilon^{(v)}(\mathbf{b})$ summarizes the entropy-regularized transport between $\nu^{(v)}$ and $\zeta_{\mathbf{b}}^{(v)}$. The optimal marginal $\mathbf{b}^*$ provides semantic consensus while preserving view-specific prototype geometry. This consensus captures cross-view semantics without coordinate-level equivalence.

Entropy Regularization and Sinkhorn Optimization. To solve Eq. (9) efficiently, we introduce the node-level transport plan $\boldsymbol{\pi}_i^{(v)} \in \mathbb{R}_+^K$ and entropy regularizer $\Omega(\boldsymbol{\pi}_i^{(v)}) = -\langle \boldsymbol{\pi}_i^{(v)}, \log \boldsymbol{\pi}_i^{(v)} \rangle$. For fixed costs and $\varepsilon > 0$, the entropy-regularized transport problem is strictly convex, yielding

$$\begin{aligned} &\min_{\boldsymbol{\pi}_i^{(v)},\, \mathbf{b}\in\Delta^{K-1}} \quad \sum_{v=1}^{V} \rho_v \sum_{i=1}^{M} \left(\langle \boldsymbol{\pi}_i^{(v)}, \mathbf{c}_i^{(v)} \rangle - \varepsilon \Omega(\boldsymbol{\pi}_i^{(v)}) \right), \\ &\text{s.t.} \quad \mathbf{1}_K^\top \boldsymbol{\pi}_i^{(v)} = \frac{1}{M},\ \forall v \in [1,V],\ \forall i \in [1,M], \qquad \sum_{i=1}^{M} \boldsymbol{\pi}_i^{(v)} = \mathbf{b},\ \forall v \in [1,V]. \end{aligned} \tag{10}$$

where the first constraint fixes each node mass to $1/M$, while the second enforces the shared marginal $\mathbf{b}$.Let $\mathbf{C}^{(v)}$ collect $\mathbf{c}_i^{(v)}$ as rows and define $\mathbf{K}^{(v)} = \exp(-\mathbf{C}^{(v)}/\varepsilon)$. Furthermore, let $\odot$ and $\oslash$ denote element-wise multiplication and division, respectively. Using $\mathbf{u}^{(v)} \in \mathbb{R}^M$ and $\mathbf{w}^{(v)} \in \mathbb{R}^K$, Iterative Bregman Projections yield

$$\begin{aligned} \mathbf{u}^{(v)} \leftarrow M^{-1}\mathbf{1}_M \oslash (\mathbf{K}^{(v)}\mathbf{w}^{(v)}), \qquad \mathbf{b} \leftarrow \prod_{v=1}^{V} \left[\mathbf{w}^{(v)} \odot (\mathbf{K}^{(v)\top}\mathbf{u}^{(v)}) \right]^{V^{-1}}, \\ \mathbf{w}^{(v)} \leftarrow \mathbf{b} \oslash (\mathbf{K}^{(v)\top}\mathbf{u}^{(v)}). \end{aligned} \tag{11}$$

At convergence, $\mathbf{b}$ yields the optimal marginal $\mathbf{b}^*$, while $\boldsymbol{\Pi}^{(v)*} = \mathrm{diag}(\mathbf{u}^{(v)})\mathbf{K}^{(v)}\,\mathrm{diag}(\mathbf{w}^{(v)}) \in \mathbb{R}_+^{M\times K}$ is the optimal transport plan, whose i-th row $\boldsymbol{\pi}_i^{(v)*}$ represents the transport mass from node i to the K prototypes.

Unified Alignment Loss. The entropy term $\varepsilon\Omega(\boldsymbol{\pi}_i^{(v)})$ enables efficient Sinkhorn optimization. Once $\boldsymbol{\Pi}^{(v)*}$ is obtained, the network minimizes the geometric transport cost:

$$\mathcal{L}_{MOT} = \sum_{v=1}^{V}\sum_{i=1}^{M} \langle \boldsymbol{\pi}_i^{(v)*}, \mathbf{c}_i^{(v)} \rangle. \tag{12}$$

3.5 THE OVERALL LOSS FUNCTION

The proposed method is trained end-to-end by jointly optimizing four complementary objectives:

$$\mathcal{L}_{total} = \alpha\mathcal{L}_{Recon} + \beta\mathcal{L}_{MOT} + \lambda_1\mathcal{L}_{str} + \lambda_2\mathcal{L}_{topo}, \tag{13}$$

where $\mathcal{L}_{Recon}$ preserves the attributes of adversarially masked nodes, $\mathcal{L}_{MOT}$ aligns heterogeneous views through a shared Wasserstein semantic geometry, $\mathcal{L}_{str}$ promotes a cohesive modular dynamic graph, and $\mathcal{L}_{topo}$ enforces structural consistency between the purified graph and reconstructed features. The non-negative coefficients α, β, λ_1, and λ_2 balance the four objectives. During inference, the unified features from the fusion module are clustered by K-Means to obtain the final partitions.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We evaluate our method on four widely used multi-view remote sensing datasets: Trento, Salinas, XuZhou, and MDAS (Peng et al., 2025; Guan et al., 2024b; 2025a; 2026; Liu et al., 2026a). More details about datasets are provided in the Appendix A.1.

Baselines. We compare AM-GMAE against eleven newest MVC methods, including LMTC (Liu et al., 2025a), DFL-Net (Chen et al., 2025), MAGIC (Zhang et al., 2025a), hubREP (Xu et al., 2025), SAMVGC (Guan et al., 2025d), ESTMC (Ji & Feng, 2025), DMAC (Wang et al., 2025), CDD (Liu & Chang, 2025), SparseMVC (Liu et al., 2026b), ETAP (Zhang et al., 2026c) and CG-MAE (Guan et al., 2026). More details about baselines are provided in the the Appendix A.2.

Implementation Details. We used the PyTorch framework to implement AM-GMAE, running on a workstation equipped with a single 24GB NVIDIA GeForce RTX 3090 GPU and 64GB of memory. More details are provided in Appendix A.3 and A.4.

Table 1: Clustering performance of twelve multi-view clustering (MVC) methods on four datasets.

Method	Trento					XuZhou				
	ACC	Kappa	NMI	ARI	PUR	ACC	Kappa	NMI	ARI	PUR
DFL-NET	62.34±8.15	46.34±12.91	48.77±11.68	40.01±13.33	64.82±8.12	47.62±6.74	33.59±6.46	38.47±5.75	23.60±9.37	52.89±4.85
DMAC	79.55±0.85	72.53±0.66	71.01±5.47	66.29±3.22	82.39±0.87	70.25±2.20	63.40±2.44	62.73±2.94	63.87±1.06	77.12±1.47
hubREP	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
LMTC	43.95±0.02	32.39±0.02	39.26±0.02	29.44±0.02	68.84±0.01	56.77±0.01	50.07±0.01	62.19±0.01	46.82±0.01	78.88±0.01
MAGIC	58.67±0.10	48.69±0.04	51.79±0.01	45.16±0.00	75.71±0.01	70.49±1.27	64.37±1.31	64.83±0.32	62.18±0.75	79.78±0.26
SparseMVC	66.81±5.64	58.60±6.65	62.71±4.22	52.40±5.53	75.83±3.97	66.37±8.62	59.93±9.12	66.16±5.44	57.48±11.30	81.26±2.93
CDD	70.83±6.38	62.43±7.81	64.45±4.69	60.93±6.89	79.28±4.41	56.65±4.45	48.15±4.65	52.30±1.75	47.59±6.50	67.17±1.69
ETAP	58.22±1.25	39.86±3.08	40.68±4.88	36.03±5.56	61.78±2.27	34.31±5.97	15.56±5.34	15.71±4.11	8.21±4.56	44.95±1.79
ESTMC	66.10±0.97	58.01±1.47	53.09±0.86	55.85±2.48	79.18±3.59	40.15±2.37	32.42±1.95	36.98±2.19	24.24±0.13	62.30±0.78
SAMVGC	95.04±0.36	93.59±0.35	93.36±0.49	93.75±0.66	95.52±0.37	75.76±0.58	70.58±0.64	72.50±0.79	61.97±1.08	83.30±0.77
CG-MAE	95.59±0.35	94.13±0.47	91.33±0.40	95.64±0.24	95.72±0.24	83.42±1.57	79.28±1.95	74.26±0.71	73.73±3.77	83.91±1.18
Ours	**95.85±0.21**	**94.75±0.34**	**94.15±0.28**	**95.88±0.19**	**95.91±0.25**	**84.35±0.85**	**80.24±1.12**	**75.18±0.55**	**74.65±0.92**	**84.72±0.65**

Method	Salinas					MDAS				
	ACC	Kappa	NMI	ARI	PUR	ACC	Kappa	NMI	ARI	PUR
DFL-NET	40.00±3.96	31.83±5.03	46.46±4.95	20.65±3.78	41.62±4.33	39.16±4.45	18.59±3.10	16.47±3.26	10.10±5.50	32.14±2.34
DMAC	74.53±0.85	71.67±1.00	78.30±0.60	60.21±2.26	76.46±1.48	21.96±1.46	12.58±1.48	14.24±1.58	4.78±0.47	31.95±0.66
hubREP	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
LMTC	64.99±0.52	61.76±0.57	73.92±0.58	54.85±0.58	73.99±0.46	22.45±0.01	16.08±0.01	17.41±0.01	6.09±0.01	39.87±0.04
MAGIC	65.68±1.45	62.37±1.45	74.53±0.49	52.90±1.58	73.92±0.47	17.97±0.14	11.33±0.18	10.97±0.16	3.20±0.04	33.06±0.19
SparseMVC	73.12±2.54	70.33±2.54	82.62±2.35	62.12±4.72	79.85±3.86	26.80±1.83	20.28±2.65	24.32±2.66	8.30±1.14	38.71±2.49
CDD	60.07±2.78	55.18±3.08	70.88±3.41	54.11±5.30	62.01±2.44	30.79±3.29	18.58±2.61	20.11±2.55	10.66±2.68	36.42±1.75
ETAP	31.30±1.06	19.37±4.08	25.51±5.10	4.78±2.75	35.01±3.15	36.59±4.07	7.28±1.72	11.76±3.71	1.38±0.81	53.50±2.05
ESTMC	44.36±2.56	40.33±3.98	50.94±3.45	29.88±1.89	57.15±1.66	22.66±1.49	16.28±1.56	17.53±3.69	7.46±2.43	41.82±1.52
SAMVGC	67.72±0.30	64.30±0.29	69.45±0.38	54.81±0.80	69.78±0.52	47.59±0.70	28.85±0.64	35.09±0.87	35.59±0.57	46.30±0.55
CG-MAE	82.74±0.54	80.89±0.59	83.27±1.04	76.65±1.71	83.11±0.62	49.09±1.44	35.23±1.41	33.00±0.72	28.98±1.06	48.25±0.15
Ours	**85.72±1.31**	**84.13±1.45**	**85.84±0.65**	**81.68±1.03**	**87.10±0.79**	**51.01±0.65**	**36.54±0.81**	**36.22±0.54**	**37.15±0.47**	**61.47±0.33**

4.2 PERFORMANCE COMPARISON

Table 1 presents the quantitative clustering results of our proposed AM-GMAE and all baseline methods across the Trento, XuZhou, Salinas, and MDAS datasets. The proposed method consistently achieves the highest scores across all five evaluation metrics on every dataset. On the Trento dataset, our approach reaches an overall accuracy of 95.85% and a Kappa of 94.75%. Similarly, on the XuZhou dataset, our model achieves 84.35% accuracy and avoids the out-of-memory errors that plague baseline methods such as hubREP. We further validate our method's superiority on the challenging Salinas and MDAS datasets. For the Salinas dataset, our method delivers an accuracy of 85.72% and an NMI score of 85.84%, establishing a clear advantage over both reconstruction-based

and structure-guided baselines. On the MDAS dataset, AM-GMAE achieves 51.01% ACC, 36.54% Kappa, 36.22% NMI, 37.15% ARI, and 61.47% PUR. Notably, PUR improves by 7.97 percentage points over the best competing result, demonstrating the effectiveness of AM-GMAE in the heterogeneous multi-sensor setting. Overall, AM-GMAE outperformed the latest SOTA methods across all 20 combinations of metrics and datasets, thereby validating its effectiveness.

4.3 Qualitative Comparison

To intuitively evaluate the clustering performance, we visualize the clustering maps of different methods alongside the Ground Truth (GT) in Figure 2. As shown, baseline methods like DFL-NET and LMTC exhibit classification errors and spotty noise, particularly at boundaries and in areas with complex structures. In comparison, the clustering maps produced by AM-GMAE show more spatially coherent regions and closer visual correspondence with the ground-truth structures. These qualitative observations are consistent with the quantitative improvements reported in Table 1.

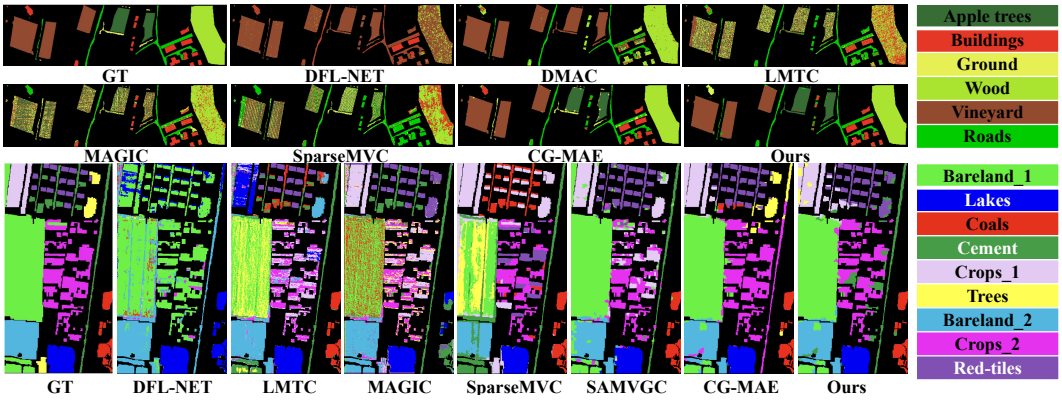


Figure 2: Clustering maps of different methods on the Trento and XuZhou datasets.

4.4 Ablation Study

To evaluate the contributions of our proposed modules, we conducted an ablation study across four datasets as detailed in Table 2. The baseline model relies on standard random masking, static graphs, and lacks cross-view alignment, which yields the lowest overall performance. Integrating the Clustering-Saliency Adversarial Mask brings substantial gains, such as increasing the accuracy on the Trento dataset from 84.12% to 89.45%. Building upon this foundation, the dynamic graph purification and topology optimization components refine the latent structure, pushing the Salinas dataset accuracy to 81.89%. Finally, the Multi-marginal Optimal Transport module aligns heterogeneous views through a shared Wasserstein semantic geometry. The complete framework integrates all components to achieve peak performance across all metrics, reaching 95.85% accuracy on Trento and 51.01% on MDAS, proving that the modules are effective and mutually reinforcing.

Table 2: Ablation study of different components on four remote sensing datasets.

Components				Trento				XuZhou				Salinas				MDAS			
CSAM	$\mathcal{L}_{str}$	$\mathcal{L}_{topo}$	MOT	ACC	NMI	ARI	PUR	ACC	NMI	ARI	PUR	ACC	NMI	ARI	PUR	ACC	NMI	ARI	PUR
				84.12	82.35	84.21	86.12	74.56	65.23	64.12	66.58	73.14	74.25	67.58	69.45	39.56	25.41	24.12	33.15
✓				89.45	87.12	89.05	90.25	78.12	69.45	68.25	71.45	78.45	79.12	71.56	73.56	43.12	29.56	29.45	37.45
✓	✓			92.15	90.34	92.45	93.15	80.56	71.89	70.45	74.25	80.12	80.45	73.12	75.89	45.45	31.12	32.56	40.12
✓	✓	✓		94.25	92.12	94.15	94.50	82.45	73.12	72.56	76.89	81.89	82.15	75.45	77.45	47.89	33.45	34.12	42.56
✓			✓	93.15	91.45	93.12	94.10	81.12	72.45	71.89	75.56	81.56	82.12	74.45	76.12	46.45	32.56	33.12	41.15
✓	✓	✓	✓	**95.85**	**94.15**	**95.88**	**95.91**	**84.35**	**75.18**	**74.65**	**84.72**	**85.72**	**85.84**	**81.68**	**87.10**	**51.01**	**36.22**	**37.15**	**61.47**

4.5 Further Analysis

Graph Topology Analysis To demonstrate the effectiveness of the DGPO module, we analyze the graph topology evolution. As shown in Figures 3(a) and (b), the initial graph on the Salinas dataset contains numerous noisy, inter-class connections, whereas the purified graph exhibits a distinct, cohesive block-diagonal structure, indicating the successful suppression of misleading edges. This

visual observation is quantitatively supported by Figure 3(c), where DGPO improves edge purity, node purity, homophily, and modularity across all four datasets, alongside a marked increase in the intra/inter-class weight ratio. Furthermore, the Kernel Density Estimation (KDE) of edge weights in Figure 3(d) reveals that DGPO effectively shifts the distribution of inter-class edge weights to lower values compared to the initial state, efficiently mitigating the impact of heterogeneous neighbors and providing a highly reliable manifold for subsequent cross-view alignment.

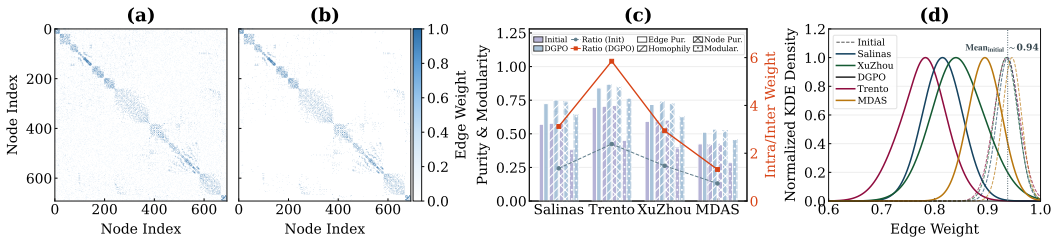


Figure 3: The graph topology analysis. (a) Initial Graph on the Salinas. (b) Purified Graph on the Salinas. (c) Topology Metrics Comparison across four datasets. (d) Inter-Class Weight Shift.

Cross-View Alignment Analysis To examine cross-view alignment stability, we conduct a controlled study on MDAS comparing MOT with three rigid alignment variants: Node MSE, InfoNCE, and cross-correlation. Here, Alignment Strength denotes the weighting coefficient of the cross-view alignment objective relative to the view-specific reconstruction and structural objectives, with larger values imposing stronger cross-view coupling. As shown in Figure 4(a), MOT maintains stable NMI across alignment strengths, whereas all rigid variants degrade under strong coupling. To assess negative transfer, we progressively corrupt one target view at a fixed alignment strength of 30 in Figure 4(b). MOT preserves positive ΔNMI, while rigid alignment becomes negative as corruption increases, indicating that rigid coupling propagates corrupted information across views. Figure 4(c) further shows that MOT causes only limited intra-view kNN geometry distortion under strong alignment, while rigid alignment induces substantially larger changes. These results demonstrate that MOT strengthens cross-view semantic interaction while preserving the geometric freedom of heterogeneous views, reducing negative transfer. Detailed experimental settings are provided in Appendix A.3, and additional experimental analyses in Appendix C.

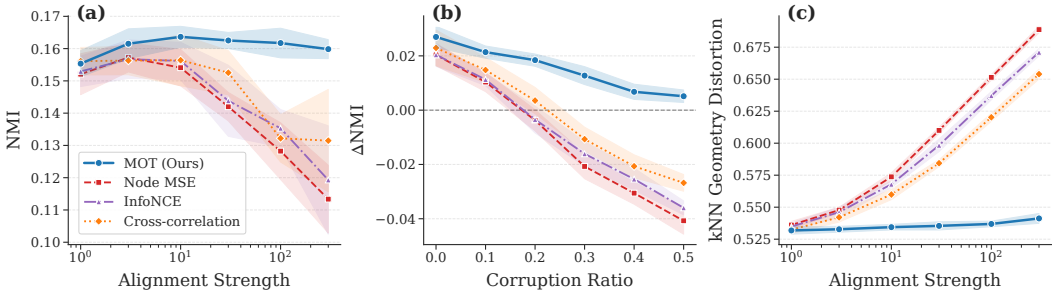


Figure 4: Cross-view alignment analysis on the MDAS dataset. (a) NMI under different alignment strengths. (b) Negative-transfer stress test under increasing corruption ratios. (c) Geometric distortion measured by changes in kNN neighborhoods under different alignment strengths.

5 Conclusion

In this paper, we propose AM-GMAE, an adaptive manifold graph masked autoencoder for multi-view remote sensing clustering. It addresses clustering-insensitive masking, unreliable graph connectivity, and heterogeneous semantic mismatch through CSAM, DGPO, and MOT, respectively. CSAM masks clustering-sensitive nodes, DGPO purifies graph topology, and MOT performs geometry-aware alignment through a shared Wasserstein semantic geometry without rigid pairwise correspondence. Experiments on four remote sensing benchmarks show consistent improvements over state-of-the-art methods across all metrics, while ablations confirm the complementary contributions of all components. Overall, AM-GMAE jointly models informative perturbations, reliable manifold structure, and heterogeneous cross-view geometry for robust unsupervised clustering.

REFERENCES

Yaoming Cai, Zijia Zhang, Pedram Ghamisi, Behnood Rasti, Xiaobo Liu, and Zhihua Cai. Transformer-based contrastive prototypical clustering for multimodal remote sensing data. *Information Sciences*, 649:119655, 2023.

Zhe Chen, Xiao-Jun Wu, Tianyang Xu, and Josef Kittler. Dfl-net: Disentangled feature learning network for multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 37 (8):4537–4547, 2025.

Yu Feng, Weixuan Liang, Xinhang Wan, Jiyuan Liu, Suyuan Liu, Qian Qu, Renxiang Guan, Huiying Xu, and Xinwang Liu. Incremental nyström-based multiple kernel clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 16613–16621, 2025.

Sichao Fu, Qinmu Peng, Yang He, Xiaorui Wang, Bin Zou, Duanquan Xu, Xiao-Yuan Jing, and Xinge You. Multilevel contrastive graph masked autoencoders for unsupervised graph-structure learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(2):3464–3478, 2024.

Renxiang Guan, Zihao Li, Wenxuan Tu, Jun Wang, Yue Liu, Xianju Li, Chang Tang, and Ruyi Feng. Contrastive multi-view subspace clustering of hyperspectral images based on graph convolutional networks. *IEEE Transactions on Geoscience and Remote Sensing*, 2024a.

Renxiang Guan, Wenxuan Tu, Zihao Li, Hao Yu, Dayu Hu, Yuzeng Chen, Chang Tang, Qiangqiang Yuan, and Xinwang Liu. Spatial-spectral graph contrastive clustering with hard sample mining for hyperspectral images. *IEEE Transactions on Geoscience and Remote Sensing*, 2024b.

Renxiang Guan, Junhong Li, Siwei Wang, Wenxuan Tu, Miaomiao Li, En Zhu, Xinwang Liu, and Ping Chen. Multi-view graph clustering with dual relation optimization for remote sensing data. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 7346–7355, 2025a.

Renxiang Guan, Tianrui Liu, Wenxuan Tu, Chang Tang, Wenhan Luo, and Xinwang Liu. Sampling enhanced contrastive multi-view remote sensing data clustering with long-short range information mining. *IEEE Transactions on Knowledge and Data Engineering*, 2025b.

Renxiang Guan, Wenxuan Tu, Dayu Hu, Weixuan Liang, Ke Liang, Yaowen Hu, Yue Liu, and Xinwang Liu. Prototype-driven multi-view attribute-missing graph clustering. *IEEE Transactions on Multimedia*, 2025c.

Renxiang Guan, Wenxuan Tu, Siwei Wang, Jiyuan Liu, Dayu Hu, Chang Tang, Yu Feng, Junhong Li, Baili Xiao, and Xinwang Liu. Structure-adaptive multi-view graph clustering for remote sensing data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 16933–16941, 2025d.

Renxiang Guan, Junhong Li, Siwei Wang, Tianrui Liu, Dayu Hu, Miaomiao Li, and Xinwang Liu. Graph masked autoencoder for multi-view remote sensing data clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21396–21404, 2026.

Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9052–9071, 2024.

Shubham Gupta, Shraban Kumar Chatterjee, and Suman Kundu. Llm-based contrastive self-supervised amr learning with masked graph autoencoders for fake news detection. *arXiv preprint arXiv:2508.18819*, 2025.

Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. Graphmae: Self-supervised masked graph autoencoders. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 594–604, 2022.

Zhenyu Hou, Yufei He, Yukuo Cen, Xiao Liu, Yuxiao Dong, Evgeny Kharlamov, and Jie Tang. Graphmae2: A decoding-enhanced masked self-supervised graph learner. In *Proceedings of the ACM web conference 2023*, pp. 737–746, 2023.

Jintian Ji and Songhe Feng. Anchors crash tensor: Efficient and scalable tensorial multi-view subspace clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2660–2675, 2025.

Wei Ju, Siyu Yi, Yifan Wang, Zhiping Xiao, Zhengyang Mao, Hourun Li, Yiyang Gu, Yifang Qin, Nan Yin, Senzhang Wang, et al. A survey of graph neural networks in real world: Imbalance, noise, privacy and ood challenges. *IEEE transactions on pattern analysis and machine intelligence*, 2025.

Haoyang Li, Xin Wang, Zeyang Zhang, Zongyuan Wu, Linxin Xiao, and Wenwu Zhu. Self-supervised masked graph autoencoder via structure-aware curriculum. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=6gX4rP6QJW.

Jintang Li, Ruofan Wu, Wangbin Sun, Liang Chen, Sheng Tian, Liang Zhu, Changhua Meng, Zibin Zheng, and Weiqiang Wang. What's behind the mask: Understanding masked graph modeling for graph autoencoders. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 1268–1279, 2023.

Bowen Liu, Xin Peng, Wenxuan Tu, Chengyao Wei, Xiangyan Tang, Jieren Cheng, and Miao Yu. Cross-view progressive feature filtering for multi-view graph clustering in remote sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 23649–23657, 2026a.

Chuang Liu, Yuyao Wang, Yibing Zhan, Xueqi Ma, Dapeng Tao, Jia Wu, and Wenbin Hu. Where to mask: Structure-guided masking for graph masked autoencoders. In Kate Larson (ed.), *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 2180–2188. International Joint Conferences on Artificial Intelligence Organization, 8 2024. doi: 10.24963/ijcai.2024/241. URL https://doi.org/10.24963/ijcai.2024/241. Main Track.

Jiyuan Liu, Xinwang Liu, Chuankun Li, Xinhang Wan, Hao Tan, Yi Zhang, Weixuan Liang, Qian Qu, Yu Feng, Renxiang Guan, et al. Large-scale multi-view tensor clustering with implicit linear kernels. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20727–20736, 2025a.

Ruimeng Liu, Xin Zou, Chang Tang, Xiao Zheng, Xingchen Hu, Kun Sun, and Xinwang Liu. Sparsemvc: probing cross-view sparsity variations for multi-view clustering. *Advances in Neural Information Processing Systems*, 38:154594–154622, 2026b.

Shujun Liu. Learning nonlinear flow matching for multimodal remote sensing image clustering. *IEEE Transactions on Geoscience and Remote Sensing*, 2026.

Shujun Liu and Ling Chang. Conditional dual diffusion for multimodal clustering of optical and sar images. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

Weiqi Liu, Yongshan Zhang, Xinxin Wang, and Lefei Zhang. Deep multi-level contrastive clustering for multi-modal remote sensing images. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 1239–1247, 2025b.

Valentino Maiorca, Luca Moschella, Antonio Norelli, Marco Fumero, Francesco Locatello, and Emanuele Rodolà. Latent space translation via semantic alignment. *Advances in Neural Information Processing Systems*, 36:55394–55414, 2023.

Abdelmalik Moujahid and Fadi Dornaika. Advanced unsupervised learning: a comprehensive overview of multi-view clustering techniques. *Artificial Intelligence Review*, 58(8):234, 2025.

Xin Peng, Bowen Liu, Renxiang Guan, and Wenxuan Tu. Multi-view graph clustering with dual structure awareness for remote sensing data. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2313–2322, 2025.

Weixuan Shen, Xiaobo Shen, and Shirui Pan. Adversarial contrastive graph masked autoencoder against graph structure and feature dual attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 12515–12523, 2025.

Jilin Shi and Peng Pu. Score-based progressive edge mask graph autoencoder. In *Fourth International Conference on Electronics Technology and Artificial Intelligence (ETAI 2025)*, volume 13692, pp. 143–150. SPIE, 2025.

Yucheng Shi, Yushun Dong, Qiaoyu Tan, Jundong Li, and Ninghao Liu. Gigamae: Generalizable graph masked autoencoder via collaborative latent space reconstruction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 2259–2269, 2023.

Qiaoyu Tan, Ninghao Liu, Xiao Huang, Rui Chen, Soo-Hyun Choi, and Xia Hu. Mgae: Masked autoencoders for self-supervised learning on graphs. *arXiv preprint arXiv:2201.02534*, 2022.

Qiaoyu Tan, Ninghao Liu, Xiao Huang, Soo-Hyun Choi, Li Li, Rui Chen, and Xia Hu. S2gae: Self-supervised graph autoencoders are generalizable learners with graph masking. In *Proceedings of the sixteenth ACM international conference on web search and data mining*, pp. 787–795, 2023.

Chuan Tang, Miaomiao Li, Jun Wang, Renxiang Guan, Siwei Wang, Chang Tang, En Zhu, and Xinwang Liu. Align then tensorize: Multi-level consistent anchor graph learning for scalable multi-view clustering. *IEEE Transactions on Image Processing*, 2026.

Yijun Tian, Kaiwen Dong, Chunhui Zhang, Chuxu Zhang, and Nitesh V Chawla. Heterogeneous graph masked autoencoders. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 9997–10005, 2023.

Yijun Tian, Chuxu Zhang, Ziyi Kou, Zheyuan Liu, Xiangliang Zhang, and Nitesh V Chawla. Ugmae: A unified framework for graph masked autoencoders. *arXiv preprint arXiv:2402.08023*, 2024.

Bocheng Wang, Chusheng Zeng, Mulin Chen, and Xuelong Li. Towards learnable anchor for deep multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21044–21052, 2025.

Liang Wang, Xiang Tao, Qiang Liu, and Shu Wu. Rethinking graph masked autoencoders through alignment and uniformity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15528–15536, 2024.

Danyang Wu, Penglei Wang, Jitao Lu, Zhanxuan Hu, Hongming Zhang, and Feiping Nie. Triangle topology enhancement for multi-view graph clustering. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

Zheming Xu, He Liu, Congyan Lang, Tao Wang, Yidong Li, and Michael C Kampffmeyer. A hubness perspective on representation learning for graph-based multi-view clustering. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15528–15537. IEEE, 2025.

Yuxuan Yao, Bo Peng, Tianyi Qin, Yanfeng Gu, Nam Ling, and Jianjun Lei. Hypergraph contrastive learning for large-scale hyperspectral image clustering. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

Jiahui Zhang, Wenjie Du, and Yang Wang. Dynamic and chemical constraints to enhance the molecular masked graph autoencoders. *Advances in Neural Information Processing Systems*, 38:174059–174085, 2026a.

Pei Zhang, Yuangang Pan, Siwei Wang, Shengju Yu, Huiying Xu, En Zhu, Xinwang Liu, and Ivor Tsang. Max-mahalanobis anchors guidance for multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 22488–22496, 2025a.

Qiqi Zhang, Chao Li, and Zhongying Zhao. Teacher-guided edge discriminator for personalized graph masked autoencoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 13269–13276, 2025b.

Yongshan Zhang, Xiaohuan Lin, Lefei Zhang, and Zhihua Cai. Orthogonal spatial-aware multi-view anchor graph clustering for incomplete remote sensing data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20532–20541, 2026b.

Yongshan Zhang, Kangyue Zheng, Shuaikang Yan, Xinxin Wang, and Zhihua Cai. Efficient tensorized multi-view anchor graph clustering with affinity propagation for remote sensing data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 28501–28509, 2026c.

Zijia Zhang, Yaoming Cai, Wenyin Gong, Xiaobo Liu, Cheng Zeng, and Gan Yu. Mmagl: Multi-objective multi-view attributed graph learning for joint clustering of hyperspectral and lidar data. *IEEE Transactions on Geoscience and Remote Sensing*, 2025c.

Ziwen Zhao, Yuhua Li, Yixiong Zou, Jiliang Tang, and Ruixuan Li. Masked graph autoencoder with non-discrete bandwidths. In *Proceedings of the ACM Web Conference 2024*, pp. 377–388, 2024.

Xizhen Zhu, Qiqi Zhang, Huanqing Cui, Chao Li, and Zhongying Zhao. M3gae: Motif-guided multi-view masked graph autoencoder for graph representation learning. *Pattern Recognition*, pp. 114495, 2026.

Luyao Zhuang, Shengyuan Chen, Yilin Xiao, Huachi Zhou, Yujing Zhang, Hao Chen, Qinggang Zhang, and Xiao Huang. Linearrag: Linear graph retrieval augmented generation on large-scale corpora. In *International Conference on Learning Representations*, volume 2026, pp. 147053–147075, 2026.

APPENDIX

This appendix provides supplementary implementation details, theoretical analyses, and additional experimental results for AM-GMAE. The contents are organized as follows:

- **Appendix A: Implementation Details.** We provide comprehensive implementation and experimental details to facilitate reproducibility. Specifically, we introduce the four multi-view remote sensing datasets and the compared baseline methods, describe the preprocessing procedure and graph construction strategy, and report the complete training configuration, hyperparameter settings, and evaluation metrics. We further provide detailed protocols for the graph topology and cross-view alignment analyses presented in the main paper, including the definitions of the corresponding diagnostic metrics. Finally, the complete optimization and inference pipeline of AM-GMAE is summarized to clarify how CSAM, DGPO, and MOT are jointly integrated during training.
- **Appendix B: Theoretical Analysis.** We provide theoretical justification for the three core components of AM-GMAE. For CSAM, we analyze the properties of the clustering-saliency score, its interpretation as a first-order measure of clustering-uncertainty sensitivity, the effect of cross-view saliency calibration, and the preference induced by Boltzmann masking. For DGPO, we analyze the properties of threshold-based graph purification, its relationship to semantic separation, and the structural properties induced by structural-entropy regularization and topology-consistent reconstruction. For MOT, we study the entropy-regularized multi-view transport formulation and the associated Wasserstein-barycentric alignment mechanism. These analyses characterize the behavior and theoretical properties of the individual components under their stated assumptions and provide further motivation for the design of AM-GMAE.
- **Appendix C: More Experiments.** We provide additional experimental evidence to complement the quantitative and qualitative results reported in the main paper. In particular, we present further qualitative clustering results to examine the spatial consistency of the learned partitions across different remote sensing scenes, together with hyperparameter sensitivity analyses that investigate the robustness of AM-GMAE under different model configurations. These supplementary experiments provide a more comprehensive assessment of the empirical behavior, stability, and robustness of the proposed framework.
- **Appendix D: Ethics Statement.** We discuss the ethical considerations associated with this work.
- **Appendix E: Reproducibility Statement.** We summarize the experimental and implementation information provided to facilitate reproducibility.
- **Appendix F: LLM Usage.** We disclose the use of large language models during the preparation of this manuscript.

A Implementation Details

A.1 Datasets

We evaluate AM-GMAE on four widely used multi-view remote sensing datasets, namely Trento, XuZhou, Salinas, and MDAS. These datasets cover heterogeneous sensing modalities and spectral-spatial feature representations, providing diverse scenarios for evaluating unsupervised multi-view clustering. The statistical characteristics of the datasets are summarized in Table 3, and visual examples of the corresponding views and ground-truth maps are presented in Figure 5. The labels are used only for evaluation, while no ground-truth information is involved during model training.

Table 3: Statistical details of the multi-view remote sensing datasets.

Datasets	Samples	Clusters	Size	Views	Features
Trento	30,214	6	600×166	HSI & LiDAR	63/2
XuZhou	68,877	9	500×260	HSI & EMP & 3D Gabor	436/45/90
Salinas	54,129	16	512×217	HSI & EMP & 3D Gabor	204/45/90
MDAS	88,026	14	300×360	HSI & MSI & SAR & DSM	242/12/2/1

Trento Dataset. The Trento dataset was acquired over a rural area in the southern part of Trento, Italy, and represents a typical heterogeneous sensor fusion scenario. It consists of a Hyperspectral image (HSI) and a LiDAR-derived Digital Surface Model (LiDAR), which provide complementary spectral and elevation information, respectively. The HS view contains 63 spectral bands, while the LiDAR view provides two-dimensional elevation-related features. The scene has a spatial size of 600×166 pixels and contains 30,214 labeled samples from 6 land-cover classes, including Apple Trees, Buildings, Ground, Woods, Vineyard, and Roads. The complementary spectral and elevation characteristics make Trento suitable for evaluating the capability of multi-view clustering methods to integrate heterogeneous physical information.

XuZhou Dataset. The XuZhou dataset was collected over a peri-urban area of XuZhou, China, and contains 68,877 labeled samples from 9 land-cover classes. The scene has a spatial size of 500×260 pixels. Following the multi-view formulation commonly adopted for this dataset, three views are constructed from complementary spectral and spatial representations. The first view is the original Hyperspectral image with 436 spectral bands. The second view is generated using Extended Morphological Profiles (EMP) with 45 features to characterize spatial and morphological structures, while the third view is constructed using 3D Gabor wavelets with 90 features to capture complementary textural information. The resulting multi-view representation provides a challenging testbed for jointly modeling high-dimensional spectral information and diverse spatial structures.

Salinas Dataset. The Salinas dataset was acquired by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) over the agricultural region of Salinas Valley, California. It contains a 512×217 pixel scene with 54,129 labeled samples distributed across 16 land-cover classes. After removing water absorption bands, the original hyperspectral data provide 204 spectral features. Similar to XuZhou, we construct three complementary views consisting of the 204-dimensional HS representation, a 45-dimensional EMP representation, and a 90-dimensional 3D Gabor representation. With its high spatial resolution and fine-grained agricultural categories, Salinas provides a challenging benchmark for evaluating whether multi-view clustering methods can effectively integrate spectral, morphological, and textural information.

MDAS Dataset. The MDAS dataset provides a more heterogeneous multi-sensor setting and contains 88,026 labeled samples from 14 land-cover classes. The scene has a spatial size of 300×360 pixels and consists of four complementary views, including Hyperspectral (HSI), Multispectral (MSI), Synthetic Aperture Radar (SAR), and Digital Surface Model (DSM) data. The corresponding views contain 242, 12, 2, and 1 features, respectively. Unlike datasets dominated by hyperspectral and derived spatial representations, MDAS combines optical, radar, and elevation information with substantially different feature characteristics. It therefore provides a challenging setting for evaluating the robustness of AM-GMAE to heterogeneous latent semantics, complementary sensor information, and potential cross-view distribution discrepancies.

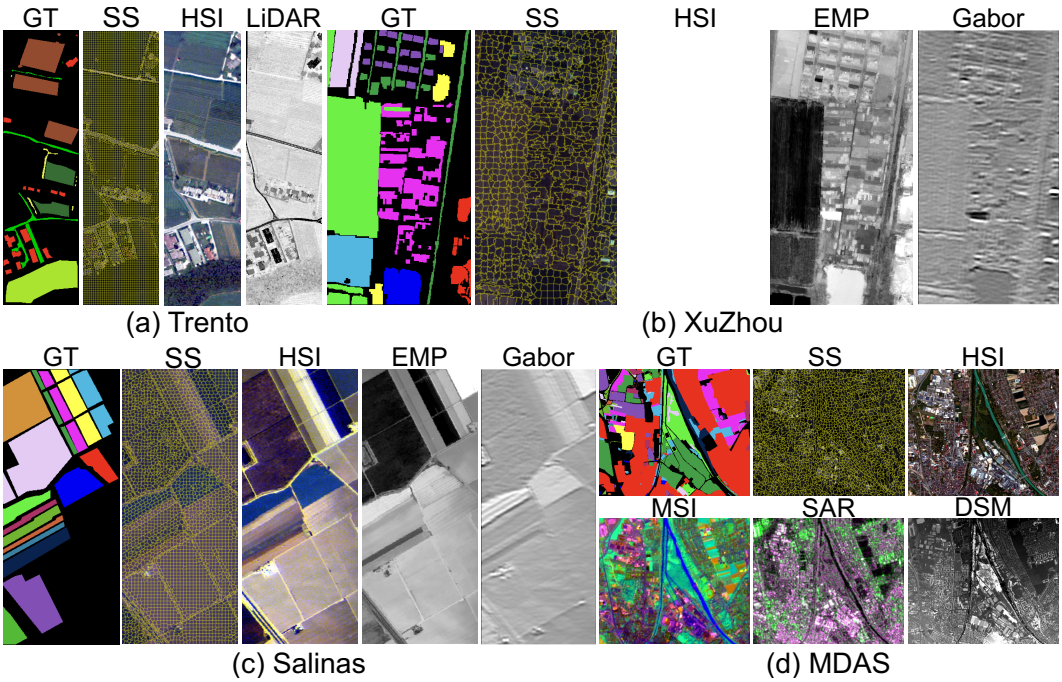


Figure 5: Visualizations of the four multi-view remote sensing datasets evaluated in our experiments: (a) Trento, (b) XuZhou, (c) Salinas, and (d) MDAS. Each panel displays the Ground Truth (GT) map, the Superpixel Segmentation (SS) map, and the corresponding multi-modal or multi-feature views.

To mitigate the prohibitive computational burden associated with pixel-wise graph construction on large-scale remote sensing imagery, and to establish a unified physical correspondence across heterogeneous views, we employ a superpixel-based preprocessing strategy. The superpixel segmentation results are shown in Figure 5. Let the multi-view dataset consist of V views, where each view $v \in \{1, \ldots, V\}$ possesses a distinct modality $\mathbf{X}^{(v)} \in \mathbb{R}^{W \times H \times D^{(v)}}$ capturing the same geographical scene. Direct pixel-to-pixel alignment is often infeasible due to varying sensor resolutions and noise distributions. To resolve this, as described in Section 3.1, we apply the Simple Linear Iterative Clustering (SLIC) algorithm on the channel-wise concatenated views to decompose the spatial domain into $M \ll N$ compact, perceptually meaningful regions (superpixels). The optimization objective of SLIC in our spectral-spatial context is to minimize the localized k-means clustering error defined by a distance measure D_s:

$$D_s(i,j) = \sqrt{\left(\frac{d_f(i,j)}{m}\right)^2 + \left(\frac{d_{xy}(i,j)}{S}\right)^2} \tag{14}$$

where d_f represents the Euclidean distance in the concatenated spectral feature space, d_{xy} is the spatial distance on the image grid, $S = \sqrt{(W \times H)/M}$ is the grid interval, and m controls the compactness of the superpixels. Upon convergence, SLIC yields a segmentation map $\mathcal{M} \in \{1, \ldots, M\}^{W \times H}$, which partitions the pixel set into M disjoint subsets $\mathcal{R} = \{\mathcal{R}_1, \ldots, \mathcal{R}_M\}$. Critically, to ensure strictly consistent semantics across different modalities, this map $\mathcal{M}$ is shared across all views $v \in \{1, \ldots, V\}$. Subsequently, a deterministic feature aggregation is performed within the boundaries defined by $\mathcal{M}$ for each view. The node feature $\mathbf{z}_i^{(v)} \in \mathbb{R}^{D^{(v)}}$ for the i-th superpixel is computed via average pooling:

$$\mathbf{z}_i^{(v)} = \frac{1}{|\mathcal{R}_i|} \sum_{\mathbf{x} \in \mathcal{R}_i} \mathbf{x}^{(v)} \tag{15}$$

This process transforms the raw imagery into a region-level feature matrix $\mathbf{Z}^{(v)} = [\mathbf{z}_1^{(v)}, \ldots, \mathbf{z}_M^{(v)}]^\top \in \mathbb{R}^{M \times D^{(v)}}$. This formulation guarantees that for any two views $v \neq u$, the

sample rows $\mathbf{z}_i^{(v)}$ and $\mathbf{z}_i^{(u)}$ correspond to the exact same physical region on the Earth's surface, providing a unified spatial support for the subsequent graph construction and cross-view alignment.

A.2 BASELINES

To evaluate the effectiveness of the proposed AM-GMAE, we compare it against eleven state-of-the-art multi-view clustering methods. These baselines encompass a wide range of advanced techniques, including graph masked autoencoders, tensor-based models, and contrastive learning approaches designed for multimodal remote sensing data.

- **CDD [TCSVT 2025]** (Liu & Chang, 2025): This conditional dual diffusion model leverages a decoupling autoencoder to effectively fuse the data manifolds of optical and SAR images into a unified latent space for multimodal clustering.
- **DFL-Net [TKDE 2025]** (Chen et al., 2025): This network utilizes a progressive learning framework and a comprehensive feature independence criterion to disentangle view-shared and view-exclusive features while adaptively filtering out clustering-irrelevant noise.
- **DMAC [AAAI 2025]** (Wang et al., 2025): This approach introduces a perturbation-driven anchor learning mechanism and an anchor graph convolution network to capture multi-view clustering consistency and optimize learnable anchors in linear time.
- **ESTMC [TPAMI 2025]** (Ji & Feng, 2025): This method integrates anchor representation learning and non-convex, function-based low-rank tensor learning to handle large-scale multi-view clustering efficiently while explicitly balancing consistency and complementarity across views.
- **hubREP [CVPR 2025]** (Xu et al., 2025): Addressing the hubness problem in high-dimensional spaces, this model utilizes a hub-aware encoder and cross-view pairing module to enhance representation embeddings and preserve local topology for graph-based multi-view clustering.
- **LMTC [CVPR 2025]** (Liu et al., 2025a): By removing the inadvertent label-leaking tensor rotation trick, this method formulates a linear-complexity algorithm that incorporates pairwise similarities with implicit linear kernels to effectively cluster large-scale multi-view tensor data.
- **MAGIC [AAAI 2025]** (Zhang et al., 2025a): This guidance method formalizes the desired properties of anchors into Diversity, Balance, and Compactness and explicitly constructs Max Mahalanobis Anchors to guide cross-view representations into a highly discriminative and compact structure.
- **SAMVGC [AAAI 2025]** (Guan et al., 2025d): This structure-adaptive approach refines graph structures and classifies samples into central and confusing categories to apply differential compactness and consistency learning, significantly improving remote sensing data clustering.
- **SparseMVC [NeurIPS 2025]** (Liu et al., 2026b): This framework addresses cross-view sparsity variations by employing adaptive sparse autoencoders and a correlation-informed sample reweighting module to dynamically balance view contributions and align feature distributions.
- **CG-MAE [AAAI 2026]** (Guan et al., 2026): This framework extends graph masked autoencoders to multi-view remote sensing clustering by selectively masking and reconstructing nodes and edges crucial for capturing key structural information.
- **ETAP [AAAI 2026]** (Zhang et al., 2026c): This method simultaneously learns view-specific and compressed anchor graphs with tensor Schatten p-norm regularization and reveals pixel clustering structures via affinity propagation for remote sensing data.

A.3 EXPERIMENT SETTINGS

All experiments are implemented using the Adam optimizer with a learning rate of 1×10^{-3} and trained for 100 epochs. Each experiment is independently repeated 10 times using random seeds $\{0, 1, \ldots, 9\}$, and the mean performance over the 10 runs is reported. The masking ratio r_m is

set to 0.5 for all datasets, and the reconstruction exponent γ is set to 1. For graph construction and superpixel preprocessing, we set the number of nearest neighbors k and the number of SLIC superpixels M to $(40, 7000)$, $(30, 3500)$, $(10, 1000)$, and $(10, 3500)$ for Trento, Salinas, XuZhou, and MDAS, respectively. For the overall objective in Eq. (13), the loss coefficients $(\alpha, \beta, \lambda_1, \lambda_2)$ are set to $(10, 100, 0.01, 0.01)$ for Trento, $(100, 1, 0.1, 0.1)$ for Salinas, $(1, 0.01, 1, 0.01)$ for XuZhou, and $(0.1, 100, 100, 1)$ for MDAS. The clustering temperature τ_c, Boltzmann masking temperature τ_m, cross-view saliency coefficient η, and Sinkhorn entropy regularization coefficient ε are set to $(1, 5, 0.5, 0.01)$, $(0.2, 5, 0.1, 0.05)$, $(0.05, 2, 0.5, 0.05)$, and $(0.05, 2, 0, 0.01)$ for the four datasets, respectively. The learnable topology threshold τ_{th} is initialized to 0.3 for all datasets, constrained to $[0, 1)$, and jointly optimized with the network during training. The isolated-node safeguard uses a small constant $\epsilon_g = 10^{-3}$ to retain the most similar non-self connection only when thresholding removes all incident edges of a node.

We evaluate clustering performance using five widely adopted metrics, including clustering accuracy (ACC), Cohen's Kappa coefficient (Kappa), normalized mutual information (NMI), adjusted Rand index (ARI), and purity (PUR). Let N_e denote the number of evaluated samples, and let $\mathbf{y} = \{y_i\}_{i=1}^{N_e}$ and $\hat{\mathbf{y}} = \{\hat{y}_i\}_{i=1}^{N_e}$ denote the ground-truth labels and predicted cluster assignments, respectively. ACC measures the proportion of correctly matched samples after optimal label permutation and is defined as $\mathrm{ACC} = \frac{1}{N_e} \max_{\sigma \in \mathfrak{S}_K} \sum_{i=1}^{N_e} \mathbb{I}(y_i = \sigma(\hat{y}_i))$, where $\mathfrak{S}_K$ denotes the set of permutations over K cluster labels and $\mathbb{I}(\cdot)$ is the indicator function. Given the optimally matched predicted labels, Kappa measures agreement after correcting for chance as $\mathrm{Kappa} = \frac{p_o - p_e}{1 - p_e}$, where $p_o = \mathrm{ACC}$ is the observed agreement and $p_e = \frac{1}{N_e^2} \sum_{k=1}^{K} a_k b_k$ is the expected agreement, with a_k and b_k denoting the numbers of samples belonging to ground-truth class k and matched predicted class k, respectively. NMI measures the mutual dependence between the two partitions and is computed as $\mathrm{NMI} = \frac{2I(\mathbf{y}; \hat{\mathbf{y}})}{\mathcal{H}(\mathbf{y}) + \mathcal{H}(\hat{\mathbf{y}})}$, where $I(\cdot;\cdot)$ and $\mathcal{H}(\cdot)$ denote mutual information and Shannon entropy, respectively. For ARI and PUR, let n_{kl} denote the number of samples shared by ground-truth class k and predicted cluster l, with $a_k = \sum_l n_{kl}$ and $b_l = \sum_k n_{kl}$. ARI evaluates pairwise partition agreement while correcting for chance and is defined as $\mathrm{ARI} = \dfrac{\sum_{k,l} \binom{n_{kl}}{2} - \frac{\left(\sum_k \binom{a_k}{2}\right)\left(\sum_l \binom{b_l}{2}\right)}{\binom{N_e}{2}}}{\frac{1}{2}\left[\sum_k \binom{a_k}{2} + \sum_l \binom{b_l}{2}\right] - \frac{\left(\sum_k \binom{a_k}{2}\right)\left(\sum_l \binom{b_l}{2}\right)}{\binom{N_e}{2}}}$, while PUR measures the dominant-class proportion within each predicted cluster as $\mathrm{PUR} = \frac{1}{N_e} \sum_{l=1}^{K} \max_k n_{kl}$. Higher values of all five metrics indicate better clustering performance.

To provide additional experimental details for the graph topology analysis presented in Section 4.5, we further specify the topology evaluation protocol and the quantitative criteria used to compare the initial graph with the topology purified by DGPO. Following the notation in Section 3.3, the initial and purified adjacency matrices of view v are denoted by $\mathbf{A}^{(v)} \in \mathbb{R}^{M \times M}$ and $\tilde{\mathcal{A}}^{(v)} \in \mathbb{R}^{M \times M}$, respectively, where M denotes the number of superpixel nodes used in the corresponding topology-analysis experiment. The purified topology is extracted at Epoch 100 after DGPO has progressively refined the edge weights and suppressed unreliable connections. The learnable topology threshold τ_{th} is initialized to 0.3 and jointly optimized during training rather than being treated as a fixed post-processing threshold. The following topology metrics are computed separately for the initial graph $\mathbf{A}^{(v)}$ and the purified graph $\tilde{\mathcal{A}}^{(v)}$. The ground-truth labels are introduced only for post-hoc structural diagnosis and are not involved in model optimization. Let $\mathbf{y} = \{y_i\}_{i=1}^{M}$ with $y_i \in \{1, \ldots, K\}$ denote the semantic labels of the graph nodes, and let $\mathbb{I}(y_i = y_j)$ denote the class-consistency indicator, which equals 1 if $y_i = y_j$ and 0 otherwise. For the initial graph, the nonzero edge set is defined as $\mathcal{E}(\mathbf{A}^{(v)}) = \{(i, j) \mid 1 \le i < j \le M,\ A_{ij}^{(v)} > 0\}$, and the same definition applies to the purified graph by replacing $\mathbf{A}^{(v)}$ with $\tilde{\mathcal{A}}^{(v)}$.

To characterize topology quality from complementary perspectives, we report four structural consistency measures together with an intra/inter weight ratio. For the initial graph, edge purity measures the discrete proportion of surviving edges connecting nodes from the same semantic class and is computed as $\mathcal{H}_{\mathrm{edge}}(\mathbf{A}^{(v)}) = \sum_{(i,j) \in \mathcal{E}(\mathbf{A}^{(v)})} \mathbb{I}(y_i = y_j) / |\mathcal{E}(\mathbf{A}^{(v)})|$. Since this metric ignores the magnitude of individual connections, we additionally calculate weighted homophily as $\mathcal{H}_{\mathrm{weight}}(\mathbf{A}^{(v)}) = \sum_{i<j} A_{ij}^{(v)} \mathbb{I}(y_i = y_j) / \sum_{i<j} A_{ij}^{(v)}$. To provide a node-centered assessment, we

define $\mathcal{N}_i = \{j \mid A_{ij}^{(v)} > 0\}$ and $\mathrm{pur}_i(\mathbf{A}^{(v)}) = \sum_{j\in\mathcal{N}_i} A_{ij}^{(v)}\mathbb{I}(y_i = y_j)/\sum_{j\in\mathcal{N}_i} A_{ij}^{(v)}$. Letting $\mathcal{V}_c = \{i \mid y_i = c\}$, the macro node purity is $\mathcal{P}_{\mathrm{macro}}(\mathbf{A}^{(v)}) = \frac{1}{K}\sum_{c=1}^{K}\frac{1}{|\mathcal{V}_c|}\sum_{i\in\mathcal{V}_c}\mathrm{pur}_i(\mathbf{A}^{(v)})$. We further evaluate the global community organization using weighted modularity. With weighted degree $d_i = \sum_{j=1}^{M} A_{ij}^{(v)}$ and total undirected edge weight $\omega = \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M} A_{ij}^{(v)}$, the modularity induced by the ground-truth partition is defined as $Q(\mathbf{A}^{(v)}) = \frac{1}{2\omega}\sum_{i=1}^{M}\sum_{j=1}^{M}\left(A_{ij}^{(v)} - \frac{d_i d_j}{2\omega}\right)\mathbb{I}(y_i = y_j)$. Finally, the relative dominance of semantically consistent connections over heterogeneous connections is quantified by $\mathcal{R}_{\mathrm{intra/inter}}(\mathbf{A}^{(v)}) = \frac{\sum_{i<j} A_{ij}^{(v)}\mathbb{I}(y_i=y_j)}{\sum_{i<j} A_{ij}^{(v)}\mathbb{I}(y_i\neq y_j)}$. The corresponding metrics for the purified graph are obtained by replacing $\mathbf{A}^{(v)}$ with $\tilde{A}^{(v)}$ in the above definitions. Figure 3(c) jointly reports these measurements to evaluate local consistency, global community organization, and intra-/inter-class separation from complementary perspectives.

To complement the aggregate statistics, Figure 3(d) analyzes how the continuous distribution of inter-class edge weights changes after DGPO purification. For the initial graph, the set of nonzero inter-class edge weights is defined as $\mathcal{S}_{\mathrm{inter}}(\mathbf{A}^{(v)}) = \{w_\ell = A_{ij}^{(v)} \mid i < j,\ A_{ij}^{(v)} > 0,\ y_i \neq y_j\}$ with $K_{\mathrm{inter}} = |\mathcal{S}_{\mathrm{inter}}(\mathbf{A}^{(v)})|$. Their empirical density is estimated using a one-dimensional Gaussian kernel density estimator, $\hat{f}_{\mathbf{A}^{(v)}}(x) = \frac{1}{K_{\mathrm{inter}}h}\sum_{\ell=1}^{K_{\mathrm{inter}}} K_G((x - w_\ell)/h)$, where $K_G(u) = \frac{1}{\sqrt{2\pi}}\exp(-u^2/2)$ and $h > 0$ is the bandwidth adaptively determined by the standard Scott/Silverman rule. The density is evaluated over the displayed edge-weight interval $x \in [0.60, 1.00]$ and peak-normalized as $f_{\mathrm{norm},\mathbf{A}^{(v)}}(x) = \hat{f}_{\mathbf{A}^{(v)}}(x)/\max_x \hat{f}_{\mathbf{A}^{(v)}}(x)$ so that the comparison emphasizes differences in distributional location and shape rather than differences in the absolute number of inter-class edges. The corresponding quantities for the purified graph are obtained by replacing $\mathbf{A}^{(v)}$ with $\tilde{\mathcal{A}}^{(v)}$. A leftward displacement of the purified distribution therefore indicates that heterogeneous edges are systematically assigned lower confidence, while a reduction of density in the high-weight region indicates that strongly misleading cross-class connections are being suppressed. This distributional view complements the scalar statistics in Figure 3(c) by showing how individual heterogeneous edge weights are redistributed during topology refinement. The analysis is conducted on Salinas, Trento, XuZhou, and MDAS, and the graph-construction setting is kept identical between the "Initial" and "DGPO" cases within each paired comparison, such that their difference reflects the effect of topology purification rather than a change in the evaluation configuration.

To provide additional experimental details for the cross-view alignment analysis presented in Section 4.5, we further clarify the definition of alignment strength and the common control protocol adopted by the three diagnostic experiments. Since β has already been assigned a different role in the original training objective, we introduce a separate scalar $\xi \geq 0$ exclusively for this analysis to denote the strength of the cross-view alignment constraint. Specifically, the controlled diagnostic objective is expressed as $\mathcal{L}_{\mathrm{diag}}^{\chi} = \mathcal{L}_{\mathrm{task}} + \xi\mathcal{L}_{\mathrm{align}}^{\chi}$, where $\mathcal{L}_{\mathrm{task}}$ denotes the view-specific task objective that is kept unchanged across all comparison groups and $\chi \in \{\mathrm{MOT}, \mathrm{MSE}, \mathrm{InfoNCE}, \mathrm{CC}\}$ denotes the alignment mechanism. The scalar ξ is therefore only an experimental control variable for adjusting the relative coupling strength and should not be identified with the coefficient β in the original model objective. When ξ is small, the view-specific objective dominates and each view retains relatively large freedom to preserve its own representation geometry; as ξ increases, cross-view consistency imposes progressively stronger coupling. For MOT, $\mathcal{L}_{\mathrm{align}}^{\mathrm{MOT}} = \mathcal{L}_{MOT}$ is the Wasserstein-barycenter transport objective introduced in Section 3.4. The node-wise MSE baseline uses $\mathcal{L}_{\mathrm{align}}^{\mathrm{MSE}} = \frac{1}{M}\sum_{i=1}^{M}\|\mathbf{h}_i^{(u)} - \mathbf{h}_i^{(v)}\|_2^2$, which directly penalizes the Euclidean discrepancy between paired representations. The InfoNCE baseline is defined as $\mathcal{L}_{\mathrm{align}}^{\mathrm{InfoNCE}} = -\frac{1}{M}\sum_{i=1}^{M}\log\frac{\exp(\mathrm{sim}(\mathbf{h}_i^{(u)},\mathbf{h}_i^{(v)})/\tau_{\mathrm{nce}})}{\sum_{j=1}^{M}\exp(\mathrm{sim}(\mathbf{h}_i^{(u)},\mathbf{h}_j^{(v)})/\tau_{\mathrm{nce}})}$, where $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity and $\tau_{\mathrm{nce}} > 0$ is the contrastive temperature. The cross-correlation baseline adopts $\mathcal{L}_{\mathrm{align}}^{\mathrm{CC}} = \sum_a((R_{\mathrm{cc}})_{aa} - 1)^2 + \lambda_{\mathrm{cc}}\sum_{a\neq b}(R_{\mathrm{cc}})_{ab}^2$, where $\mathbf{R}_{\mathrm{cc}}$ is the normalized cross-correlation matrix between latent dimensions and λ_{cc} controls the off-diagonal decorrelation term. The distinct symbols τ_{nce} and λ_{cc} are used to avoid conflicts with the temperature parameters and loss coefficients already defined elsewhere in the manuscript. All three experiments use the paired random-seed set $\mathcal{S} = \{0, 1, \ldots, 9\}$, and for a fixed seed s, all alignment mechanisms share the same pa-

rameter initialization and DataLoader shuffling order so that differences between methods are not confounded by stochastic initialization or batch sampling.

In the alignment-strength sensitivity experiment shown in Figure 4(a), all views are kept clean and the diagnostic alignment strength is varied over $\xi \in \{1, 3, 10, 30, 100, 300\}$. For each alignment mechanism χ and strength ξ, the reported clustering performance is the ten-seed mean $\mu_{\mathrm{NMI}}(\chi, \xi) = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \mathrm{NMI}(\chi, \xi, s)$, while the corresponding variability is measured by $\sigma_{\mathrm{NMI}}(\chi, \xi) = \sqrt{\frac{1}{|\mathcal{S}|-1} \sum_{s \in \mathcal{S}} [\mathrm{NMI}(\chi, \xi, s) - \mu_{\mathrm{NMI}}(\chi, \xi)]^2}$. This experiment jointly evaluates performance and stability as the coupling between views becomes stronger. In particular, the comparison is designed to examine whether rigid point-wise or dimension-wise alignment increasingly interferes with view-specific structures when the alignment constraint becomes dominant, and whether distribution-level transport can preserve greater geometric flexibility while still promoting semantic consistency. Figure 4(b) further evaluates robustness to asymmetric view degradation. The diagnostic alignment strength is fixed at $\xi = 30$, and one target view v_c is corrupted through feature dropout with $p \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5\}$. Given the superpixel-level feature matrix $\mathbf{Z}^{(v_c)} \in \mathbb{R}^{M \times D^{(v_c)}}$, a binary mask $\mathbf{B}^{(p)} \in \{0, 1\}^{M \times D^{(v_c)}}$ is sampled with $\Pr(B_{ij}^{(p)} = 0) = p$, and the degraded input is constructed as $\tilde{\mathbf{Z}}^{(v_c)}(p) = \mathbf{Z}^{(v_c)} \odot \mathbf{B}^{(p)}$. Negative transfer is quantified by the seed-wise paired difference $\Delta_{\mathrm{NT}}(\chi, p, s) = \mathrm{NMI}_{\text{all-views}}(\chi, p, s) - \mathrm{NMI}_{\text{clean-only}}(s)$, where $\mathrm{NMI}_{\text{all-views}}(\chi, p, s)$ is obtained when the corrupted view remains involved in joint alignment and $\mathrm{NMI}_{\text{clean-only}}(s)$ is obtained from a separately trained reference system using only the clean views indexed by $\{1, \ldots, V\} \setminus \{v_c\}$. Accordingly, $\Delta_{\mathrm{NT}} < 0$ means that retaining the degraded view results in lower clustering performance than removing it. Performing this subtraction within the same random seed reduces seed-specific performance offsets and isolates the effect introduced by the corrupted view, allowing the experiment to distinguish useful multi-view information transfer from harmful propagation of unreliable semantics.

Finally, Figure 4(c) provides a mechanism-level analysis by measuring how strongly cross-view alignment alters the original local geometry within each view. Let $\mathbf{h}_i^{(v),\mathrm{pre}}$ denote the representation of node i in view v obtained during the task-only stage before the diagnostic cross-view alignment constraint is applied, and let $\mathbf{h}_i^{(v),\mathrm{post}}(\xi)$ denote the corresponding representation after optimization with alignment strength ξ. Using cosine distance and a fixed neighborhood size $k_{\mathrm{geo}} = 15$, the geometry distortion is measured as $D_{\mathrm{geo}}^{(v)}(\xi) = 1 - \frac{1}{M} \sum_{i=1}^{M} \frac{|\mathcal{N}_{k_{\mathrm{geo}}}(\mathbf{h}_i^{(v),\mathrm{pre}}) \cap \mathcal{N}_{k_{\mathrm{geo}}}(\mathbf{h}_i^{(v),\mathrm{post}}(\xi))|}{k_{\mathrm{geo}}}$, where $\mathcal{N}_{k_{\mathrm{geo}}}(\mathbf{h})$ denotes the indices of the k_{geo} nearest neighbors of representation $\mathbf{h}$ in the corresponding latent space. This quantity measures the fraction of the original local neighborhood that is replaced after alignment: a value close to 0 indicates that most local neighbors are preserved, whereas a larger value indicates stronger deformation of the view-specific manifold. The pre- and post-alignment representations are extracted from the same training-run instances used in the alignment-strength sensitivity experiment, which establishes a seed-wise correspondence between clustering behavior and geometric deformation. Together, the three controlled experiments examine three related but distinct properties of cross-view alignment: sensitivity to increasingly strong coupling, susceptibility to negative transfer when one view becomes unreliable, and preservation of the intrinsic local geometry of individual views. This design makes the empirical trends reported in Section 4.5 interpretable not only in terms of final clustering performance, but also in terms of how different alignment mechanisms transfer information across views and reshape their underlying representation spaces.

A.4 Algorithm Pipeline

The complete optimization procedure of AM-GMAE is summarized in Algorithm 1. Given the multi-view remote sensing data, the model first constructs a shared superpixel partition and the corresponding view-specific graphs. CSAM computes clustering-saliency scores from the unmasked representations and calibrates the masking priority using information from correlated views. The masked graph inputs are then encoded to obtain the view-specific latent representations $\{\mathbf{H}^{(v)}\}_{v=1}^{V}$, while the decoder reconstructs the corresponding node attributes. Based on these latent representations, DGPO dynamically constructs purified graph structures and regularizes their structural and topological consistency. Finally, MOT aligns the heterogeneous view-specific representations

Algorithm 1 Optimization Pipeline of AM-GMAE

Input: Multi-view data $\mathcal{X} = \{\mathbf{X}^{(v)}\}_{v=1}^{V}$; number of clusters K; number of superpixels M; number of neighbors k; masking ratio r_m; training epochs T; Sinkhorn iterations R; hyperparameters $\tau_c, \tau_m, \eta, \varepsilon, \gamma, \alpha, \beta, \lambda_1, \lambda_2$.
Output: Pixel-level clustering partition $\hat{\mathbf{y}}$.

Apply SLIC to the concatenated multi-view features to obtain the shared superpixel partition $\mathcal{R} = \{\mathcal{R}_i\}_{i=1}^{M}$.
for $v = 1, \ldots, V$ **do**
  Obtain the superpixel-level feature matrix $\mathbf{Z}^{(v)}$ by region-wise average pooling.
  Construct the view-specific k-NN graph $\mathbf{A}^{(v)}$ and its normalized adjacency $\tilde{\mathbf{A}}^{(v)}$.
end for
Initialize all learnable parameters Θ, including the encoder–decoder parameters, view-specific prototypes, mask tokens, and the learnable topology threshold τ_{th}.
for $t = 1, \ldots, T$ **do**
  for $v = 1, \ldots, V$ **do**
    Obtain the pre-mask representation $\tilde{\mathbf{H}}^{(v)}$ from the unmasked graph input.
    Compute the soft cluster assignments according to Eq. (1).
    Compute the clustering-saliency scores according to Eq. (2).
  end for
  for $v = 1, \ldots, V$ **do**
    Calibrate the node saliency using cross-view information according to Eq. (3).
    Sample $\lfloor r_m M \rfloor$ nodes without replacement and construct the masked input according to Eq. (4).
    Encode the masked graph to obtain $\mathbf{H}^{(v)}$ and decode the node attributes to obtain $\hat{\mathbf{Z}}^{(v)}$.
  end for
  Compute the masked reconstruction objective according to Eq. (5).
  for $v = 1, \ldots, V$ **do**
    Construct the purified graph from $\mathbf{H}^{(v)}$ according to Eq. (6), including the isolated-node safeguard described in Section 3.3.
  end for
  Compute the structural regularization according to Eq. (7).
  Compute the topology-consistent reconstruction objective according to Eq. (8).
  Construct the view-specific transport costs and formulate the entropy-regularized barycentric transport problem according to Eq. (10).
  Perform R Sinkhorn iterations according to Eq. (11) and obtain the corresponding transport plans.
  Compute the cross-view transport objective according to Eq. (12).
  Compute the overall training objective according to Eq. (13).
  Update all learnable parameters Θ using Adam.
end for
Obtain the trained view-specific representations $\{\mathbf{H}^{(v)}\}_{v=1}^{V}$.
Construct the unified representation $\mathbf{H}$ according to Eq. (16).
Apply K-Means with K clusters to $\mathbf{H}$ to obtain the superpixel-level cluster assignments.
Propagate each superpixel assignment to all pixels belonging to the corresponding superpixel region.
return the pixel-level partition $\hat{\mathbf{y}}$.

through a shared Wasserstein semantic marginal. All learnable parameters are jointly optimized according to Eq. (13).

After optimization, the learned view-specific representations are integrated by a parameter-free feature fusion operation. Specifically, the representations from all views are concatenated along the feature dimension to construct the unified representation:

$$\mathbf{H} = \mathrm{Concat}\left(\mathbf{H}^{(1)}, \mathbf{H}^{(2)}, \ldots, \mathbf{H}^{(V)}\right) \in \mathbb{R}^{M \times Vd}. \tag{16}$$

Since the fusion operation does not impose coordinate-wise averaging across views, it retains the complementary information contained in the view-specific representations while preserving their individual latent geometry. The resulting unified representation $\mathbf{H}$ is used for the final K-Means clustering.

B THEORETICAL ANALYSIS

This section analyzes the theoretical properties of the three core components of AM-GMAE. For CSAM, we establish that the clustering-saliency score characterizes the local worst-case sensitivity of clustering uncertainty and that Boltzmann masking preferentially selects clustering-sensitive nodes. For DGPO, we characterize threshold-based purification, establish a soft semantic-cut interpretation of the structural objective, and analyze the denoising effect of topology-consistent reconstruction under a homophilous purified graph. For MOT, we prove existence and uniqueness of the entropy-regularized transport solution for fixed costs, derive the corresponding iterative Bregman projections, and characterize its invariance to independent view-wise Euclidean isometries. These results establish properties of the individual mechanisms under their stated assumptions; they do not imply global convergence of the complete non-convex end-to-end training problem.

B.1 PROPERTIES OF CLUSTERING-SALIENCY ADVERSARIAL MASK

CSAM measures the sensitivity of clustering uncertainty rather than clustering uncertainty itself. This distinction is important: a perfectly uniform assignment maximizes Shannon entropy, yet its entropy gradient with respect to the logits is zero. Hence a high-entropy node is not necessarily a locally sensitive node. The following analysis shows that the proposed saliency score directly quantifies the first-order variation of clustering uncertainty under latent perturbations.

For notational simplicity, node and view superscripts are omitted whenever no ambiguity arises. Throughout this subsection, $\tau_c > 0$, $\tilde{\mathbf{h}} \neq \mathbf{0}$, and $\boldsymbol{\mu}_k \neq \mathbf{0}$ for every k. When differentiating with respect to $\tilde{\mathbf{h}}$, the prototypes and temperature are regarded as fixed.

Lemma 1 (Entropy Gradient and Saliency Bound). *Let $\mathbf{Q} = (Q_1, \dots, Q_K)^\top \in \Delta^{K-1}$ be generated by*

$$Q_k = \frac{\exp(a_k)}{\sum_{j=1}^{K} \exp(a_j)}, \qquad a_k = \frac{1}{\tau_c} \cos(\tilde{\mathbf{h}}, \boldsymbol{\mu}_k), \tag{17}$$

and define

$$\mathcal{H}(\mathbf{Q}) = -\sum_{k=1}^{K} Q_k \log Q_k, \qquad S = \left\| \nabla_{\tilde{\mathbf{h}}} \mathcal{H} \right\|_2. \tag{18}$$

Then $0 \leq \mathcal{H}(\mathbf{Q}) \leq \log K$, and $\mathcal{H}$ is strictly concave on the relative interior of Δ^{K-1}. Moreover,

$$\frac{\partial \mathcal{H}}{\partial a_k} = -Q_k \left(\log Q_k + \mathcal{H} \right), \tag{19}$$

and therefore

$$\nabla_{\tilde{\mathbf{h}}} \mathcal{H} = -\frac{1}{\tau_c} \sum_{k=1}^{K} Q_k \left(\log Q_k + \mathcal{H} \right) \nabla_{\tilde{\mathbf{h}}} \cos(\tilde{\mathbf{h}}, \boldsymbol{\mu}_k). \tag{20}$$

If $\|\tilde{\mathbf{h}}\|_2 \geq r_h > 0$, then

$$S \leq \frac{2\mathcal{H}}{\tau_c r_h} \leq \frac{2 \log K}{\tau_c r_h}. \tag{21}$$

In particular, $S = 0$ when $Q_k = 1/K$ for all k.

Proof. Since $-x \log x \geq 0$ for $x \in [0, 1]$, with the convention $0 \log 0 = 0$, we have $\mathcal{H}(\mathbf{Q}) \geq 0$. Let $\mathbf{u} = K^{-1} \mathbf{1}_K$. The nonnegativity of the Kullback–Leibler divergence gives

$$D_{\mathrm{KL}}(\mathbf{Q} \| \mathbf{u}) = \sum_{k=1}^{K} Q_k \log \frac{Q_k}{1/K} = \log K - \mathcal{H}(\mathbf{Q}) \geq 0, \tag{22}$$

and hence $\mathcal{H}(\mathbf{Q}) \leq \log K$. For $\mathbf{Q}$ in the relative interior of the simplex,

$$\nabla^2_{\mathbf{Q}} \mathcal{H} = -\operatorname{diag}(Q_1^{-1}, \dots, Q_K^{-1}). \tag{23}$$

Thus, for every nonzero $\mathbf{z} \in \mathbb{R}^K$,

$$\mathbf{z}^\top \nabla^2_{\mathbf{Q}} \mathcal{H}\, \mathbf{z} = -\sum_{k=1}^{K} \frac{z_k^2}{Q_k} < 0, \tag{24}$$

so $\mathcal{H}$ is strictly concave on the relative interior of Δ^{K-1}. For the softmax mapping,

$$\frac{\partial Q_j}{\partial a_k} = Q_j\big(\mathbb{I}(j=k) - Q_k\big). \tag{25}$$

Therefore,

$$\begin{aligned}\frac{\partial \mathcal{H}}{\partial a_k} &= -\sum_{j=1}^{K} (\log Q_j + 1)\frac{\partial Q_j}{\partial a_k} \\ &= -Q_k(\log Q_k + 1) + Q_k \sum_{j=1}^{K} Q_j(\log Q_j + 1) \\ &= -Q_k(\log Q_k + \mathcal{H}),\end{aligned} \tag{26}$$

where we used $\sum_j Q_j = 1$ and $\sum_j Q_j \log Q_j = -\mathcal{H}$. Applying the chain rule to $a_k = \tau_c^{-1}\cos(\tilde{\mathbf{h}}, \boldsymbol{\mu}_k)$ yields Eq. (20). The Jacobian form used in Section 3.2 is equivalent because

$$\nabla_{\tilde{\mathbf{h}}} \mathcal{H} = -\mathcal{J}_{\tilde{\mathbf{h}}}(\mathbf{Q})^\top (\log \mathbf{Q} + \mathbf{1}_K) = -\mathcal{J}_{\tilde{\mathbf{h}}}(\mathbf{Q})^\top \log \mathbf{Q}, \tag{27}$$

where the second equality follows from

$$\mathcal{J}_{\tilde{\mathbf{h}}}(\mathbf{Q})^\top \mathbf{1}_K = \nabla_{\tilde{\mathbf{h}}} \sum_{k=1}^{K} Q_k = \mathbf{0}. \tag{28}$$

For nonzero $\tilde{\mathbf{h}}$ and $\boldsymbol{\mu}_k$,

$$\nabla_{\tilde{\mathbf{h}}} \cos(\tilde{\mathbf{h}}, \boldsymbol{\mu}_k) = \frac{\boldsymbol{\mu}_k}{\|\tilde{\mathbf{h}}\|_2 \|\boldsymbol{\mu}_k\|_2} - \frac{\cos(\tilde{\mathbf{h}}, \boldsymbol{\mu}_k)}{\|\tilde{\mathbf{h}}\|_2^2}\tilde{\mathbf{h}}, \tag{29}$$

and hence

$$\left\|\nabla_{\tilde{\mathbf{h}}} \cos(\tilde{\mathbf{h}}, \boldsymbol{\mu}_k)\right\|_2^2 = \frac{1 - \cos^2(\tilde{\mathbf{h}}, \boldsymbol{\mu}_k)}{\|\tilde{\mathbf{h}}\|_2^2} \le \frac{1}{r_h^2}. \tag{30}$$

Using Eq. (20) and the triangle inequality,

$$\begin{aligned} S &\le \frac{1}{\tau_c r_h} \sum_{k=1}^{K} Q_k \left|\log Q_k + \mathcal{H}\right| \\ &\le \frac{1}{\tau_c r_h}\left(-\sum_{k=1}^{K} Q_k \log Q_k + \mathcal{H}\sum_{k=1}^{K} Q_k\right) = \frac{2\mathcal{H}}{\tau_c r_h} \le \frac{2\log K}{\tau_c r_h}. \end{aligned} \tag{31}$$

If $Q_k = 1/K$ for every k, then $\log Q_k + \mathcal{H} = -\log K + \log K = 0$, and Eq. (20) gives $S = 0$. □

Lemma 1 shows why entropy and saliency should not be conflated. A uniform assignment is maximally uncertain but locally stationary in entropy, whereas a large S identifies a representation at which clustering uncertainty changes rapidly under perturbation. The next theorem makes this interpretation precise.

Theorem 1 (First-Order Clustering-Uncertainty Sensitivity)**.** *Assume that $\mathcal{H}(\tilde{\mathbf{h}})$ is twice continuously differentiable on*

$$\mathcal{B}_{\epsilon_0} = \left\{\tilde{\mathbf{h}} + \boldsymbol{\delta} : \|\boldsymbol{\delta}\|_2 \le \epsilon_0\right\}, \tag{32}$$

and that

$$\sup_{\mathbf{x} \in \mathcal{B}_{\epsilon_0}} \|\nabla^2 \mathcal{H}(\mathbf{x})\|_2 \le L_H \tag{33}$$

for some finite L_H. Then, for every $0 < \epsilon \le \epsilon_0$,

$$\left| \max_{\|\boldsymbol{\delta}\|_2 \le \epsilon} \left[\mathcal{H}(\tilde{\mathbf{h}} + \boldsymbol{\delta}) - \mathcal{H}(\tilde{\mathbf{h}}) \right] - \epsilon S \right| \le \frac{L_H}{2} \epsilon^2. \tag{34}$$

Consequently,

$$\max_{\|\boldsymbol{\delta}\|_2 \le \epsilon} \left[\mathcal{H}(\tilde{\mathbf{h}} + \boldsymbol{\delta}) - \mathcal{H}(\tilde{\mathbf{h}}) \right] = \epsilon S + \mathcal{O}(\epsilon^2). \tag{35}$$

Proof. Taylor's theorem gives, for every $\|\boldsymbol{\delta}\|_2 \le \epsilon$,

$$\mathcal{H}(\tilde{\mathbf{h}} + \boldsymbol{\delta}) - \mathcal{H}(\tilde{\mathbf{h}}) = \nabla \mathcal{H}(\tilde{\mathbf{h}})^\top \boldsymbol{\delta} + R(\boldsymbol{\delta}), \tag{36}$$

where

$$|R(\boldsymbol{\delta})| \le \frac{L_H}{2} \|\boldsymbol{\delta}\|_2^2 \le \frac{L_H}{2} \epsilon^2. \tag{37}$$

By Cauchy–Schwarz,

$$\nabla \mathcal{H}(\tilde{\mathbf{h}})^\top \boldsymbol{\delta} \le \epsilon \|\nabla \mathcal{H}(\tilde{\mathbf{h}})\|_2 = \epsilon S, \tag{38}$$

which gives

$$\max_{\|\boldsymbol{\delta}\|_2 \le \epsilon} \left[\mathcal{H}(\tilde{\mathbf{h}} + \boldsymbol{\delta}) - \mathcal{H}(\tilde{\mathbf{h}}) \right] \le \epsilon S + \frac{L_H}{2} \epsilon^2. \tag{39}$$

If $S > 0$, choose

$$\boldsymbol{\delta}^* = \epsilon \frac{\nabla \mathcal{H}(\tilde{\mathbf{h}})}{\|\nabla \mathcal{H}(\tilde{\mathbf{h}})\|_2}. \tag{40}$$

Then

$$\mathcal{H}(\tilde{\mathbf{h}} + \boldsymbol{\delta}^*) - \mathcal{H}(\tilde{\mathbf{h}}) \ge \epsilon S - \frac{L_H}{2} \epsilon^2, \tag{41}$$

which gives the matching lower bound. If $S = 0$, choosing $\boldsymbol{\delta} = \mathbf{0}$ gives a nonnegative maximum, while the preceding upper bound becomes $L_H \epsilon^2 / 2$. Therefore Eq. (34) holds in both cases. □

Theorem 1 provides the main justification for the CSAM score: S is exactly the first-order coefficient of the largest possible increase in clustering entropy inside a sufficiently small ℓ_2 perturbation ball. Thus masking a high-saliency node removes information whose local perturbation has a strong effect on clustering confidence. This is a local sensitivity result; it does not claim monotonic entropy growth under arbitrarily large perturbations.

The cross-view calibration preserves this interpretation while increasing the priority of nodes that are simultaneously sensitive in positively correlated views.

Proposition 1 (Cross-View Saliency Amplification)**.** *Assume $\eta \ge 0$. Then $\tilde{S}_i^{(v)} \ge S_i^{(v)}$. For fixed latent representations, $\tilde{S}_i^{(v)}$ is nondecreasing in every $S_i^{(u)}$, $u \ne v$, satisfying $\cos(\tilde{\mathbf{h}}_i^{(v)}, \tilde{\mathbf{h}}_i^{(u)}) > 0$. If $S_i^{(u)} \le S_{\max}$ for all views, then*

$$S_i^{(v)} \le \tilde{S}_i^{(v)} \le [1 + \eta(V - 1)] S_{\max}. \tag{42}$$

Proof. From the definition,

$$\tilde{S}_i^{(v)} = S_i^{(v)} + \eta \sum_{u \ne v} \left[\cos(\tilde{\mathbf{h}}_i^{(v)}, \tilde{\mathbf{h}}_i^{(u)}) \right]_+ S_i^{(u)}. \tag{43}$$

Every additional term is nonnegative, proving the lower bound and monotonicity. Since $[\cos(\cdot, \cdot)]_+ \le 1$, the upper bound follows immediately. □

This result does not assert that every positively correlated view is semantically useful. It states only that the calibration rule cannot suppress the original saliency and that, for fixed representations, positively correlated cross-view saliency increases masking priority in a controlled and bounded manner.

Theorem 2 (Conditional Saliency Preference of Boltzmann Masking). *Consider one draw of the without-replacement masking procedure, conditional on the current remaining node-index set $\mathcal{I}_{\mathrm{rem}}$. For fixed saliency scores, define*

$$p_i(\beta_m) = \frac{\exp(\beta_m \tilde{S}_i)}{\sum_{j \in \mathcal{I}_{\mathrm{rem}}} \exp(\beta_m \tilde{S}_j)}, \qquad \beta_m = \tau_m^{-1}. \tag{44}$$

and $\bar{S}(\beta_m) = \sum_{i \in \mathcal{I}_{\mathrm{rem}}} p_i(\beta_m)\tilde{S}_i$. Then

$$\frac{d\bar{S}}{d\beta_m} = \mathrm{Var}_{p(\beta_m)}(\tilde{S}) \geq 0. \tag{45}$$

Consequently, for every $\beta_m \geq 0$,

$$\bar{S}(\beta_m) \geq \frac{1}{|\mathcal{I}_{\mathrm{rem}}|} \sum_{i \in \mathcal{I}_{\mathrm{rem}}} \tilde{S}_i. \tag{46}$$

with strict inequality for $\beta_m > 0$ whenever the remaining saliency scores are not all identical.

Proof. Differentiating the softmax distribution gives $\frac{dp_i}{d\beta_m} = p_i(\tilde{S}_i - \bar{S})$. Therefore,

$$\frac{d\bar{S}}{d\beta_m} = \sum_{i \in \mathcal{I}_{\mathrm{rem}}} p_i \tilde{S}_i^2 - \left(\sum_{i \in \mathcal{I}_{\mathrm{rem}}} p_i \tilde{S}_i \right)^2 = \mathrm{Var}_{p(\beta_m)}(\tilde{S}) \geq 0. \tag{47}$$

At $\beta_m = 0$, $p_i = 1/|\mathcal{I}_{\mathrm{rem}}|$, so $\bar{S}(0)$ is the arithmetic mean of the remaining scores. Monotonicity in β_m proves Eq. (46). For finite β_m, all p_i are strictly positive, so the variance is positive whenever the scores are not all identical. □

As $\tau_m \to \infty$, the conditional distribution approaches uniform sampling over the remaining nodes. As $\tau_m \to 0^+$, its mass concentrates on the maximizers of $\tilde{S}_i$ in the current remaining set. Hence τ_m continuously controls the strength of preference toward high-saliency nodes. The result is deliberately conditional on each sequential draw: it establishes a saliency bias relative to uniform sampling over the same remaining set, but does not claim that the final without-replacement subset globally maximizes total saliency.

B.2 Properties of Dynamic Graph Purification

DGPO replaces the initial graph by a latent-similarity graph obtained through soft thresholding. We first analyze the thresholding operator itself and then connect it to semantic separation. The statements below concern the thresholded graph before the isolated-node safeguard unless explicitly stated otherwise.

Proposition 2 (Properties of the Thresholded Graph). *Assume $\mathbf{h}_i \neq \mathbf{0}$ for all i. For $i \neq j$,*

$$\tilde{\mathcal{A}}_{ij} > 0 \iff \cos(\mathbf{h}_i, \mathbf{h}_j) > \tau_{th}, \qquad 0 \leq \tilde{\mathcal{A}}_{ij} \leq 1 - \tau_{th}. \tag{48}$$

The thresholded adjacency is symmetric. Furthermore, if $\tau_1 < \tau_2$, then

$$\tilde{\mathcal{A}}_{ij}(\tau_2) \leq \tilde{\mathcal{A}}_{ij}(\tau_1), \qquad \operatorname{supp} \tilde{\mathcal{A}}(\tau_2) \subseteq \operatorname{supp} \tilde{\mathcal{A}}(\tau_1). \tag{49}$$

Proof. For $i \neq j$,

$$\tilde{\mathcal{A}}_{ij} = \left[\cos(\mathbf{h}_i, \mathbf{h}_j) - \tau_{th}\right]_+. \tag{50}$$

The positive-part operator is positive exactly when its argument is positive. Since cosine similarity is symmetric and bounded above by one, symmetry and Eq. (48) follow. Finally, for fixed x, $[x-\tau]_+$ is nonincreasing in τ, which proves both inequalities in Eq. (49). □

Thus increasing τ_{th} can only weaken or remove edges in the pre-safeguard graph. The isolated-node safeguard is excluded from this monotonicity statement because it deliberately reintroduces one small edge when thresholding removes all incident edges of a node. Its purpose is to make the subsequent degree normalization well defined rather than to strengthen the purification guarantee. The thresholding operation is piecewise differentiable in both the cosine score and τ_{th}; at the threshold boundary, the usual subgradient of the positive-part operator is sufficient for gradient-based optimization.

Theorem 3 (Sufficient Semantic-Margin Purification)**.** *Suppose there exist constants* s_{inter} *and* s_{intra} *such that*

$$s_{\text{inter}} < \tau_{th} < s_{\text{intra}}, \tag{51}$$

and

$$y_i = y_j \Longrightarrow \cos(\mathbf{h}_i, \mathbf{h}_j) \geq s_{\text{intra}}, \qquad y_i \neq y_j \Longrightarrow \cos(\mathbf{h}_i, \mathbf{h}_j) \leq s_{\text{inter}}. \tag{52}$$

Then every inter-class edge is removed by thresholding, while every intra-class pair satisfies

$$\tilde{\mathcal{A}}_{ij} \geq s_{\text{intra}} - \tau_{th} > 0. \tag{53}$$

If every node has at least one same-class neighbor, the isolated-node safeguard is never activated.

Proof. If $y_i \neq y_j$, then

$$\cos(\mathbf{h}_i, \mathbf{h}_j) - \tau_{th} \leq s_{\text{inter}} - \tau_{th} < 0, \tag{54}$$

so $\tilde{\mathcal{A}}_{ij} = 0$. If $y_i = y_j$, then

$$\cos(\mathbf{h}_i, \mathbf{h}_j) - \tau_{th} \geq s_{\text{intra}} - \tau_{th} > 0, \tag{55}$$

which gives Eq. (53). Under the final assumption, every node already has positive degree after thresholding, so the safeguard is not invoked. □

The labels in Theorem 3 are used only to state a sufficient separation condition and never enter training. The condition is intentionally strong: real latent spaces need not exhibit a global pairwise semantic margin. The theorem therefore does not claim that DGPO always reconstructs the ground-truth graph; instead, it establishes that the proposed thresholding rule is exactly consistent with semantic separation whenever such a latent margin exists.

B.3 Structural Entropy and Topology-Consistent Reconstruction

The two structural terms play different roles. $\mathcal{L}_{1str}$ depends only on the normalized node degrees and regularizes the organization of graph volume; by itself it contains no cluster-semantic information. In contrast, $\mathcal{L}_{2str}$ depends jointly on graph connectivity and the soft assignments $\mathbf{Q}$ and admits a direct semantic-cut interpretation. This distinction is important because minimizing $\mathcal{L}_{1str}$ alone cannot imply semantic community recovery.

Under the isolated-node safeguard, $d_i > 0$ for every node. Since softmax assignments satisfy $Q_{ik} > 0$ and $\sum_k Q_{ik} = 1$, we also have $\Gamma_G > 0$, $\Gamma_k > 0$, and

$$\sum_{k=1}^{K} \Gamma_k = \sum_{k=1}^{K} \sum_{i=1}^{M} Q_{ik} d_i = \sum_{i=1}^{M} d_i = \Gamma_G. \tag{56}$$

Consequently, $r_k = \Gamma_k / \Gamma_G$ is a strictly positive probability distribution over the K semantic groups.

Theorem 4 (Soft Semantic-Cut Identity and Structural-Entropy Bounds)**.** *For every semantic group* k*,*

$$\Gamma_k - m_k = \sum_{i=1}^{M} \sum_{j=1}^{M} Q_{ik}(1 - Q_{jk}) \tilde{\mathcal{A}}_{ij} \geq 0. \tag{57}$$

If, conditionally on the soft assignments, auxiliary labels are independently sampled as $Y_i \sim \text{Categorical}(\mathbf{Q}_i)$*, then* $\Gamma_k - m_k$ *equals the expected directed edge weight leaving group* k*. Moreover,*

$$0 \leq \mathcal{L}_{1str}^{(v)} \leq \log_2 M, \qquad 0 \leq \mathcal{L}_{2str}^{(v)} \leq \log_2 K. \tag{58}$$

Proof. Using $d_i = \sum_j \tilde{\mathcal{A}}_{ij}$,

$$\begin{aligned} \Gamma_k - m_k &= \sum_i Q_{ik} d_i - \sum_{i,j} Q_{ik} Q_{jk} \tilde{\mathcal{A}}_{ij} \\ &= \sum_{i,j} Q_{ik} \tilde{\mathcal{A}}_{ij} - \sum_{i,j} Q_{ik} Q_{jk} \tilde{\mathcal{A}}_{ij} \\ &= \sum_{i,j} Q_{ik}(1 - Q_{jk}) \tilde{\mathcal{A}}_{ij}, \end{aligned} \tag{59}$$

which proves Eq. (57). Every term is nonnegative.

Under the auxiliary categorical model,

$$\Pr(Y_i = k, Y_j \neq k) = Q_{ik}(1 - Q_{jk}), \tag{60}$$

so the right-hand side of Eq. (57) is

$$\mathbb{E}\left[\sum_{i,j} \tilde{\mathcal{A}}_{ij} \mathbb{I}(Y_i = k, Y_j \neq k)\right], \tag{61}$$

which is the expected directed outgoing edge weight of semantic group k.

Because $0 \leq 1 - Q_{jk} \leq 1$,

$$0 \leq \Gamma_k - m_k \leq \Gamma_k. \tag{62}$$

Using $r_k = \Gamma_k / \Gamma_G$,

$$\begin{aligned} 0 \leq \mathcal{L}_{2str}^{(v)} &= -\sum_{k=1}^{K} \frac{\Gamma_k - m_k}{\Gamma_G} \log_2 r_k \\ &\leq -\sum_{k=1}^{K} r_k \log_2 r_k \leq \log_2 K. \end{aligned} \tag{63}$$

Similarly, $\varrho_i = d_i / \Gamma_G$ is a probability distribution over M nodes, so

$$\mathcal{L}_{1str}^{(v)} = -\sum_{i=1}^{M} \varrho_i \log_2 \varrho_i \tag{64}$$

satisfies $0 \leq \mathcal{L}_{1str}^{(v)} \leq \log_2 M$. □

The auxiliary random labels above are introduced solely to interpret the deterministic algebraic quantity $\Gamma_k - m_k$ and are not sampled or used during optimization. The theorem shows that $\mathcal{L}_{2str}$ penalizes a soft semantic boundary mass, weighted by the information content $-\log_2 r_k$ associated with each semantic volume. In contrast, $\mathcal{L}_{1str}$ should be interpreted only as a complementary degree-organization regularizer. Therefore the semantic-cut interpretation belongs specifically to $\mathcal{L}_{2str}$ rather than to $\mathcal{L}_{1str}$ in isolation.

Corollary 5 (Hard-Assignment Reduction)**.** *Suppose $Q_{ik} \in \{0, 1\}$ and every group $C_k = \{i : Q_{ik} = 1\}$ has positive volume. Define*

$$\mathrm{cut}(C_k, \overline{C}_k) = \sum_{i \in C_k} \sum_{j \notin C_k} \tilde{\mathcal{A}}_{ij}. \tag{65}$$

Then

$$\Gamma_k = \mathrm{vol}(C_k), \qquad \Gamma_k - m_k = \mathrm{cut}(C_k, \overline{C}_k), \tag{66}$$

and consequently

$$\mathcal{L}_{2str}^{(v)} = -\sum_{k=1}^{K} \frac{\mathrm{cut}(C_k, \overline{C}_k)}{\Gamma_G} \log_2 \frac{\mathrm{vol}(C_k)}{\Gamma_G}. \tag{67}$$

Proof. For hard assignments, $Q_{ik} = \mathbb{I}(i \in C_k)$, and therefore

$$\Gamma_k = \sum_{i \in C_k} d_i = \mathrm{vol}(C_k). \tag{68}$$

Applying Eq. (57),

$$\Gamma_k - m_k = \sum_{i \in C_k} \sum_{j \notin C_k} \tilde{\mathcal{A}}_{ij} = \mathrm{cut}(C_k, \overline{C}_k). \tag{69}$$

Substitution into Eq. (7) proves Eq. (67). □

Corollary 5 shows that, when the assignments become discrete, the soft quantity $\Gamma_k - m_k$ reduces exactly to the usual outgoing graph cut. This establishes the semantic meaning of the two-dimensional structural term without claiming that its minimization alone recovers the ground-truth partition, because the assignments $\mathbf{Q}$ are themselves learned and may be imperfect.

We next analyze the topology-consistent reconstruction term. Define

$$\mathbf{P} = \tilde{\mathbf{D}}^{-1}\tilde{\mathcal{A}}. \tag{70}$$

The isolated-node safeguard guarantees $d_i > 0$, so $P_{ij} = \tilde{\mathcal{A}}_{ij}/d_i \geq 0$ and $\sum_j P_{ij} = 1$. Hence

$$(\mathbf{PZ})_i = \sum_{j=1}^{M} P_{ij}\mathbf{z}_j \tag{71}$$

is a convex combination of pristine features connected to node i in the purified graph. Thus the target of $\mathcal{L}_{topo}$ is a one-step random-walk average rather than an arbitrary transformed feature.

Theorem 6 (Conditional Noise Reduction under Homophilous Purification)**.** *Fix a row-stochastic transition matrix $\mathbf{P}$ and regard it as deterministic with respect to the feature-noise variables considered below. Suppose node i and every node j satisfying $P_{ij} > 0$ belong to the same semantic class c, whose class-conditional feature mean is denoted by $\boldsymbol{\mu}_c$, and*

$$\mathbf{z}_j = \boldsymbol{\mu}_c + \boldsymbol{\epsilon}_j, \tag{72}$$

where the noise vectors are mutually independent and satisfy

$$\mathbb{E}[\boldsymbol{\epsilon}_j] = \mathbf{0}, \qquad \mathrm{Cov}(\boldsymbol{\epsilon}_j) = \boldsymbol{\Sigma} \succeq \mathbf{0}. \tag{73}$$

Assume also that $\mathbf{z}_i = \boldsymbol{\mu}_c + \boldsymbol{\epsilon}_i$ follows the same covariance model. Then

$$\mathbb{E}[(\mathbf{PZ})_i] = \boldsymbol{\mu}_c, \tag{74}$$

and

$$\mathrm{Cov}[(\mathbf{PZ})_i] = \left(\sum_j P_{ij}^2\right)\boldsymbol{\Sigma} \preceq \boldsymbol{\Sigma} = \mathrm{Cov}(\mathbf{z}_i). \tag{75}$$

If at least two coefficients P_{ij} are strictly positive, then $\sum_j P_{ij}^2 < 1$; consequently, the covariance is strictly reduced whenever $\boldsymbol{\Sigma} \succ \mathbf{0}$.

Proof. Since $\sum_j P_{ij} = 1$,

$$(\mathbf{PZ})_i = \sum_j P_{ij}(\boldsymbol{\mu}_c + \boldsymbol{\epsilon}_j) = \boldsymbol{\mu}_c + \sum_j P_{ij}\boldsymbol{\epsilon}_j. \tag{76}$$

Taking expectation proves Eq. (74). By mutual independence,

$$\mathrm{Cov}[(\mathbf{PZ})_i] = \sum_j P_{ij}^2\,\mathrm{Cov}(\boldsymbol{\epsilon}_j) = \left(\sum_j P_{ij}^2\right)\boldsymbol{\Sigma}. \tag{77}$$

Because $P_{ij} \geq 0$ and $\sum_j P_{ij} = 1$,

$$\sum_j P_{ij}^2 \leq \left(\sum_j P_{ij}\right)^2 = 1. \tag{78}$$

This proves Eq. (75). If at least two coefficients are positive, then the cross terms in $(\sum_j P_{ij})^2$ are strictly positive, so $\sum_j P_{ij}^2 < 1$. For $\boldsymbol{\Sigma} \succ \mathbf{0}$, multiplication by a scalar strictly smaller than one yields a strict covariance reduction. □

The fixed-$\mathbf{P}$ qualification in Theorem 6 is essential. In AM-GMAE, the purified graph is learned from representations that ultimately depend on the observed data, so $\mathbf{P}$ need not be statistically independent of the feature noise. The theorem is therefore a conditional smoothing result rather than an unconditional guarantee for the full graph-learning process. Together with Theorem 3, it yields the intended theoretical chain: if the latent geometry becomes sufficiently semantically separated, thresholding can remove heterogeneous edges; once the resulting neighborhood is homophilous, topology averaging preserves the semantic mean while reducing independent feature noise.

B.4 Properties of Multi-marginal Optimal Transport

The exact multi-marginal formulation introduced in Section 3.4 requires an exponentially large joint coupling tensor. AM-GMAE instead optimizes the barycentric relaxation consisting of one node-to-prototype transport plan per view and a shared prototype-mass marginal. The following analysis concerns this implemented entropy-regularized barycentric problem.

Throughout this subsection, the cost matrices $\{\mathbf{C}^{(v)}\}_{v=1}^{V}$ are assumed fixed and finite, $\varepsilon > 0$, and the view weights satisfy $\rho_v > 0$ and $\sum_v \rho_v = 1$. AM-GMAE uses $\rho_v = 1/V$. These assumptions concern the inner transport problem and do not imply convexity of the full neural-network training objective.

Theorem 7 (Existence and Uniqueness of the Entropic Transport Solution)**.** *For fixed cost matrices, the entropy-regularized barycentric transport problem in Section 3.4 admits a unique optimal collection of transport plans $\{\mathbf{\Pi}^{(v)*}\}_{v=1}^{V}$ and a unique common marginal $\mathbf{b}^*$.*

Proof. The feasible set is nonempty because

$$\Pi_{ik}^{(v)} = \frac{1}{MK}, \qquad b_k = \frac{1}{K} \tag{79}$$

satisfies all constraints. The feasibility conditions consist of linear equalities and nonnegativity constraints, so the feasible set is closed and convex. The row constraints imply

$$0 \le \Pi_{ik}^{(v)} \le \frac{1}{M}, \tag{80}$$

hence the feasible set is bounded and therefore compact.

With $0 \log 0 = 0$, the objective can be written as

$$\sum_{v=1}^{V} \rho_v \left[\left\langle \mathbf{\Pi}^{(v)}, \mathbf{C}^{(v)} \right\rangle + \varepsilon \sum_{i,k} \Pi_{ik}^{(v)} \log \Pi_{ik}^{(v)} \right]. \tag{81}$$

It is continuous on the compact feasible set, so a minimizer exists.

The scalar function $x \mapsto x \log x$ is strictly convex on $[0, \infty)$. Because $\varepsilon > 0$ and every $\rho_v > 0$, the entropy term in Eq. (81) is strictly convex in the complete collection $\{\mathbf{\Pi}^{(v)}\}_{v=1}^{V}$, while the transport-cost term is linear. Hence two distinct feasible collections of transport plans cannot both be optimal, proving uniqueness of $\{\mathbf{\Pi}^{(v)*}\}_{v=1}^{V}$.

Finally, every feasible plan satisfies

$$\mathbf{b} = \mathbf{\Pi}^{(v)\top} \mathbf{1}_M. \tag{82}$$

Therefore the unique optimal transport plans determine a unique common marginal $\mathbf{b}^*$. □

Theorem 7 establishes that the inner entropy-regularized transport subproblem is well posed. It does not imply uniqueness of the encoder, decoder, or prototype parameters because those parameters determine the cost matrices nonlinearly and the complete AM-GMAE objective remains non-convex.

To analyze the solver, define

$$\begin{aligned} \mathcal{C}_1 &= \left\{ \{\mathbf{\Pi}^{(v)}\}_{v=1}^{V} : \mathbf{\Pi}^{(v)} \mathbf{1}_K = M^{-1} \mathbf{1}_M,\ \forall v \right\}, \\ \mathcal{C}_2 &= \left\{ \{\mathbf{\Pi}^{(v)}\}_{v=1}^{V} : \mathbf{\Pi}^{(1)\top} \mathbf{1}_M = \cdots = \mathbf{\Pi}^{(V)\top} \mathbf{1}_M \right\}. \end{aligned} \tag{83}$$

The set $\mathcal{C}_2$ enforces equality of the column marginals but does not by itself constrain their total mass. However, every element of $\mathcal{C}_1 \cap \mathcal{C}_2$ has unit total mass because each matrix in $\mathcal{C}_1$ has M rows of mass $1/M$. Hence the common marginal belongs to Δ^{K-1} at feasibility.

Define the generalized KL divergence

$$D_{\mathrm{KL}}(\mathbf{P} \| \mathbf{G}) = \sum_{i,k} \left[P_{ik} \log \frac{P_{ik}}{G_{ik}} - P_{ik} + G_{ik} \right], \tag{84}$$

and its weighted product-space extension

$$\mathcal{D}_\rho\left(\{\mathbf{P}^{(v)}\}\middle\|\{\mathbf{G}^{(v)}\}\right)=\sum_{v=1}^{V}\rho_v D_{\mathrm{KL}}(\mathbf{P}^{(v)}\|\mathbf{G}^{(v)}). \tag{85}$$

Theorem 8 (Bregman Projections and Convergence of the MOT Solver)**.** *Let*

$$\mathbf{K}^{(v)}=\exp\left(-\mathbf{C}^{(v)}/\varepsilon\right) \tag{86}$$

element-wise. For positive scaling vectors, the weighted KL projection onto $\mathcal{C}_1$ is

$$\mathbf{u}^{(v)}\leftarrow M^{-1}\mathbf{1}_M\oslash(\mathbf{K}^{(v)}\mathbf{w}^{(v)}). \tag{87}$$

After this projection, define the current column marginal

$$\mathbf{c}^{(v)}=\mathbf{w}^{(v)}\odot(\mathbf{K}^{(v)\top}\mathbf{u}^{(v)}). \tag{88}$$

The weighted KL projection onto $\mathcal{C}_2$ is obtained by

$$\mathbf{b}\leftarrow\prod_{v=1}^{V}\left(\mathbf{c}^{(v)}\right)^{\rho_v},\qquad \mathbf{w}^{(v)}\leftarrow\mathbf{b}\oslash(\mathbf{K}^{(v)\top}\mathbf{u}^{(v)}), \tag{89}$$

where products, powers, and divisions are element-wise. For $\rho_v=1/V$, these updates reduce to Eq. (11). Moreover, for fixed finite costs, $\varepsilon>0$, and positive initialization, cyclic application of these projections converges to the unique solution of Theorem 7.

Proof. Since

$$K_{ik}^{(v)}=\exp(-C_{ik}^{(v)}/\varepsilon), \tag{90}$$

we have

$$\begin{aligned}\varepsilon D_{\mathrm{KL}}(\mathbf{\Pi}^{(v)}\|\mathbf{K}^{(v)})=&\left\langle\mathbf{\Pi}^{(v)},\mathbf{C}^{(v)}\right\rangle+\varepsilon\sum_{i,k}\Pi_{ik}^{(v)}\log\Pi_{ik}^{(v)}\\&-\varepsilon\sum_{i,k}\Pi_{ik}^{(v)}+\varepsilon\sum_{i,k}K_{ik}^{(v)}.\end{aligned} \tag{91}$$

Every feasible transport plan has total mass one, so the third term in Eq. (91) is constant on $\mathcal{C}_1\cap\mathcal{C}_2$, while the final term is independent of $\mathbf{\Pi}^{(v)}$. Thus the entropy-regularized transport problem is equivalent to the weighted KL projection of the Gibbs kernels onto $\mathcal{C}_1\cap\mathcal{C}_2$.

Consider first a positive current matrix

$$\mathbf{\Gamma}^{(v)}=\mathrm{diag}(\mathbf{u}_{\mathrm{old}}^{(v)})\mathbf{K}^{(v)}\,\mathrm{diag}(\mathbf{w}^{(v)}). \tag{92}$$

Projection onto the row constraint is separable over rows. The Lagrangian stationarity condition has the form

$$\Pi_{ik}^{(v)}=\lambda_i^{(v)}\Gamma_{ik}^{(v)}. \tag{93}$$

Enforcing $\sum_k\Pi_{ik}^{(v)}=1/M$ gives

$$\lambda_i^{(v)}=\frac{M^{-1}}{\sum_k\Gamma_{ik}^{(v)}}. \tag{94}$$

Because

$$\sum_k\Gamma_{ik}^{(v)}=u_{\mathrm{old},i}^{(v)}(\mathbf{K}^{(v)}\mathbf{w}^{(v)})_i, \tag{95}$$

the previous row scaling cancels, yielding Eq. (87). After this projection,

$$\mathbf{\Gamma}^{(v)}=\mathrm{diag}(\mathbf{u}^{(v)})\mathbf{K}^{(v)}\,\mathrm{diag}(\mathbf{w}^{(v)}), \tag{96}$$

whose column marginal is

$$\mathbf{\Gamma}^{(v)\top}\mathbf{1}_M=\mathbf{w}^{(v)}\odot(\mathbf{K}^{(v)\top}\mathbf{u}^{(v)})=\mathbf{c}^{(v)}. \tag{97}$$

We now project the collection $\{\mathbf{\Gamma}^{(v)}\}$ onto $\mathcal{C}_2$. Let $\mathbf{b} \geq \mathbf{0}$ denote the common column marginal. For fixed $\mathbf{b}$, KL minimization is separable over views and columns, and its stationarity condition gives

$$\Pi_{ik}^{(v)} = \Gamma_{ik}^{(v)} \frac{b_k}{c_k^{(v)}}. \tag{98}$$

Substitution into Eq. (85) leaves, up to terms independent of $\mathbf{b}$,

$$\sum_{k=1}^{K} \sum_{v=1}^{V} \rho_v \left[b_k \log \frac{b_k}{c_k^{(v)}} - b_k \right]. \tag{99}$$

For positive $\mathbf{c}^{(v)}$, this expression is strictly convex in each $b_k > 0$. Its derivative vanishes when

$$\sum_{v=1}^{V} \rho_v \log \frac{b_k}{c_k^{(v)}} = 0. \tag{100}$$

Because $\sum_v \rho_v = 1$,

$$b_k = \prod_{v=1}^{V} (c_k^{(v)})^{\rho_v}, \tag{101}$$

which proves the first update in Eq. (89). The required column-rescaling factor is $b_k / c_k^{(v)}$, so

$$\begin{aligned} \mathbf{w}_{\text{new}}^{(v)} &= \mathbf{w}^{(v)} \odot \left[\mathbf{b} \oslash \mathbf{c}^{(v)} \right] \\ &= \mathbf{b} \oslash (\mathbf{K}^{(v)\top} \mathbf{u}^{(v)}), \end{aligned} \tag{102}$$

which proves the second update. Finally, finite costs and $\varepsilon > 0$ imply $K_{ik}^{(v)} > 0$ for all i, k, v, and positive initialization keeps all iterates in the positive domain. The sets $\mathcal{C}_1$ and $\mathcal{C}_2$ are affine equality-constraint sets with a nonempty intersection. Therefore the standard convergence result for cyclic iterative Bregman projections onto affine marginal constraints applies (**?**), and the iterates converge to the weighted KL projection onto $\mathcal{C}_1 \cap \mathcal{C}_2$. By Theorem 7, this projection is unique and coincides with the unique entropy-regularized transport solution. □

The intermediate $\mathbf{b}$ in Eq. (89) is not required to have unit mass because a projection onto $\mathcal{C}_2$ may temporarily violate the row constraints. At convergence, both constraint sets hold simultaneously, and therefore

$$\mathbf{1}_K^\top \mathbf{b}^* = \mathbf{1}_M^\top \mathbf{\Pi}^{(v)*} \mathbf{1}_K = \mathbf{1}_M^\top (M^{-1} \mathbf{1}_M) = 1. \tag{103}$$

Hence $\mathbf{b}^* \in \Delta^{K-1}$. The shared-marginal constraint has a specific interpretation:

$$\mathbf{\Pi}^{(v)*\top} \mathbf{1}_M = \mathbf{b}^*, \qquad v = 1, \ldots, V. \tag{104}$$

Thus all views assign the same total transport mass to each common prototype index. However, the formulation imposes neither $\boldsymbol{\pi}_i^{(u)*} = \boldsymbol{\pi}_i^{(v)*}$ nor $\mathbf{h}_i^{(u)} = \mathbf{h}_i^{(v)}$. Cross-view coupling therefore acts at the shared marginal level rather than by forcing node-wise coordinate equality. This interpretation is relative to the prototype indexing learned by the model; as in clustering generally, the transport constraints alone do not identify ground-truth class labels or eliminate label-switching symmetries.

Theorem 9 (View-Wise Euclidean-Isometry Invariance)**.** *For each view v, let*

$$T_v(\mathbf{x}) = \mathbf{R}_v \mathbf{x} + \mathbf{t}_v, \qquad \mathbf{R}_v^\top \mathbf{R}_v = \mathbf{I}, \tag{105}$$

and apply the same view-specific transformation to latent nodes and their corresponding prototypes:

$$\mathbf{h}_i^{(v)\prime} = \mathbf{R}_v \mathbf{h}_i^{(v)} + \mathbf{t}_v, \qquad \boldsymbol{\mu}_k^{(v)\prime} = \mathbf{R}_v \boldsymbol{\mu}_k^{(v)} + \mathbf{t}_v. \tag{106}$$

Then $\mathbf{C}^{(v)\prime} = \mathbf{C}^{(v)}$ for every view. Consequently, the unique entropy-regularized transport plans, the common marginal $\mathbf{b}^*$, and $\mathcal{L}_{MOT}$ are invariant under independent view-wise Euclidean isometries.

In contrast, the normalized node-wise MSE alignment

$$\mathcal{L}_{\text{MSE}} = \frac{1}{M} \sum_{i=1}^{M} \|\mathbf{h}_i^{(u)} - \mathbf{h}_i^{(v)}\|_2^2 \tag{107}$$

is not invariant to independent view-wise translations.

Proof. For every i, k, v,

$$\begin{aligned} C_{ik}^{(v)\prime} &= \|\mathbf{h}_i^{(v)\prime} - \boldsymbol{\mu}_k^{(v)\prime}\|_2^2 \\ &= \|\mathbf{R}_v(\mathbf{h}_i^{(v)} - \boldsymbol{\mu}_k^{(v)})\|_2^2 \\ &= (\mathbf{h}_i^{(v)} - \boldsymbol{\mu}_k^{(v)})^\top \mathbf{R}_v^\top \mathbf{R}_v(\mathbf{h}_i^{(v)} - \boldsymbol{\mu}_k^{(v)}) \\ &= \|\mathbf{h}_i^{(v)} - \boldsymbol{\mu}_k^{(v)}\|_2^2 = C_{ik}^{(v)}. \end{aligned} \tag{108}$$

Hence the transformed transport problem has exactly the same objective and feasible set as the original problem. By Theorem 7, its optimal plans and common marginal are unchanged, and therefore $\mathcal{L}_{MOT}$ is unchanged. For the MSE counterexample, suppose $\mathbf{h}_i^{(u)} = \mathbf{h}_i^{(v)}$ for every i, so $\mathcal{L}_{\mathrm{MSE}} = 0$. Leave view u unchanged and translate view v by a nonzero vector $\mathbf{t}$, i.e.,

$$\mathbf{h}_i^{(v)\prime} = \mathbf{h}_i^{(v)} + \mathbf{t}. \tag{109}$$

Then

$$\mathcal{L}'_{\mathrm{MSE}} = \frac{1}{M}\sum_{i=1}^{M} \|\mathbf{t}\|_2^2 = \|\mathbf{t}\|_2^2 > 0. \tag{110}$$

Thus node-wise MSE changes under an independent view translation, whereas the MOT objective remains invariant when each view and its prototypes undergo the same Euclidean isometry. □

Theorem 9 formalizes one sense in which MOT preserves heterogeneous geometric freedom: different views are not required to share the same latent coordinate frame. This property supports the motivation for reducing rigid cross-view coupling, but it should not be interpreted as a universal guarantee against negative transfer. Negative transfer in the complete model can still depend on representation quality, prototype estimation, graph construction, and non-convex optimization. The theorem therefore explains a geometric mechanism that can mitigate one source of harmful alignment pressure rather than proving that harmful transfer is impossible.

B.5 Computational Complexity

We finally analyze the additional complexity of the proposed components, excluding the encoder–decoder computations shared with the underlying GMAE.

For CSAM, evaluating node–prototype similarities and the corresponding clustering-saliency terms requires $\mathcal{O}(VMKd)$ arithmetic under direct evaluation. Cross-view saliency calibration compares corresponding node representations across view pairs and requires at most $\mathcal{O}(V^2Md)$ additional time. Thus, for fixed V, K, and d, the principal CSAM computations scale linearly with the number of graph nodes M. For DGPO, dense pairwise latent similarities require $\mathcal{O}(VM^2d)$ time and $\mathcal{O}(VM^2)$ adjacency memory. The soft intra-cluster quantity can be written as $m_k^{(v)} = \mathbf{q}_k^{(v)\top}\tilde{\mathcal{A}}^{(v)}\mathbf{q}_k^{(v)}$, so direct dense evaluation over all K groups requires $\mathcal{O}(VKM^2)$ time. Dense topology smoothing requires $\mathcal{O}(\sum_{v=1}^{V} M^2D^{(v)})$ time. If $\tilde{\mathcal{A}}^{(v)}$ is stored sparsely with E_v nonzero directed entries, these two costs reduce respectively to $\mathcal{O}(K\sum_v E_v)$ and $\mathcal{O}(\sum_v E_vD^{(v)})$. Therefore the dense DGPO implementation is the principal quadratic component with respect to M, which also motivates the superpixel formulation $M \ll N$.

For MOT, constructing the node-to-prototype cost matrices requires $\mathcal{O}(VMKd)$ time and $\mathcal{O}(VMK)$ storage. Each Bregman iteration requires matrix–vector products with $\mathbf{K}^{(v)}$ and $\mathbf{K}^{(v)\top}$ and therefore costs $\mathcal{O}(VMK)$. With R iterations, the transport solver requires $\mathcal{O}(RVMK)$ time. By comparison, an explicit V-way prototype coupling contains $\mathcal{O}(K^V)$ entries. Hence the barycentric formulation avoids the exponentially sized exact multi-marginal coupling tensor while retaining linear dependence on M and V for fixed K and R.

B.6 Summary of Theoretical Properties

The above results establish complementary properties for the three modules. For CSAM, Lemma 1 and Theorem 1 show that the saliency score is bounded and has a precise local worst-case perturbation interpretation, while Theorem 2 establishes that the sampling mechanism preferentially selects

high-saliency nodes. Thus the masking strategy is supported by a perturbation-sensitivity argument rather than by uncertainty magnitude alone.

For DGPO, Proposition 2 and Theorem 3 characterize the edge-removal behavior of the thresholding operator. Theorem 4 further shows that $\mathcal{L}_{2str}$ penalizes a soft semantic boundary quantity and reduces to a conventional graph cut under hard assignments. Finally, Theorem 6 establishes that, conditional on a fixed homophilous purified topology, random-walk averaging preserves the class mean while reducing independent feature-noise covariance. These results explain the complementary roles of purification, structural regularization, and topology-consistent reconstruction without assuming access to ground-truth labels during training.

For MOT, Theorem 7 establishes well-posedness of the fixed-cost entropy-regularized transport subproblem, while Theorem 8 derives the implemented Bregman updates and their fixed-cost convergence. Theorem 9 further shows that semantic coupling through the shared transport marginal does not require heterogeneous views to share an identical Euclidean coordinate frame. Overall, the theoretical analysis supports the three design principles of AM-GMAE: *perturbation-sensitive masking*, *semantically interpretable topology refinement*, and *geometry-flexible cross-view alignment*.

C More Experiments

C.1 Additional Qualitative Results

To further evaluate the spatial quality of the obtained clustering results, we provide additional clustering-map comparisons on the Salinas and MDAS datasets in Figures 6 and 7, respectively. These two datasets provide complementary evaluation scenarios. Salinas contains fine-grained agricultural parcels with elongated and spatially adjacent categories, while MDAS contains heterogeneous observations acquired from multiple sensing modalities. Therefore, the two datasets allow us to further examine whether AM-GMAE can preserve spatial structures under both fine-grained land-cover distributions and heterogeneous multi-sensor observations.

As shown in Figure 6, several competing methods exhibit noticeable scattered errors inside otherwise homogeneous agricultural regions. In particular, DFL-NET, LMTC, and MAGIC introduce substantial pixel-level fragmentation, while SparseMVC and CG-MAE recover clearer semantic structures but still produce visible local inconsistencies. In comparison, AM-GMAE generates more spatially compact clustering regions and better preserves the elongated shapes and boundaries of agricultural parcels. The clustering map is also visually closer to the Ground Truth (GT), which is consistent with the quantitative improvements reported in Table 1.

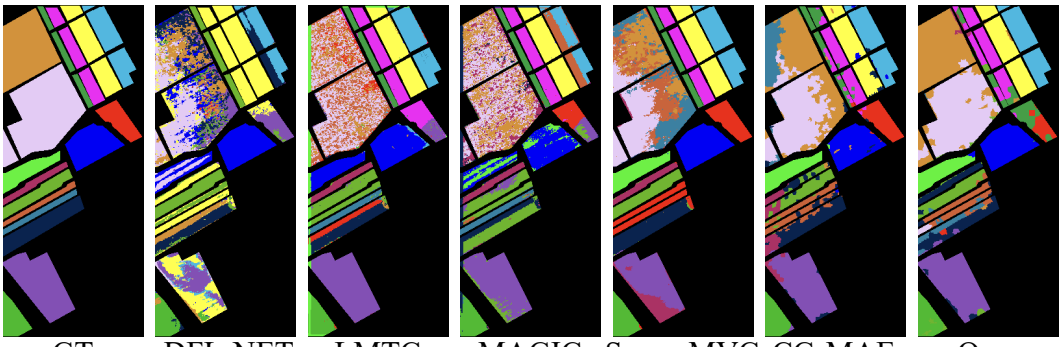


Figure 6: Additional qualitative comparison on the Salinas dataset. From left to right: Ground Truth (GT), DFL-NET, LMTC, MAGIC, SparseMVC, CG-MAE, and AM-GMAE (Ours). AM-GMAE produces more spatially coherent agricultural regions and suppresses scattered clustering errors while preserving the major class boundaries.

The MDAS results in Figure 7 further demonstrate the behavior of different methods in a more heterogeneous multi-sensor setting. The clustering maps obtained by several baselines contain frag-

mented regions and local semantic mixing, indicating the difficulty of directly integrating representations generated from substantially different sensing modalities. By contrast, AM-GMAE recovers larger homogeneous regions and preserves the dominant spatial structures more consistently. This observation supports the motivation of the proposed framework: CSAM improves the informativeness of masked reconstruction, DGPO suppresses unreliable neighborhood propagation, and MOT enables cross-view interaction without forcing heterogeneous representations into an identical latent geometry.

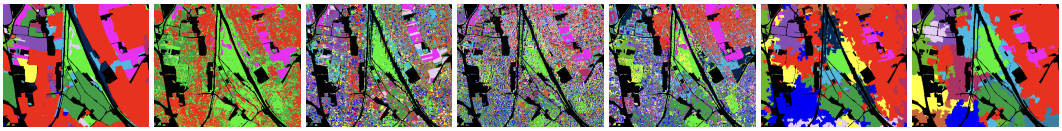


Figure 7: Additional qualitative comparison on the MDAS dataset. From left to right: Ground Truth (GT), DFL-NET, LMTC, MAGIC, SparseMVC, CG-MAE, and AM-GMAE (Ours). Compared with the competing methods, AM-GMAE yields more coherent semantic regions and reduces local fragmentation in the heterogeneous multi-sensor scene.

C.2 Hyperparameter Sensitivity Analysis

We further investigate the sensitivity of AM-GMAE to the major hyperparameters associated with the overall objective, structural regularization, masked reconstruction, and cross-view alignment. Specifically, we analyze the loss coefficients α, β, λ_1, and λ_2, the reconstruction exponent γ, the entropy regularization coefficient ε of MOT, the clustering temperature τ_c, and the Boltzmann masking temperature τ_m. Clustering accuracy (ACC) is used as the evaluation criterion on all four datasets. Overall, the results show that the response to individual hyperparameters is dataset dependent, which is expected because the four datasets differ substantially in the number of views, feature dimensionality, sensing modality, and graph structure. Nevertheless, relatively strong performance is maintained over multiple neighboring configurations rather than being restricted to a single isolated parameter setting.

Sensitivity to α and β. Figure 8 studies the interaction between the reconstruction coefficient α and the MOT coefficient β. The four datasets exhibit different response surfaces, indicating that the desired balance between view-specific reconstruction and cross-view semantic alignment depends on the degree of view heterogeneity. Trento exhibits a comparatively stable high-accuracy region across multiple coefficient combinations, whereas XuZhou, Salinas, and particularly MDAS show more pronounced variations when the relative contribution of the two objectives changes. These results indicate that neither reconstruction nor cross-view alignment should dominate the optimization indiscriminately. Instead, an appropriate balance allows the model to preserve informative view-specific structures while benefiting from semantic information shared across views.

Sensitivity to λ_1 and λ_2. We next evaluate the coefficients associated with the two DGPO objectives. As shown in Figure 9, changing λ_1 and λ_2 leads to different ACC variations across the four datasets. Trento remains relatively stable over a broad range, while the more challenging XuZhou, Salinas, and MDAS datasets show greater dependence on the balance between structural regularization and topology-consistent reconstruction. Excessively emphasizing either objective can weaken the complementary effect of the two structural constraints. In contrast, suitable combinations allow DGPO to simultaneously suppress unreliable graph connections and retain useful neighborhood information for reconstruction.

Sensitivity to the reconstruction exponent γ. Figure 10 reports the influence of γ, which controls the sensitivity of the masked reconstruction objective to reconstruction discrepancies. The results demonstrate a clear dataset-dependent behavior. Trento maintains comparatively stable performance across the tested settings, whereas the remaining datasets show larger variations. This observation

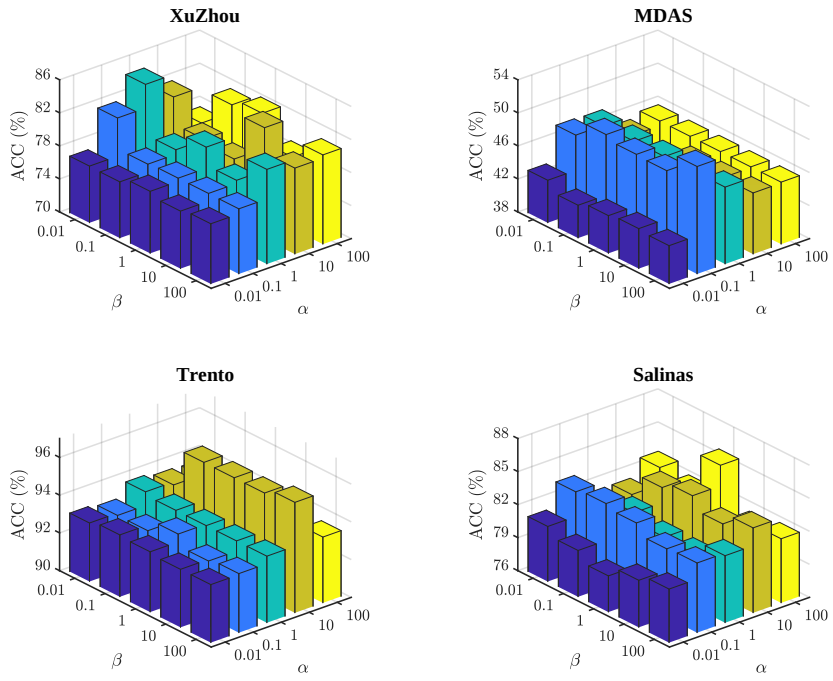


Figure 8: Sensitivity of clustering accuracy (ACC) to the loss coefficients α and β on XuZhou, MDAS, Trento, and Salinas. The two coefficients control the contributions of masked reconstruction and multi-marginal optimal transport, respectively. The preferred balance varies across datasets because of their different degrees of cross-view heterogeneity.

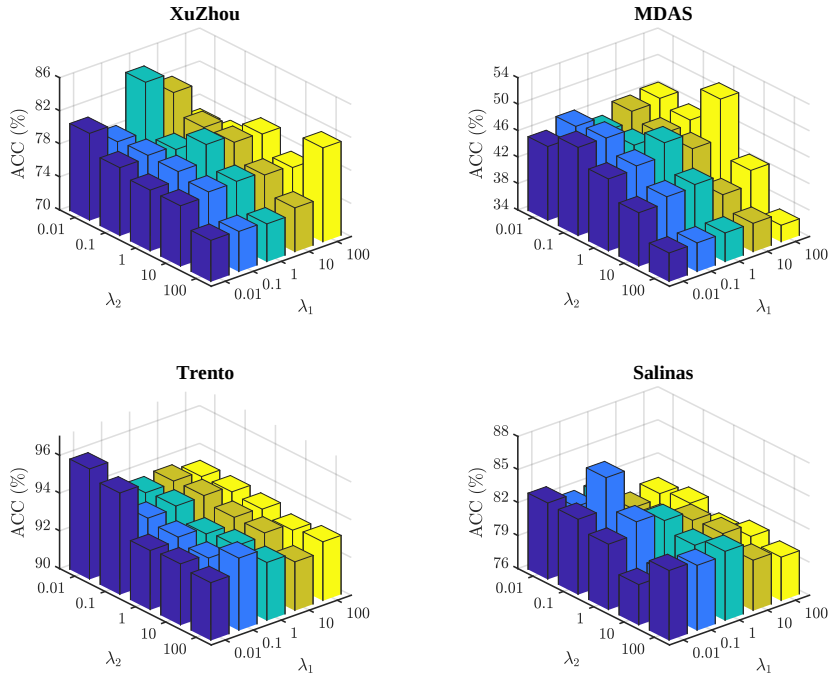


Figure 9: Sensitivity of clustering accuracy (ACC) to λ_1 and λ_2 on four datasets. Here, λ_1 weights the structural entropy objective $\mathcal{L}_{str}$ and λ_2 weights the topology-consistent reconstruction objective $\mathcal{L}_{topo}$.

suggests that the desired reconstruction emphasis depends on the difficulty of the underlying semantic structure and the amount of complementary information available from neighboring nodes and views. Importantly, the performance changes remain smooth as γ varies, indicating that AM-GMAE does not exhibit abrupt degradation under moderate changes of this parameter.

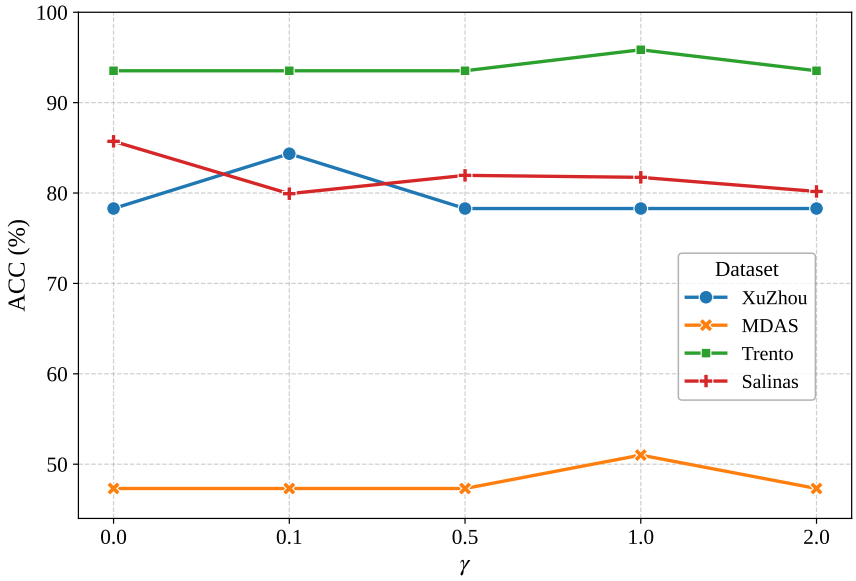


Figure 10: Sensitivity of clustering accuracy (ACC) to the reconstruction exponent γ on XuZhou, MDAS, Trento, and Salinas. Different datasets favor different reconstruction sensitivities, reflecting differences in feature distributions and reconstruction difficulty.

Sensitivity to the entropy regularization coefficient ε. The entropy regularization coefficient ε controls the smoothness of the transport plan in MOT and also affects the numerical behavior of Sinkhorn optimization. Figure **??** shows that AM-GMAE remains effective across a range of entropy regularization strengths, although the preferred region differs across datasets. A suitable ε provides a balance between concentrated semantic transport and sufficiently smooth optimization. When the regularization is altered substantially, the resulting transport plan may become either overly diffuse or insufficiently adaptive, leading to reduced clustering accuracy on some datasets. The different trends across the four datasets further support the use of dataset-specific transport regularization for heterogeneous multi-view data.

Sensitivity to the clustering temperature τ_c. Figure 12 analyzes the temperature τ_c used in the prototype-based soft cluster assignments of Eq. (1). Since τ_c determines the sharpness of the assignment distribution, it directly affects both the clustering uncertainty and the saliency information used by CSAM. The experimental results show distinct preferred temperature regions across the four datasets. A temperature that is too small can produce overly sharp assignments, whereas a larger temperature generates smoother semantic probabilities. The observed dataset-dependent optima indicate that an appropriate level of assignment confidence is important for estimating informative clustering-saliency signals.

Sensitivity to the masking temperature τ_m. Finally, Figure 13 evaluates the Boltzmann masking temperature τ_m in Eq. (4). This parameter controls how strongly the node sampling distribution concentrates on high-saliency nodes. A lower temperature increases the preference for nodes with large calibrated saliency scores, whereas a higher temperature produces a smoother masking distribution. The results show that the preferred masking selectivity differs among datasets: some datasets benefit from more concentrated saliency-guided sampling, while others favor a smoother distribution. This behavior is consistent with differences in cluster ambiguity and graph complexity across the four remote sensing scenes. Overall, the results demonstrate that CSAM can operate effectively under a

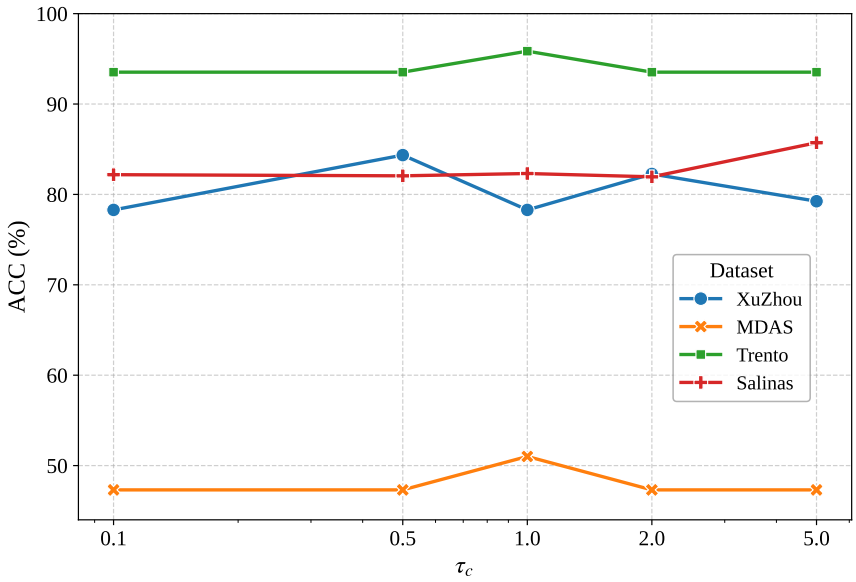


Figure 12: Sensitivity of clustering accuracy (ACC) to the clustering temperature τ_c on XuZhou, MDAS, Trento, and Salinas. The temperature controls the sharpness of the prototype-based soft cluster assignments used to estimate clustering uncertainty and saliency.

range of masking temperatures while retaining the ability to adapt the masking difficulty to dataset characteristics.

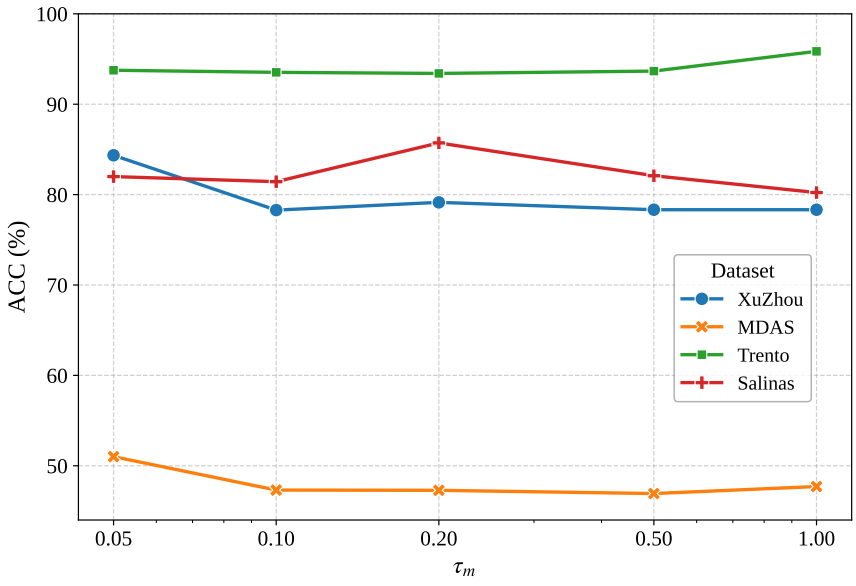


Figure 13: Sensitivity of clustering accuracy (ACC) to the Boltzmann masking temperature τ_m on four datasets. Smaller values produce more concentrated saliency-guided masking distributions, while larger values result in smoother node-selection probabilities.

In summary, the additional sensitivity experiments demonstrate that the main components of AM-GMAE exhibit interpretable responses to their corresponding hyperparameters. The objective coefficients determine the balance among masked reconstruction, structural refinement, and cross-view semantic alignment; γ regulates reconstruction sensitivity; ε controls the smoothness of the

optimal-transport coupling; and τ_c and τ_m respectively control clustering-assignment sharpness and adversarial masking selectivity. Although the preferred parameter regions differ among datasets, the results generally vary smoothly over neighboring configurations, showing that the effectiveness of AM-GMAE is not restricted to a single isolated hyperparameter configuration.

D Ethics Statement

This work adheres to the ICLR Code of Ethics. In this study, no human subjects or animal experimentation was involved. All datasets used, including Trento, Salinas, XuZhou, and MDAS, were sourced in compliance with relevant usage guidelines, ensuring no violation of privacy. We have taken care to avoid any biases or discriminatory outcomes in our research process. No personally identifiable information was used, and no experiments were conducted that could raise privacy or security concerns. We are committed to maintaining transparency and integrity throughout the research process.

E Reproducibility Statement

We have made every effort to ensure that the results presented in this paper are reproducible. The experimental setup, including training steps, model configurations, and hardware details, is described in detail in the paper. We have also provided a full description of CSAM, DGPO, MOT, to assist others in reproducing our experiments.

Additionally, Trento, Salinas, XuZhou, and MDAS are publicly available, ensuring consistent and reproducible evaluation results. We believe these measures will enable other researchers to reproduce our work and further advance the field.

F LLM Usage

Large Language Models (LLMs) were used to aid in the writing and polishing of the manuscript. Specifically, we used an LLM to assist in refining the language, improving readability, and ensuring clarity in various sections of the paper. The model helped with tasks such as sentence rephrasing, grammar checking, and enhancing the overall flow of the text.

It is important to note that the LLM was not involved in the ideation, research methodology, or experimental design. All research concepts, ideas, and analyses were developed and conducted by the authors. The contributions of the LLM were solely focused on improving the linguistic quality of the paper, with no involvement in the scientific content or data analysis.

The authors take full responsibility for the content of the manuscript, including any text generated or polished by the LLM. We have ensured that the LLM-generated text adheres to ethical guidelines and does not contribute to plagiarism or scientific misconduct.