Learning Subspace-Preserving Sparse Attention Graphs from Heterogeneous Multiview Data

Jie Chen[1,2], **Yuanbiao Gou**[2], **Chuanbin Liu**[3], **Zhu Wang**[4], **Xi Peng**[5*]
[1]National Key Laboratory of Fundamental Algorithms and Models for Engineering Simulation, Sichuan University, Chengdu, China
[2]College of Computer Science, Sichuan University, Chengdu, China
[3]School of Economics and Management, China University of Petroleum (Beijing), Beijing, China
[4]Law School, Sichuan University, Chengdu, China
[5]School of Artificial Intelligence, Sichuan University, Chengdu, China

Abstract

The high-dimensional features extracted from large-scale unlabeled data via various pretrained models with diverse architectures are referred to as heterogeneous multiview data. Most existing unsupervised transfer learning methods fail to faithfully recover intrinsic subspace structures when exploiting complementary information across multiple views. Therefore, a fundamental challenge involves constructing sparse similarity graphs that preserve these underlying subspace structures for achieving semantic alignment across heterogeneous views. In this paper, we propose a sparse attention graph learning (SAGL) method that learns subspace-preserving sparse attention graphs from heterogeneous multiview data. Specifically, we introduce a bilinear attention factorization scheme to capture asymmetric similarities among the high-dimensional features, which breaks the symmetry bottleneck that is inherent in the traditional representation learning techniques. A dynamic sparsity gating mechanism then predicts a feature-specific compression factor for adaptively controlling the topological contributions of neighbors. Furthermore, we employ a structured sparse projection via α-entmax to generate subspace-preserving sparse attention graphs for individual views. SAGL leverages these view-specific graphs to conduct sparse information aggregation, yielding discriminative representations for multiview learning tasks. In addition, we provide a rigorous theoretical analysis that bridges differentiable sparse attention and probability simplex constraints. Extensive experiments conducted on multiple benchmark datasets demonstrate that SAGL consistently outperforms the state-of-the-art unsupervised transfer learning approaches.

1 Introduction

By leveraging transfer learning with pretrained models possessing diverse architectures, high-dimensional features can be obtained from large-scale unlabeled data, yielding heterogeneous multiview data [1, 2]. These features typically lie close to low-dimensional structures corresponding to latent semantic categories [3]. Unsupervised knowledge transfer aims to bridge the discrepancy between such pretrained features and the intrinsic subspace structures for multiview learning.

Recent advances in unsupervised knowledge transfer techniques have fundamentally reshaped the representation learning process for downstream tasks [4–6]. By transferring knowledge from pretrained models, it has become possible to extract transferable features from large-scale unlabeled visual data across different domains. Numerous unsupervised transfer learning approaches have been developed

*Corresponding author: Xi Peng (pengx.gm@gmail.com)

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

to enforce the semantic alignment of features across multiple views, ensuring that their representations remain discriminative without the need for manual annotations [7–9]. By taking advantage of the extensive semantic priors of foundation vision-language models, these unsupervised transfer methods consistently produce results that are competitive with those of the traditional supervised fine-tuning approaches [10, 11]. When samples approximately lie in a union of low-dimensional subspaces, the samples derived from the same subspace tend to exhibit stronger intrinsic correlations with each other than with samples obtained from different subspaces. However, most existing unsupervised transfer learning approaches fail to effectively capture such subspace structures when completing unsupervised knowledge transfer tasks.

Multiview self-representation learning has achieved remarkable progress in terms of capturing the subspace structures that are embedded in multiview data [12–17]. Most existing self-representation learning-based methods can be broadly categorized into graph representation learning [14–16] and self-supervised graph learning techniques [12, 13, 18, 19]. These methods focus primarily on exploiting complementary information by recovering subspace structures while simultaneously enforcing a shared consensus graph across multiple views. However, when ℓ_0-norm or low-rank relaxations are integrated into shallow learning architectures, graph representation learning approaches require computationally prohibitive iterative solvers [15, 3, 20]. As a result, this severely restricts their scalability to large-scale unlabeled data. In addition, the traditional contrastive learning architectures adopted by self-supervised graph representation learning approaches limit their ability to capture rich semantic knowledge from heterogeneous multiview data. The α-entmax transformation was proposed as a formal generalization of the standard softmax function for inducing structured sparsity in attention mechanisms [21, 22]. The theoretical relationship between α-entmax projection and ℓ_0-norm sparse self-representation across multiple views deserves further investigation in heterogeneous multiview learning scenarios. Therefore, a fundamental challenge lies in constructing sparse similarity graphs that faithfully capture the intrinsic subspace structures embedded in heterogeneous multiview data.

In this paper, we propose a sparse attention graph learning (SAGL) method that learns subspace-preserving sparse attention graphs from heterogeneous multiview data. By reformulating the ℓ_0-norm sparse self-representation as a simplex-constrained differentiable optimization, we present an adaptive sparse self-representation learning framework that performs sparse information aggregation to produce discriminative representations. Specifically, we first introduce a bilinear attention factorization scheme to capture asymmetric similarities among features. This breaks the symmetry bottleneck inherent in traditional representation learning techniques. Then, we present a dynamic sparsity gating mechanism that adaptively controls the topological contribution of neighbors to facilitate rigorous intrasubspace isolation. Furthermore, we employ a structured sparse projection via α-entmax to generate the subspace-preserving sparse attention graphs for individual views. Compared to existing multiview graph learning methods, SAGL enables fully unsupervised transfer learning without computationally expensive iterative solvers. In addition, we provide a rigorous theoretical analysis bridging differentiable sparse attention and probability simplex constraints, demonstrating how structured sparsity enforces subspace-preserving representations. Extensive experiments demonstrate that SAGL consistently outperforms state-of-the-art unsupervised transfer learning approaches.

The key contributions of this work are summarized as follows.

- By reformulating sparse self-representation under probability simplex constraints, we present a sparse self-representation framework that learns subspace-preserving sparse attention graphs from heterogeneous multiview data. These graphs are employed to perform sparse information aggregation, yielding discriminative representations for performing unsupervised transfer learning without requiring computationally expensive iterative solvers.

- We introduce a bilinear attention factorization scheme to capture the asymmetric similarities among features. This scheme is enhanced by a dynamic sparsity gating mechanism that predicts a feature-specific compression factor to adaptively control the topological contributions of neighbors for implementing rigorous intrasubspace isolation.

- We provide rigorous theoretical guarantees that bridge differentiable sparse attention and probability simplex constraints, demonstrating how structured sparsity enforces subspace-preserving representations in sparse self-representation learning.

2 Related Work

2.1 Multiview Self-Representation Learning

The traditional self-representation learning approaches are aimed at uncovering the subspace structures embedded in high-dimensional data [23, 24, 3]. They typically assume that such data lie near low-dimensional subspaces corresponding to latent semantic categories [25, 3]. Inspired by these classic self-representation learning approaches, multiview self-representation learning methods have been developed to exploit complementary subspace information across multiple views [26–28, 15]. For example, Chen *et al.* [15] proposed a low-rank tensor graph learning approach that enforces the alignment of spectral block-diagonal representations for high-dimensional multiview data. However, the relaxations developed based on the nuclear norm require computationally prohibitive iterative solvers [25, 3, 20]. Xu *et al.* [28] introduced a self-representation learning framework (SRLF) that mitigates cross-modal discrepancies by deriving high-level semantic representations from low-level representations. However, this approach struggles to align heterogeneous feature distributions across multiple views. These studies indicate that integrating self-representation learning with view-specific structural constraints is essential for capturing the subspace structures of high-dimensional data.

2.2 Unsupervised Knowledge Transfer Learning

Transfer learning provides a fundamental paradigm for leveraging the semantic knowledge encoded in diverse pretrained models to enhance the generalizability of representation learning techniques. Unsupervised knowledge transfer is achieved without fine-tuning the large pretrained models. Recent advances in unsupervised knowledge transfer have been aimed at learning discriminative representations from large-scale unlabeled data without manual supervision [7, 9, 29]. For example, Gadetsky *et al.* [9] proposed an unsupervised transfer learning method that exploits transferable features to learn linearly separable representations without manual annotation. Chen *et al.* [7] introduced a multiview self-representation learning (MSRL) method that captures view-invariant representations by exploiting the self-representation property of features across heterogeneous views. These studies demonstrate that unsupervised knowledge transfer improves the general applicability of transfer learning in real-world scenarios without the need for manual annotation.

3 Sparse Attention Graph Learning

3.1 Adaptive Sparse Self-Representation Learning

Let $\mathcal{D}_{tr} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ be a training dataset consisting of n unlabeled samples, with each sample $\mathbf{x}_i \in \mathbb{R}^m$. By utilizing L distinct pretrained models $\{\phi_l\}_{l=1}^{L}$, a multiview dataset $\mathcal{H} = \{\mathbf{H}^{(1)}, \mathbf{H}^{(2)}, \ldots, \mathbf{H}^{(L)}\}$ is generated through transfer learning, where $\mathbf{H}^{(l)} = \left[\mathbf{h}_1^{(l)}, \mathbf{h}_2^{(l)}, ..., \mathbf{h}_n^{(l)}\right]$ $(1 \leq l \leq L)$, and $\mathbf{h}_i^{(l)} \in \mathbb{R}^{d_l}$ denotes the feature of sample $\mathbf{x}_i$, i.e.,

$$\mathbf{h}_i^{(l)} = \phi_l\left(\mathbf{x}_i\right). \tag{1}$$

When at least two of these L pretrained models adopt different pretraining objectives, $\mathcal{H}$ is referred to as heterogeneous multiview data. For simplicity, we utilize two pretrained models, namely, $\phi_\mu\left(\cdot\right)$ and $\phi_\nu\left(\cdot\right)$. Considering two linear models $\delta^{(\mu)}\left(\cdot\,;\,\cdot\right)$ and $\delta^{(\nu)}\left(\cdot\,;\,\cdot\right)$, the corresponding features $\mathbf{z}_i^{(\mu)} \in \mathbb{R}^C$ and $\mathbf{z}_i^{(\nu)} \in \mathbb{R}^C$ can be produced by

$$\mathbf{z}_i^{(\mu)} = \delta^{(\mu)}(\mathbf{h}_i^{(\mu)}; \mathbf{W}_\mu), \qquad \mathbf{z}_i^{(\nu)} = \delta^{(\nu)}(\mathbf{h}_i^{(\nu)}; \mathbf{W}_\nu) \tag{2}$$

where C represents the number of categories, and $\mathbf{W}_\mu \in \mathbb{R}^{d_\mu \times C}$ and $\mathbf{W}_\nu \in \mathbb{R}^{d_\nu \times C}$ denote the parameters of $\delta^{(\mu)}\left(\cdot\,;\,\cdot\right)$ and $\delta^{(\nu)}\left(\cdot\,;\,\cdot\right)$, respectively.

For the lth view, let $\mathbf{Z}^{(l)} = \left[\mathbf{z}_1^{(l)}, \mathbf{z}_2^{(l)}, \ldots, \mathbf{z}_n^{(l)}\right]^\top \in \mathbb{R}^{n \times C}$ denote the matrix of features. Each feature can be represented as a sparse linear combination of other features belonging to the same latent subspace. Specifically, the optimal sparse coefficient vector $\mathbf{a}_i^{(l)}$ can be obtained by solving the

following ℓ_0-norm minimization problem:

$$\widetilde{\mathbf{a}}_i^{(l)} = \arg\min_{\mathbf{a}_i^{(l)}} \left\| \mathbf{z}_i^{(l)} - \left(\mathbf{Z}^{(l)}\right)^\top \mathbf{a}_i^{(l)} \right\|_2^2 + \lambda \left\| \mathbf{a}_i^{(l)} \right\|_0 \tag{3}$$

where $\lambda > 0$ is a trade-off parameter, $\|\cdot\|_2$ denotes the Euclidean norm, and $\|\cdot\|_0$ denotes the number of nonzero entries contained in the vector. As the optimization problem in Eq. (3) is nonconvex and NP-hard, it is typically relaxed to a sparsity-inducing ℓ_1-norm minimization problem. However, relying on traditional iterative solvers for this convex relaxation is computationally prohibitive when applied to large-scale multiview data.

To integrate the ℓ_0-norm self-representation strategy into an end-to-end differentiable learning framework, we reformulate the sparse representation problem as a simplex-projected attention scheme. Let $\Delta^{n-1} := \left\{ \mathbf{a}_i^{(l)} \mid \mathbf{1}^\top \mathbf{a}_i^{(l)} = 1,\ \mathbf{a}_i^{(l)} \geq \mathbf{0} \right\}$ denote the $(n-1)$-dimensional probability simplex. We formulate the simplex-constrained sparse self-representation problem as follows:

$$\min_{\mathbf{a}_i^{(l)} \in \Delta^{n-1}} \|\mathbf{a}_i^{(l)}\|_0 + \lambda \left\| \mathbf{z}_i^{(l)} - \left(\mathbf{Z}^{(l)}\right)^\top \mathbf{a}_i^{(l)} \right\|_2^2 \quad \text{s.t.} \quad \mathbf{z}_i^{(l)} = \left(\mathbf{Z}^{(l)}\right)^\top \mathbf{a}_i^{(l)} + \mathbf{e}_i^{(l)} \tag{4}$$

where $\mathbf{e}_i^{(l)} \in \mathbb{R}^C$ denotes the sparse error term. Instead of algebraic optimization solutions, our goal is to adaptively learn subspace-preserving sparse attention graphs $\left\{ \mathbf{A}^{(l)} \in \mathbb{R}^{n \times n} \right\}_{l=1}^{L}$ for the multiview dataset $\mathcal{H}$, where $\mathbf{A}^{(l)} = \left\{ \mathbf{a}_1^{(l)}, \mathbf{a}_2^{(l)}, ..., \mathbf{a}_n^{(l)} \right\}$, and an attention coefficient $a_{ij}^{(l)} \in \mathbf{A}^{(l)}$ $(1 \leq i, j \leq n)$ denotes the similarity between the features $\mathbf{h}_i^{(l)}$ and $\mathbf{h}_j^{(l)}$ in the lth view.

Given a structured sparse attention graph $\mathbf{A}^{(l)}$ $(l \in \{\mu, \nu\})$, the sparse information aggregation operation on $\mathbf{z}_i^{(l)}$, denoted as $f\left(\mathbf{z}_i^{(l)}; \mathbf{a}_i^{(l)}\right)$, is defined as follows:

$$\mathbf{p}_i^{(l)} = f\left(\mathbf{z}_i^{(l)}; \mathbf{a}_i^{(l)}\right) \tag{5}$$

where $\mathbf{a}_i^{(l)}$ represents the ith column vector of $\mathbf{A}^{(l)}$, and $\mathbf{p}_i{}^{(l)} \in \mathbb{R}^C$ denotes the aggregated representation of the sample $\mathbf{x}_i$ for the lth view. To align the heterogeneous distributions across multiple views, we present an adaptive sparse self-representation learning framework. Specifically, we formulate the process of constructing subspace-preserving sparse attention graphs as an adaptive sparse representation learning problem:

$$\min_{\mathbf{W}_\mu, \mathbf{W}_\nu, \mathbf{A}^{(\mu)}, \mathbf{A}^{(\nu)}} \sum_{i=1}^{n} \mathcal{L}\left(\mathbf{p}_i^{(\mu)}, \mathbf{p}_i^{(\nu)}\right) + \lambda \Omega(\mathbf{A}^{(\mu)}, \mathbf{A}^{(\nu)}) \text{ s.t. } \left(\mathbf{A}^{(l)}\right)^\top \mathbf{1} = \mathbf{1}, \mathbf{A}^{(l)} \geq \mathbf{0}, l \in \{\mu, \nu\} \tag{6}$$

where $\mathcal{L}$ denotes a self-supervised alignment loss, and $\Omega(\cdot, \cdot)$ denotes the structural sparsity constraint imposed on the attention graphs. The fundamental challenge encountered when minimizing the objective function in Eq. (6) lies in efficiently optimizing the network weights to learn the subspace-preserving sparse attention graphs $\mathbf{A}^{(l)}$ while satisfying the imposed sparse structural constraints.

3.2 Learning Subspace-Preserving Sparse Attention Graphs

According to Eq. (4), subspace-preserving sparse attention graphs determine the topological neighborhoods for feature-specific aggregation. These graphs can be constructed by evaluating the similarities between the features. However, the standard attention mechanisms inevitably result in densely connected graphs. Such dense connectivity leads to heavily entangled representations across different latent subspaces. We take advantage of the underlying subspace structures embedded in high-dimensional features to learn subspace-preserving sparse attention graphs.

3.2.1 Bilinear Attention Factorization

We introduce a bilinear attention factorization scheme that projects the encoded features $\mathbf{Z}^{(l)}$ into two latent semantic subspaces via two learnable projection matrices $\mathbf{U}^{(l)} \in \mathbb{R}^{C \times C}$ and $\mathbf{V}^{(l)} \in \mathbb{R}^{C \times C}$.

The raw similarity matrix $\mathbf{S}^{(l)} \in \mathbb{R}^{n\times n}$ is computed as follows:

$$\mathbf{S}^{(l)} = \frac{(\mathbf{Z}^{(l)}\mathbf{U}^{(l)})(\mathbf{Z}^{(l)}\mathbf{V}^{(l)})^\top}{\sqrt{C}} \tag{7}$$

where the scaling factor $\frac{1}{\sqrt{C}}$ prevents the dot products from growing excessively large in high-dimensional feature spaces. The linear projection matrices $\mathbf{U}^{(l)}$ and $\mathbf{V}^{(l)}$ transform the encoded linear features into different latent semantic subspaces before computing global pairwise dot products. The similarity matrix $\mathbf{S}^{(l)}$ captures the global pairwise semantic structure among all the samples in the lth view. Unlike the standard inner products that inherently produce symmetric undirected graphs, the introduction of two distinct projection matrices $\mathbf{U}^{(l)}$ and $\mathbf{V}^{(l)}$ generates an asymmetric similarity matrix $\mathbf{S}^{(l)}$. This fundamental asymmetry effectively facilitates discriminative information aggregation among the features for self-representation learning.

3.2.2 Dynamic Sparsity Gate

Since the features derived from distinct latent subspaces exhibit varying degrees of topological uncertainty, applying an isotropic threshold across all features inevitably compromises the structural integrity of the learned graphs. To achieve precise intrasubspace isolation, it is imperative to dynamically adapt the sparsity level for each individual feature. To this end, we propose a dynamic sparsity gate that is formulated as a lightweight multilayer perceptron (MLP). For each feature $\mathbf{z}_i^{(l)}$, the gate predicts a feature-specific compression factor $\omega_i^{(l)} \in [0, 1]$:

$$\omega_i^{(l)} = \sigma\left(\mathbf{W}_2\text{ReLU}\left(\mathbf{W}_1\mathbf{z}_i^{(l)}\right)\right) \tag{8}$$

where $\sigma(\cdot)$ denotes the sigmoid activation function, and $\mathbf{W}_1$ and $\mathbf{W}_2$ are learnable weight matrices. The directed similarities are subsequently adapted by this dynamic gate as follows:

$$\tilde{\mathbf{s}}_i^{(l)} = \mathbf{s}_i^{(l)} \cdot \left(1 - \omega_i^{(l)}\right) \tag{9}$$

where $\mathbf{s}_i^{(l)}$ denotes the ith column of $\mathbf{S}^{(l)}$. A larger $\omega_i^{(l)}$ imposes a stronger penalty on the similarities, effectively forcing the model to concentrate on a narrower set of highly confident neighbors in the same latent subspace. As the feature-specific compression factor $\omega_i^{(l)}$ is dynamically learned during the optimization, we also present an alternative scheme to compute the directed similarities as follows:

$$\tilde{s}_{ij}^{(l)} = \frac{s_{ij}^{(l)}}{1 - \omega_i^{(l)} + \epsilon} \tag{10}$$

where ϵ is a small constant, e.g., $\epsilon = 10^{-6}$. This scheme explicitly breaks the structural symmetry, i.e., $\tilde{s}_{ij}^{(l)} \neq \tilde{s}_{ji}^{(l)}$. Consequently, the factor $\omega_i^{(l)}$ adaptively controls the topological contributions of the neighbors during the feature aggregation.

3.2.3 Structured Sparse Projection

The bilinear attention factorization scheme can capture the global pairwise semantic information among all the features. However, it inevitably assigns nonzero similarities to all feature pairs regardless of their underlying subspace affiliations. This indicates that the similarity matrix $\mathbf{S}^{(l)}$ remains densely connected, which violates the independent subspace assumption underlying the sparse self-representation learning. Structural sparsity can be enforced for the similarity matrix $\mathbf{S}^{(l)}$ by eliminating any spurious connections between features derived from different latent subspaces. Specifically, we project each column of the similarity matrix $\mathbf{S}^{(l)}$ onto the $(n-1)$-dimensional probability simplex Δ^{n-1} to obtain a subspace-preserving sparse attention graph $\mathbf{A}^{(l)}$. The standard softmax function is intrinsically incapable of producing absolute zeros. Instead of utilizing the standard softmax function, the subspace-preserving sparse attention graph $\mathbf{A}^{(l)}$ is generated by employing the α-entmax transformation [21, 22]: $\mathbf{a}_i^{(l)} = \text{entmax}_\alpha\left(\tilde{\mathbf{s}}_i^{(l)}\right)$. For any $\alpha > 1$, the entmax mapping process induces sparse solutions via a thresholding mechanism:

$$\mathbf{a}_i^{(l)} = \left[(\alpha-1)\left(\tilde{\mathbf{s}}_i^{(l)} - \tau_i^{(l)}\mathbf{1}\right)\right]_+^{\frac{1}{\alpha-1}} \tag{11}$$

where $\tau_i^{(l)} \in \mathbb{R}$ is a unique threshold. By setting $\alpha = 1.5$, the transformation acts as a strict continuous thresholding operator. This operator forces the marginal intersubspace similarities to exactly zero. By adaptively adjusting the factor $\omega_i^{(l)}$, the projection assigns stronger inhibition to features with higher topological uncertainty while preserving the meaningful intrasubspace connections for features with distinct subspace affiliations. Consequently, it produces a subspace-preserving sparse attention graph $\mathbf{A}^{(l)}$ that reveals the block-diagonal structures of the underlying subspaces.

Discussion. The strict thresholding property of α-entmax with $\alpha = 1.5$ provides a principled mechanism for recovering the block-diagonal subspace structures from the given features. Specifically, the α-entmax projection inherently maps low-confidence similarities to exact zeros via Tsallis entropy regularization. The discrete ℓ_0-norm sparsity and simplex constraints in Eq. (4) are transformed into an end-to-end differentiable formulation in Eq. (6) by parameterizing $\mathbf{A}^{(l)}$ with the α-entmax transformation. The structural sparsity constraint $\Omega(\mathbf{A}^{(\mu)}, \mathbf{A}^{(\nu)})$ in Eq. (6) is implicitly implemented through this differentiable thresholding mechanism, which produces sparse attention graphs with $\mathbf{A}^{(l)} \neq \mathbf{I}$. Consequently, this projection ensures that the proposed SAGL model consistently produces directed sparse attention graphs without additional regularization terms. SAGL reformulates ℓ_0-norm self-representation as a differentiable projection onto a continuous probability simplex via α-entmax. By integrating bilinear asymmetric factorization with dynamic gating, it jointly optimizes high-level feature representations and sparse self-representation graphs in an end-to-end manner, while preserving the underlying subspace structure without relying on iterative solvers.

3.2.4 Sparse Information Aggregation

To obtain view-specific representations, we introduce a sparse information aggregation strategy that integrates sparse representation learning with a residual connection. The aggregated representation $\mathbf{p}_i^{(l)}$ is formulated as follows:

$$\mathbf{p}_i^{(l)} = f\left(\mathbf{z}_i^{(l)}; \mathbf{a}_i^{(l)}\right) = \sum_{j=1}^{n} a_{ij}^{(l)} \mathbf{z}_j^{(l)} + \mathbf{z}_i^{(l)} \tag{12}$$

where $\mathbf{a}_i^{(l)}$ represents the ith column vector of $\mathbf{A}^{(l)}$. Unlike aggregating the information derived from all features, the sparse aggregation operation leverages the highly sparse nature of $\mathbf{a}_i^{(l)}$. This ensures that $\mathbf{p}_i^{(l)}$ is represented as a sparse linear combination of spatially proximate neighbors belonging to the same latent subspace. By strictly restricting the aggregation procedure to the intrinsic subspace, this operation filters out intersubspace noise. This significantly enhances the discriminative capabilities of representations.

The probability simplex constraint $\left(\mathbf{A}^{(l)}\right)^\top \mathbf{1} = \mathbf{1}$, $\mathbf{A}^{(l)} \geq \mathbf{0}$ directly corresponds to the sparse self-representation constraint in Eq. (6). The sparse information aggregation uses the sparsity of $\mathbf{A}^{(l)}$ to refine the representations in Eq. (12). By restricting the aggregation stage to intrinsic subspaces, the mechanism filters out intersubspace noise and produces more discriminative representations. For the lth view, the assignment probability distribution of the representation $\mathbf{p}_i^{(l)}$, i.e., $\mathbf{q}_i^{(l)} \in \Delta^{C-1}$, corresponding to sample $\mathbf{x}_i$ $(1 \leq i \leq n)$, is calculated as follows:

$$\mathbf{q}_i^{(l)} = \rho\left(\mathbf{p}_i^{(l)}\right) \tag{13}$$

where $\rho(\cdot)$ is a softmax activation function and Δ^{C-1} denotes the probability simplex. The complete optimization objective of SAGL is provided in the appendix.

3.3 Theoretical Justification

3.3.1 Asymmetry of Bilinear Attention Factorization

The underlying relationship among high-dimensional features is typically directed. The directional dependency implies that the similarity relationship from $\mathbf{z}_j^{(l)}$ to $\mathbf{z}_i^{(l)}$ is inherently asymmetric. We provide a theoretical analysis of the asymmetric nature of bilinear attention factorization. The assumptions and proofs of all presented theorems are provided in the appendix.

Theorem 3.1 (Universal Representation Capacity for Asymmetric Subspaces)**.** *Let a linear feature matrix $\mathbf{Z} \in \mathbb{R}^{n\times C}$ have full column rank, and let $\mathbf{A}^* \in \mathbb{R}^{n\times n}$ be an arbitrary asymmetric adjacency matrix defining the directed similarity relationships among different subspaces. If $\mathbf{A}^*$ resides within the bilinear span of the features, i.e., $\mathbf{A}^* = \mathbf{Z}\mathbf{W}^*\mathbf{Z}^\top$, for some latent structural matrix $\mathbf{W}^* \in \mathbb{R}^{C\times C}$, the bilinear attention factorization scheme implemented via projection matrices $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{C\times C}$ is guaranteed to possess universal expressive power for the exact recovery of $\mathbf{A}^*$.*

Theorem 3.1 provides a rigorous theoretical justification for employing two distinct projection matrices, i.e., $\mathbf{U}^{(l)}$ and $\mathbf{V}^{(l)}$, in Eq. (7). Specifically, it demonstrates that bilinear attention factorization has universal expressive power for pursuing the arbitrary directed similarity relationships of latent subspaces. The traditional representation learning techniques often evaluate similarity relationships via symmetric projections, which forces the underlying structural matrix $\mathbf{W}^*$ to be strictly symmetric. In contrast, the proposed bilinear attention factorization scheme successfully breaks the symmetry bottleneck of standard inner product operations. This enables the proposed SAGL model to faithfully capture the directed relationships inherent in high-dimensional features.

3.3.2 Rethinking ℓ_0-Norm Self-Representation as Simplex-Projected Attention

We establish a theoretical bridge between the ℓ_0-norm self-representation and the differentiable probability simplex-projected attention. The probability simplex constraint provides a principled convex relaxation that preserves the subspace-preserving property of the original ℓ_0-norm solution.

Definition 3.2 (Subspace-Preserving Representation)**.** A sparse coefficient vector $\mathbf{a}_i^{(l)} \in \mathbb{R}^n$ is said to be *subspace-preserving* for a feature $\mathbf{z}_i^{(l)} \in \mathcal{S}_k$ if $a_{ij}^{(l)} \neq 0$ implies that $\mathbf{z}_j^{(l)} \in \mathcal{S}_k$; i.e., all nonzero coefficients index features from the same latent subspace.

Theorem 3.3 (ℓ_0-Norm-to-Simplex Relaxation)**.** *Let the features $\{\mathbf{z}_i^{(l)}\}_{i=1}^n$ be drawn from a union of linear subspaces $\{\mathcal{S}_k\}_{k=1}^K$. Under the independent subspace, sufficient intrasubspace sampling, and similarity separation assumptions (Assumptions A.1, A.2, and A.3, respectively), the following subspace-preserving properties hold:*

(1) Subspace-Preserving Property of the ℓ_0-Norm Solution: *Consider the ℓ_0-norm sparse self-representation problem:*

$$\widetilde{\mathbf{a}}_i^{(l)} = \arg\min_{\mathbf{a}_i^{(l)}} \left\|\mathbf{a}_i^{(l)}\right\|_0 \quad \textit{s.t.} \quad \mathbf{z}_i^{(l)} = \left(\mathbf{Z}^{(l)}\right)^\top \mathbf{a}_i^{(l)}, \quad a_{ii}^{(l)} = 0. \tag{14}$$

Then the solution $\widetilde{\mathbf{a}}_i^{(l)}$ is guaranteed to be subspace-preserving, i.e., $\widetilde{a}_{ij}^{(l)} \neq 0 \implies \mathbf{z}_j^{(l)} \in \mathcal{S}_{y_i}$.

(2) Exact Subspace Preservation via Simplex Projection: *Let $\widetilde{\mathbf{a}}_i^{(l)} = \textit{entmax}_\alpha(\tilde{\mathbf{s}}_i^{(l)})$ for $\alpha > 1$, where $\tilde{\mathbf{s}}_i^{(l)}$ is the similarity vector. Consider the entmax-regularized attention projection:*

$$\widetilde{\mathbf{a}}_i^{(l)} = \arg\max_{\mathbf{a}_i^{(l)} \in \Delta^{n-1}} \langle \mathbf{a}_i^{(l)}, \tilde{\mathbf{s}}_i^{(l)} \rangle - \Omega_\alpha(\mathbf{a}_i^{(l)}), \tag{15}$$

where $\Omega_\alpha(\mathbf{a}_i^{(l)})$ is the Tsallis entropy for $\alpha > 1$. If the directed similarities satisfy the similarity separation condition (Assumption A.3) with a sufficiently large margin $\epsilon > 0$, then the resulting sparse attention is exactly subspace-preserving, i.e.,

$$\widetilde{a}_{ij}^{(l)} > 0 \implies \mathbf{z}_j^{(l)} \in \mathcal{S}_{y_i}. \tag{16}$$

Theorem 3.3 establishes the fact that the probability simplex constraint $\left(\mathbf{A}^{(l)}\right)^\top \mathbf{1} = \mathbf{1}, \mathbf{A}^{(l)} \geq \mathbf{0}$ in Eq. (6) provides a principled convex relaxation of the ℓ_0-norm self-representation constraint. Under strict separation between intra- and intersubspace similarities, this thresholding mechanism guarantees that only intrasubspace connections remain active, thereby preserving the subspace structure. By employing α-entmax with $\alpha = 1.5$, we obtain a continuous analog of the ℓ_0 sparsity operator. This formulation facilitates the end-to-end differentiable optimization process without relying on iterative algebraic solvers. Consequently, the subspace-preserving property established for simplex projection inherently extends to entmax-based attention under analogous separation conditions.

Table 1: Clustering performance of competing methods on eight datasets.

Methods	Pets			KITTI			Flowers			Caltech101			EuroSAT			SUN397			Food101			ImageNet-1K		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
SRLF	95.45	95.76	91.97	42.18	13.64	8.60	94.06	98.48	95.74	65.27	87.69	46.27	95.68	90.25	90.76	63.68	82.89	53.59	91.53	92.20	85.72	69.41	85.07	56.09
PRR	85.28	80.99	78.87	45.06	16.14	11.42	94.59	93.17	90.92	78.28	72.95	64.59	80.86	56.41	59.50	64.37	80.88	51.18	92.81	92.48	88.28	72.97	68.53	58.92
MIM	83.89	87.89	76.49	45.30	15.46	10.61	97.50	96.77	95.33	86.90	83.62	78.66	87.66	56.34	61.34	66.40	75.32	57.13	92.20	79.62	80.61	63.24	84.72	40.18
TURTLE	96.05	96.14	92.79	45.59	16.51	11.88	99.59	99.66	99.40	90.14	94.59	88.16	96.38	91.61	92.17	71.18	85.13	59.01	95.02	94.97	90.95	71.98	88.12	61.72
MSRL	97.38	97.18	95.00	53.74	19.39	16.94	99.72	99.76	99.59	90.38	94.69	86.92	99.88	99.67	99.73	70.26	85.33	59.61	95.07	95.13	91.09	74.65	89.17	65.14
SAGL	**97.57**	**97.33**	**95.35**	**56.95**	**24.73**	**26.12**	**99.76**	**99.80**	**99.65**	**91.54**	**95.35**	**89.01**	**99.92**	**99.80**	**99.82**	**73.34**	**86.29**	**62.27**	**97.68**	**97.42**	**95.53**	**79.34**	**90.98**	**70.50**

Table 2: A comparison of performance between SAGL and the supervised LR baselines.

Methods	Pets	KITTI	Flowers	Caltech101	EuroSAT	SUN397	Food101	ImageNet-1K
LR+SigLIP 2	96.76	73.33	99.66	**97.22**	96.14	**83.52**	96.91	**87.04**
LR+DINOv3	97.08	**76.60**	99.72	94.51	97.40	79.80	95.81	86.87
SAGL	**97.57**	56.95	**99.76**	91.54	**99.92**	73.34	**97.68**	79.34

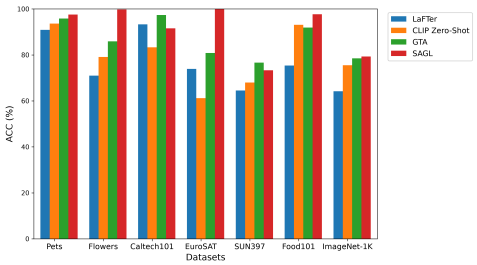


Figure 1: A comparison of performance between SAGL and the zero-shot transfer learning baselines.

4 Experiments

We conduct extensive experiments to evaluate the effectiveness of the proposed SAGL method on eight publicly available datasets: Pets, KITTI, Flowers, Caltech101, EuroSAT, SUN397, Food101 and ImageNet-1K [30]. Specifically, we consider three transfer learning tasks: self-supervised, zero-shot, and supervised transfer learning. For self-supervised transfer learning, three standard metrics are employed to evaluate the clustering performance of all competing methods, i.e., clustering accuracy (ACC), normalized mutual information (NMI) and adjusted rand index (ARI) [7]. The source code of SAGL is publicly available at `https://github.com/chenjie20/SAGL`. We compare SAGL against representative state-of-the-art methods, including SRLF [28], principle of rate reduction (PRR) [2], masked image modeling (MIM) [1], MSRL [7] and TURTLE [9]. For zero-shot transfer learning tasks, CLIP zero-shot transfer [9], LaFTer [31] and a generation, transduction and adaptation (GTA) framework [32] are included as baselines. For supervised transfer learning, we employ a supervised linear probe as the baseline. We utilize two representative pretrained models, i.e., DINOv3 [33] and SigLIP 2 [34], as backbones. The source codes of all the baselines are provided by their respective authors.

4.1 Performance Evaluation

4.1.1 Self-Supervised Transfer Learning

For self-supervised transfer learning, we follow the standard unsupervised evaluation protocol [9, 7], where label information is unavailable during training. The proposed SAGL model is trained on the training set of each dataset and evaluated on the test set without any fine-tuning. The number of clusters is set to the ground-truth number of semantic categories for each dataset. The clustering results of all competing methods on the self-supervised learning task are reported in Table 1, where the best and second-best results are highlighted in bold and underlined, respectively. We observe that SAGL consistently outperforms the competing methods in terms of ACC, NMI, and ARI. In particular, the performance gap becomes increasingly evident on larger datasets, such as SUN397, Food101 and ImageNet-1K. For example, SAGL improves ACC by 2.16%, 2.61%, and 4.69% on SUN397, Food101, and ImageNet-1K, respectively. SRLF does not exploit the generalization capability of pretrained backbones for clustering tasks. Although PRR and MIM incorporate pretrained backbones, they rely on individual views that often fail to provide sufficient information for clustering tasks. In

contrast, SAGL, MSRL and TURTLE employ various pretrained backbones on large-scale unlabeled data to produce multiple views, thereby effectively capturing complementary information across different views. Additionally, SAGL consistently outperforms both MSRL and TURTLE across all the datasets. This is because SAGL further considers the intrinsic subspace structures embedded in the high-dimensional multiview data. Consequently, these results validate the effectiveness of the proposed SAGL approach for self-supervised learning tasks.

Table 3: Clustering performance of SAGL with different pretrained model combinations.

Methods	Pets			KITTI			Flowers			Caltech101			EuroSAT			SUN397			Food101			ImageNet-1K		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
(1)	94.63	95.82	91.73	52.14	21.01	20.01	99.71	99.75	99.56	85.39	92.46	83.72	99.62	99.00	99.16	63.49	82.09	52.39	92.11	93.52	87.36	72.45	87.97	62.07
(2)	93.19	95.07	89.79	52.41	18.60	15.34	69.78	87.47	65.28	77.38	87.44	82.38	92.60	88.35	85.48	59.88	81.48	49.93	**98.32**	**98.08**	**96.72**	71.36	87.59	61.60
(3)	90.79	94.15	85.72	48.93	24.53	14.84	84.55	93.41	45.75	88.53	94.21	82.46	88.76	81.66	77.22	69.81	84.17	32.59	94.02	94.67	88.45	75.85	89.36	47.75
(4)	92.31	95.58	89.95	52.21	22.16	17.85	88.24	95.10	59.99	89.74	94.22	85.44	99.72	99.32	99.38	65.71	81.58	11.83	97.60	97.59	95.51	75.75	89.72	65.50
SAGL	**97.57**	**97.33**	**95.35**	**56.95**	**24.73**	**26.12**	**99.76**	**99.80**	**99.65**	**91.54**	**95.35**	**89.01**	**99.92**	**99.80**	**99.82**	**73.34**	**86.29**	**62.27**	97.68	97.42	95.53	**79.34**	**90.98**	**70.50**

Table 4: Training time (seconds) with different pretrained backbone combinations.

Combination	Pets	KITTI	Flowers	Caltech101	EuroSAT	SUN397	Food101	ImageNet-1K
(1)	25	20	8	4	18	100	147	2,404
(2)	16	24	128	18	71	62	59	4,724
(3)	32	4	21	14	77	155	785	6,886
(4)	268	25	13	17	46	181	311	6,296
SAGL	113	20	17	85	73	99	362	1,409

4.1.2 Zero-Shot Transfer Learning

The zero-shot transfer learning baselines, such as CLIP zero-shot transfer (ZST), LaFTer, and GTA, utilize descriptions of ground-truth classes as a form of supervision. Under the inductive setting, CLIP ZST and GTA employ CLIP ViT-L/14 as their backbone, whereas LaFTer adopts ViT-B/32. Fig. 1 shows the clustering accuracies of SAGL and these zero-shot transfer learning baselines across all datasets except KITTI. The KITTI dataset is not included due to its relatively small number of clusters. SAGL outperforms these weakly supervised baselines across most datasets without requiring any class descriptions or label information.

4.1.3 Supervised Transfer Learning

We employ a supervised linear probe as an important evaluation protocol to measure the performance gap between SAGL and supervised baselines. The probe is implemented via a standard logistic regression (LR) model. Specifically, for the supervised LR baselines, we freeze the pretrained backbones and train only a linear classifier on top of the features extracted by the individual pretrained backbones (i.e., DINOv3 and SigLIP 2) using all available labels. Table 2 shows a comparison between the clustering accuracy of SAGL and the classification accuracy of the supervised LR baselines across eight datasets. Despite the absence of label information during training, SAGL outperforms the supervised LR baselines on the Pets, Flowers, EuroSAT, and Food101 datasets. While the supervised baselines are often limited by the features produced by individual pretrained backbones, SAGL effectively leverages the complementary information of multiple views to produce more discriminative representations, even without label supervision.

4.2 Evaluating Various Pretrained Model Combinations

We construct feature pairs from different pretrained backbones, including DINOv3, SigLIP 2, CLIP ViT-L/14, and ConvNeXt V2. Specifically, we consider four combinations: (1) ConvNeXt V2 and DINOv3, (2) ConvNeXt V2 and SigLIP 2, (3) SigLIP 2, DINOv3, and ConvNeXt V2, and (4) SigLIP 2, DINOv3, and CLIP ViT-L/14. Table 3 shows the clustering performance of SAGL with four pretrained model combinations across eight datasets. Incorporating additional views is generally expected to provide richer complementary information. However, the results indicate that increasing the number of views does not consistently lead to improved clustering performance. For example, SAGL with two-view combinations achieves better clustering performance than with three-view combinations, such as (3) and (4). This is because highly uncorrelated views provide limited complementary information while increasing computational cost. These findings highlight the importance of carefully selecting view combinations that maximize complementary information rather than simply increasing the number of views. In addition, model pairs sharing the same architecture achieve substantially

higher clustering results than heterogeneous pairs combining different architectures. For example, the results show that a combination of ViT-based models consistently outperforms a mixed-architecture combination across most datasets. The large distributional gap between heterogeneous architecture pairs makes it difficult for the SAGL model to recover the intrinsic subspace structures across views.

Table 5: Ablation study on the key components of the SAGL model.

Methods	Pets			KITTI			Flowers			Caltech101			EuroSAT			SUN397			Food101			ImageNet-1K		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
SAGL_i	96.18	96.07	92.95	50.53	17.57	15.56	99.07	99.54	97.98	90.19	94.14	87.28	96.40	91.72	92.22	70.21	84.53	58.02	95.13	94.99	91.11	73.97	88.79	64.28
SAGL_b	63.61	72.98	50.04	52.47	21.85	19.36	99.32	99.66	99.41	80.08	88.20	70.45	86.34	76.86	73.43	31.99	57.39	2.70	67.72	77.91	54.66	18.86	47.04	0.24
SAGL_w	97.22	**97.34**	94.96	55.08	21.71	22.98	99.37	99.71	99.48	90.52	94.54	87.79	97.12	93.65	93.76	71.24	86.11	61.17	95.68	95.74	92.06	77.54	90.32	68.32
SAGL_s	96.57	96.78	93.80	54.55	23.80	23.96	99.06	99.47	99.20	90.35	93.54	84.72	99.04	97.67	97.89	71.83	85.94	60.27	96.02	96.62	93.54	77.21	90.74	69.12
SAGL	**97.57**	97.33	**95.35**	**56.95**	**24.73**	**26.12**	**99.76**	**99.80**	**99.65**	**91.54**	**95.35**	**89.01**	**99.92**	**99.80**	**99.82**	**73.34**	**86.29**	**62.27**	**97.68**	**97.42**	**95.53**	**79.34**	**90.98**	**70.50**

We evaluate how computational cost grows as additional pretrained backbones are incorporated. As reported in Table 4, increasing the number of views from two (Combinations (1) and (2)) to three (Combinations (3) and (4)) dramatically increases training time on large-scale datasets. For example, on ImageNet-1K, the training time increases from 2,404s for SAGL with two views (Combination (1)) to 6,886s and 6,296s for Combinations (3) and (4), respectively. Moreover, incorporating additional backbones does not consistently improve clustering performance over the default two-view configuration (SigLIP 2 + DINOv3). On ImageNet-1K, our two-view SAGL achieves an ACC of 79.34%, whereas the three-view combinations obtain lower ACC values of 75.85% and 75.75%, respectively. These results demonstrate that the default two-view configuration (SigLIP 2 + DINOv3) provides a favorable balance between clustering accuracy and computational efficiency.

4.3 Ablation Study

We conduct an ablation study to verify the contribution of each key component in SAGL. Specifically, we consider three variants of SAGL: (1) replacing the sparse attention graphs with identity matrices, denoted as SAGL_i, (2) retaining only the dense attention graphs generated by bilinear attention factorization, denoted as SAGL_b, and (3) removing the dynamic sparsity gating mechanism, denoted as SAGL_w. Additionally, we evaluate a symmetric variant of SAGL, denoted by SAGL_s, which enforces identical projection matrices. Table 5 reports the clustering performance of these four variants against the full SAGL model. The full SAGL model significantly outperforms its variants, SAGL_i, SAGL_b, SAGL_w and SAGL_s. For example, the performance gap between SAGL and SAGL_i becomes particularly significant on the large-scale datasets. When the structured sparse projection is replaced with dense attention graphs, the clustering performance of SAGL_b drops dramatically across most datasets compared to the full SAGL model. The clustering results of SAGL_w also highlight the crucial role of the dynamic sparsity gate in adaptive neighbor selection. Additionally, the full asymmetric bilinear attention consistently outperforms its symmetric variant across all benchmark datasets. The asymmetric factorization enhances the ability of sparse representations to capture directed inter-sample dependencies. The results demonstrate the effectiveness of the sparse information aggregation operation in Eq. (12).

5 Conclusion

In this paper, we present the SAGL method for learning subspace-preserving sparse attention graphs from heterogeneous multiview data. SAGL utilizes bilinear attention factorization to capture asymmetric similarities among high-dimensional features, successfully breaking the symmetry constraints in traditional representation learning techniques. Moreover, we introduce a dynamic sparsity gating mechanism that adaptively controls the topological contributions of neighbors by predicting a feature-specific compression factor. To enforce structured sparsity, we adopt a structured sparse projection via α-entmax to generate subspace-preserving sparse attention graphs. By leveraging these view-specific graphs, SAGL conducts sparse information aggregation to produce discriminative representations for unsupervised transfer learning without the need for computationally expensive iterative solvers. Finally, we provide a rigorous theoretical analysis bridging the gap between differentiable sparse attention and probability simplex constraints. This analysis demonstrates how structured sparsity enforces subspace-preserving representations. Extensive experiments demonstrate that SAGL consistently outperforms state-of-the-art unsupervised transfer learning approaches.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (NSFC) under Grants 62576231 and 62677067, in part by Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China under Grant JYB2025XDXM610, in part by the Fundamental Research Funds for the Central Universities under Grant CJ202303 and 20826041K4049, in part by the Sichuan University Graduate 'AI+X' Curriculum Development Project under Grant GSAIX2026034, and in part by the Open Project Program of the Omni-media Technology and Communication Cognition Key Laboratory of Sichuan Province.

References

[1] B. Alkin, L. Miklautz, S. Hochreiter, and J. Brandstetter, "MIM-Refiner: A contrastive learning boost from intermediate pre-trained representations," in *Proc. 13th Int. Conf. Learn. Represent.*, Singapore, Apr. 2025, pp. 1–37.

[2] T. Chu, S. Tong, T. Ding, X. Dai, B. D. Haeffele, R. Vidal, and Y. Ma, "Image clustering via the principle of rate reduction in the age of pretrained models," in *Proc. 12th Int. Conf. Learn. Represent.*, Vienna, Austria, May 2024, pp. 1–12.

[3] M. Brbić and I. Kopriva, "l_0-motivated low-rank sparse subspace clustering," *IEEE Trans. on Cyber.*, vol. 50, no. 4, pp. 1711–1725, 2020.

[4] I. Ng, Y. Li, Z. Li, Y. Zheng, G. Chen, and K. Zhang, "A general representation-based approach to multi-source domain adaptation," in *Proc. 42nd Int. Conf. Mach. Learn.*, Vancouver, Canada, Jul. 2025, pp. 45 911–45 933.

[5] J. Wang, J. Li, W. Zhuang, C. Chen, L. Lyu, and F. Ma, "Enhancing foundation models with federated domain knowledge infusion," in *Proc. 42nd Int. Conf. Mach. Learn.*, Vancouver, Canada, Jul. 2025, pp. 63 621–63 635.

[6] Z. Chen, Y. Deng, Y. Li, and Q. Gu, "Understanding transferable representation learning and zero-shot transfer in CLIP," in *Proc. 12th Int. Conf. Learn. Represent.*, Vienna, Austria, May 2024, pp. 1–12.

[7] J. Chen, Z. Wang, C. Liu, and X. Peng, "Multiview self-representation learning across heterogeneous views," *arXiv preprint arXiv:2602.04328*, pp. 1–12, Jan. 2026.

[8] C. Liu, R. Bing, X. Xi, W. Dai, and G. Yuan, "Heterogeneous graph structure learning for experts selection in academic evaluation," *IEEE Trans. Comput. Soc. Syst.*, vol. 12, no. 6, pp. 4677–4688, 2025.

[9] A. Gadetsky, Y. Jiang, and M. Brbic, "Let go of your labels with unsupervised transfer," in *Proc. 41st Int. Conf. Mach. Learn.*, Vienna, Austria, Jul. 2024, pp. 14 382–14 407.

[10] M. Fu, K. Zhu, Z. Ding, and J. Wu, "DTL: Parameter- and memory-efficient disentangled vision learning," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 48, no. 2, pp. 1736–1749, Feb. 2026.

[11] C. Shang, M. Li, Y. Zhang, Z. Chen, J. Wu, F. Gu, Y. Lu, and Y. Cheung, "PRO-VPT: Distribution-adaptive visual prompt tuning via prompt relocation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Hawaii, USA, Oct. 2025, pp. 1558–1568.

[12] Q. Wang, B. Zhao, Z. Zhang, Q. Gao, and L. Jiao, "Dual self-supervised deep graph clustering," *IEEE Trans. Multi.*, pp. 1–10, Jan. 2026.

[13] G. Li, Z. Yu, K. Yang, J. Lv, and C. L. P. Chen, "Pseudo-label similarity graph-driven multi-view contrastive clustering," *IEEE Trans. Multi.*, pp. 1–13, Feb. 2026.

[14] B. Jiang, C. Zhang, X. Liang, P. Zhou, J. Yang, X. Wu, J. Guan, W. Ding, and W. Sheng, "Collaborative similarity fusion and consistency recovery for incomplete multi-view clustering," in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 17, Philadelphia, Pennsylvania, USA, Feb. 2025, pp. 17 617–17 625.

[15] J. Chen, H. Mao, W. L. Woo, C. Liu, Z. Wang, and X. Peng, "One-step adaptive graph learning for incomplete multiview subspace clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 37, no. 5, pp. 2771–2783, May 2025.

[16] M. Zhang, X. Liu, T. Han, X. Qu, and S. Niu, "Adaptive anchor-guided representation learning for efficient multi-view subspace clustering," *IEEE Trans. Image Process.*, vol. 34, pp. 6053–6067, Sept. 2025.

[17] L. Fei, J. He, Q. Zhu, S. Zhao, J. Wen, and Y. Xu, "Deep multi-view contrastive clustering via graph structure awareness," *IEEE Trans. Image Process.*, vol. 34, pp. 3805–3816, Jun. 2025.

[18] B. Deng, T. Wang, L. Fu, S. Huang, C. Chen, and T. Zhang, "THESAURUS: contrastive graph clustering by swapping fused gromov-wasserstein couplings," in *Proc. AAAI Conf. Artif. Intell.*, Philadelphia, Pennsylvania, USA, Feb. 2025, pp. 16 199–16 207.

[19] C. Liu, X. Zhang, H. Zhao, Z. Liu, X. Xi, and L. Yu, "LMCBert: An automatic academic paper rating model based on large language models and contrastive learning," *IEEE Trans. Cybern.*, vol. 55, no. 6, pp. 2970–2979, 2025.

[20] E. Elhamifar and R. Vidal, "Sparse subspace clustering algorithm, theory, and applications," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, 2013.

[21] B. Peters, V. Niculae, and A. F. T. Martins, "Sparse sequence-to-sequence models," in *Proc. 57th Annu. Meet. Assoc. Comput. Linguist.*, Florence, Italy, Jul. 2019, pp. 1504–1519.

[22] A. F. F. Martins and R. F. Astudillo, "From softmax to sparsemax: A sparse model of attention and multi-label classification," in *Proc. 33rd Int. Conf. Mach. Learn.*, New York, USA, Jun. 2016, pp. 1614–1623.

[23] J. Chen, H. Mao, Y. Gou, and X. Peng, "Hierarchical sparse representation clustering for high-dimensional data streams," *IEEE Trans. Neural. Netw. Learn. Syst.*, vol. 36, no. 10, pp. 18 035–18 047, Oct. 2025.

[24] S. Huang, Y. Zhang, L. Fu, and S. Wang, "Learnable multi-view matrix factorization with graph embedding and flexible loss," *IEEE Trans. on Multi.*, vol. 25, pp. 3259–3272, 2022.

[25] L. Fu, Z. Chen, Y. Chen, and S. Wang, "Unified low-rank tensor learning and spectral embedding for multi-view subspace clustering," *IEEE Trans. on Multi.*, vol. 25, pp. 4972–4985, 2022.

[26] J. Wen, J. Long, X. Lu, C. Liu, X. Fang, and Y. Xu, "Partial multiview incomplete multilabel learning via uncertainty-driven reliable dynamic fusion," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 48, no. 1, pp. 236–250, Jan. 2026.

[27] J. Yang, C. Y. Lu, Z. Wang, H. T. Chen, G. K. Xu, C. Zhang, S. Dong, X. Liang, and B. Jiang, "Multi-view clustering with granularity-aware pseudo supervision," in *Proc. AAAI Conf. Artif. Intell.*, vol. 40, no. 19, Singapore, Jan. 2026, pp. 27 538–27 546.

[28] H. Xu, X. Su, S. Chen, G. Chen, and X. Chen, "Bridging optimization and neural networks for efficient multi-view clustering," in *Proc. AAAI Conf. Artif. Intell.*, vol. 40, no. 19, Singapore, Jan. 2026, pp. 16 066–16 074.

[29] C. Liu, J. Guo, X. Zhang, D. Wu, and L. Yu, "Expert credibility prediction model based on fuzzy C-means clustering and similarity association," *IEEE Trans. on Cyber.*, vol. 33, no. 8, pp. 2719–2729, 2025.

[30] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, and et.al, "Pytorch: an imperative style, high-performance deep learning library," in *Proc. 33rd Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2019, pp. 8026–8037.

[31] M. J. Mirza, L. Karlinsky, W. Lin, H. Possegger, M. K. R. Feris, and H. Bischof, "LaFTer: Label–free tuning of zero-shot classifier using language and unlabeled image collections," in *Proc. 37th Adv. Neural Inf. Process. Syst.*, New Orleans, Louisiana, USA, Dec. 2023, pp. 5765–5777.

[32] O. Saha, L. Lawrence, G. V. Horn, and S. Maji, "Generate, transduct, adapt: Iterative transduction with vlms," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Hawaii, USA, Oct. 2025, pp. 1369–1379.

[33] O. Sim'eoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, F. Massa, D. Haziza, L. Wehrstedt, J. Wang, T. Darcet, T. Moutakanni, L. Sentana, C. Roberts, A. Vedaldi, J. Tolan, J. Brandt, C. Couprie, J. Mairal, H. J'egou, P. Labatut, and P. Bojanowski, "DINOv3," pp. 1–67, 2025.

[34] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa, O. Hénaff, J. Harmsen, A. Steiner, and X. Zhai, "SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features," pp. 1–20, 2025.

[35] D. P. Kingma and J. L. Ba, "ADAM: a method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent.*, San Diego, CA, USA, May 2015, pp. 1–15.

[36] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Mar. 2011.

[37] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton, "Similarity of neural network representations revisited," in *Proc. 36th Int. Conf. Mach. Learn.*, California, USA, Jun. 2019, pp. 3519–3529.

[38] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res*, vol. 9, no. 11, pp. 2579–2605, 2008.

Algorithm 1 Optimization Procedure for SAGL

Input: An unlabeled training dataset $\mathcal{D}_{tr} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ and a testing dataset $\mathcal{D}_{ts} = \{\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots, \hat{\mathbf{x}}_{\hat{n}}\}$ belonging to C categories, and L pretrained models.
Parameters: The number of training epochs $epochs$, batch size B, and hyperparameters $\gamma > 0$ and $\beta > 0$.
Output: The predicted labels $\mathbf{Y} = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{\hat{n}}]$ for $\mathcal{D}_{ts}$.

Construct the multiview data $\mathcal{H}_{tr}$ from $\mathcal{D}_{tr}$ via Eq. (1);
for $t = 1$ **to** $epochs$ **do**
  for each mini-batch $\mathcal{B}$ of size B sampled from $\mathcal{H}_{tr}$ **do**
    for $l = 1$ **to** L **do**
      Compute the linear features $\mathbf{Z}^{(l)}$ for samples in $\mathcal{B}$ via Eq. (2);
      Compute the raw similarity matrix $\mathbf{S}^{(l)}$ via Eq. (7);
      Compute the directed similarities $\tilde{\mathbf{S}}^{(l)}$ via Eq. (9);
      Construct the subspace-preserving sparse attention graph $\mathbf{A}^{(l)}$ via Eq. (11);
      Compute the assignment probability distribution $\mathbf{q}_i^{(l)}$ for each sample $\mathbf{x}_i \in \mathcal{B}$ via Eq. (13);
    end for
    Determine the pseudolabel $\hat{y}_i$ for each sample $\mathbf{x}_i \in \mathcal{B}$ via Eq. (17);
    Update the network parameters by minimizing $\mathcal{L}$ in Eq. (22);
  end for
end for
Construct the multiview data $\mathcal{H}_{ts}$ from $\mathcal{D}_{ts}$ via Eq. (1);
Perform forward computation on $\mathcal{H}_{ts}$ to predict the labels $\hat{y}_i$ for all samples $\hat{\mathbf{x}}_i \in \mathcal{D}_{ts}$ via Eq. (17).

A Technical Appendices

A.1 Enforcing Multiview Alignment under Structural Sparsity Constraints

To align the heterogeneous representations across multiple views in a fully unsupervised manner, SAGL is optimized by jointly minimizing three complementary components: pseudolabel prediction, cluster diversity regularization, and cross-view consistency alignment [7]. We optimize the entire network using an adaptive momentum-based mini-batch gradient descent strategy [35], with the pretrained backbones remaining frozen during training. The complete optimization procedure of the proposed SAGL method is summarized in Algorithm 1.

Pseudolabel Prediction Loss. Consider the view-specific representations $\mathbf{q}_i^{(l)}$ produced by the sparse information aggregation process described in Eq. (13). To enable self-supervised training to be implemented in the absence of ground-truth annotations, we construct pseudolabels by aggregating the predicted soft assignments across all L views:

$$\hat{y}_i = \arg\max_k \frac{1}{L} \sum_{l=1}^{L} q_{ik}^{(l)} \tag{17}$$

where $q_{ik}^{(l)}$ denotes the predicted probability of sample $\mathbf{x}_i$ belonging to class k ($1 \leq k \leq C$) in the lth view. A negative log-likelihood loss is applied to refine each view-specific prediction:

$$\mathcal{L}_{pseudo} = -\frac{1}{n} \sum_{l=1}^{L} \sum_{i=1}^{n} \log q_{i\hat{y}_i}^{(l)}. \tag{18}$$

This loss encourages each independent view to produce predictions that align with the global consensus distribution.

Cluster Diversity Regularization. Unsupervised transfer learning-based clustering models are highly prone to producing trivial solutions, where all instances collapse into a single underlying subspace (i.e., a single cluster). We explicitly maximize the marginal entropy of the global cluster assignments. Let $\bar{\mathbf{q}}_j^{(l)}$ denote the average assignment probability of all instances that are assigned to

category j. Cluster diversity can be achieved by minimizing the negative entropy:

$$\mathcal{L}_{div} = \sum_{l=1}^{L} \sum_{j=1}^{C} \bar{q}_j^{(l)} \log \bar{q}_j^{(l)} \tag{19}$$

where

$$\bar{q}_j^{(l)} = \frac{1}{n} \sum_{i=1}^{n} q_{i,j}^{(l)}. \tag{20}$$

Cross-View Consistency Alignment. The semantic assignment probability distributions $\mathbf{q}_i^{(a)}$ and $\mathbf{q}_i^{(b)}$ produced by view a and view b, respectively, should exhibit strong semantic alignment. To enforce semantic consistency across heterogeneous views (a, b), we minimize the cross-entropy among the soft assignment distributions of different pretrained models:

$$\mathcal{L}_{align} = -\frac{1}{n} \sum_{i=1}^{n} \sum_{a=1}^{L} \sum_{b=1, a \neq b}^{L} \sum_{j=1}^{C} q_{i,j}^{(a)} \log q_{i,j}^{(b)}. \tag{21}$$

Overall Objective. The overall optimization objective of SAGL is defined as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{pseudo} + \gamma \mathcal{L}_{div} + \beta \mathcal{L}_{align} \tag{22}$$

where γ and β are the trade-off parameters. The overall objective $\mathcal{L}_{total}$ represents a specific implementation of the general self-supervised alignment loss $\mathcal{L}\left(\mathbf{p}_i^{(\mu)}, \mathbf{p}_i^{(\nu)}\right)$ presented in Eq. (6).

Limitations. SAGL relies on the availability of sufficient neighbors within the same latent subspace to perform sparse information aggregation. A limitation is introduced during mini-batch training when sample shuffling is applied, as it may lead to an insufficient number of samples from certain semantic categories within a single batch. This violates the sufficient intrasubspace sampling assumption, which requires each feature's neighborhood to contain at least d_k linearly independent samples from its corresponding subspace to ensure faithful recovery of the underlying subspace structures. In such cases, the α-entmax projection may fail to establish robust and meaningful connections. As a result, this may degrade the separability of the learned representations and reduce overall clustering performance.

A.2 Proof of Theorem 3.1

Proof. Let $\mathbf{A}^* = \mathbf{Z}\mathbf{W}^*\mathbf{Z}^\top$ denote the target adjacency matrix. The objective is to show that there exist matrices $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{C \times C}$ such that

$$\frac{1}{\sqrt{C}} \mathbf{Z}\mathbf{U}\mathbf{V}^\top \mathbf{Z}^\top = \mathbf{A}^*. \tag{23}$$

Since $\mathbf{W}^*$ is an arbitrary (potentially asymmetric) matrix, we can perform singular value decomposition (SVD) on the scaled target $\sqrt{C}\mathbf{W}^*$:

$$\sqrt{C}\mathbf{W}^* = \mathbf{P}\mathbf{\Sigma}\mathbf{Q}^\top \tag{24}$$

where $\mathbf{P}, \mathbf{Q} \in \mathbb{R}^{C \times C}$ are orthogonal matrices containing the left and right singular vectors, respectively, and $\mathbf{\Sigma} \in \mathbb{R}^{C \times C}$ is the diagonal matrix of singular values. We can explicitly construct the optimal projection matrices as follows:

$$\mathbf{U} = \mathbf{P}\mathbf{\Sigma}^{1/2}, \quad \mathbf{V} = \mathbf{Q}\mathbf{\Sigma}^{1/2}. \tag{25}$$

Since the diagonal entries of $\mathbf{\Sigma}$ are nonnegative singular values, $\mathbf{\Sigma}^{1/2}$ is a strictly real-valued matrix, which ensures that $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{C \times C}$. Then, we have

$$\begin{aligned} \mathbf{U}\mathbf{V}^\top &= (\mathbf{P}\mathbf{\Sigma}^{1/2})(\mathbf{Q}\mathbf{\Sigma}^{1/2})^\top \\ &= \mathbf{P}\mathbf{\Sigma}^{1/2}\mathbf{\Sigma}^{1/2}\mathbf{Q}^\top \\ &= \mathbf{P}\mathbf{\Sigma}\mathbf{Q}^\top \\ &= \sqrt{C}\mathbf{W}^* \end{aligned} \tag{26}$$

Upon substituting Eq. (26) back into the bilinear attention factorization formulation given in Eq. (7), we obtain

$$\begin{aligned}\mathbf{A} &= \frac{1}{\sqrt{C}}\mathbf{Z}(\mathbf{U}\mathbf{V}^\top)\mathbf{Z}^\top \\ &= \frac{1}{\sqrt{C}}\mathbf{Z}(\sqrt{C}\mathbf{W}^*)\mathbf{Z}^\top \\ &= \mathbf{Z}\mathbf{W}^*\mathbf{Z}^\top = \mathbf{A}^*\end{aligned} \tag{27}$$

Therefore, $\mathbf{U}$ and $\mathbf{V}$ always exist such that the bilinear attention factorization exactly recovers any arbitrary asymmetric adjacency matrix $\mathbf{A}^*$, where $\mathbf{U}$ and $\mathbf{V}$ are constructed independently from the left and right singular vectors, respectively. □

A.3 Proof of Theorem 3.3

Proof. **(1)** Let $\mathbf{z}_i^{(l)} \in \mathcal{S}_k$. Under the independent subspace assumption, any linear feature $\mathbf{z}_i^{(l)} \in \mathcal{S}_k$ cannot be represented as a linear combination of features derived from $\mathcal{S}_{k'}$ $(k' \neq k)$. Consequently, assigning coefficients for the features located outside $\mathcal{S}_k$ does not reduce the reconstruction error $\left\|\mathbf{z}_i^{(l)} - \mathbf{Z}^{(l)}\mathbf{a}_i^{(l)}\right\|_2^2$. Conversely, this strictly increases the ℓ_0-norm penalty. Thus, the optimal solution $\widetilde{\mathbf{a}}_i^{(l)}$ that minimizes Eq. (14) must assign zero to all cross-subspace features, resulting in a subspace-preserving representation.

(2) The entmax operator for $\alpha > 1$ is defined as follows:

$$\text{entmax}_\alpha(\tilde{\mathbf{s}}_i^{(l)})_j = [(\alpha-1)\tilde{s}_{ij}^{(l)} - \tau_i^{(l)}]_+^{1/(\alpha-1)}, \tag{28}$$

where $[\cdot]_+ = \max(0, \cdot)$ and $\tau_i^{(l)} \in \mathbb{R}$ is the unique threshold determined by the simplex constraint $\sum_j a_{ij}^{(l)} = 1$.

Let $\mathbf{z}_i^{(l)} \in \mathcal{S}_{y_i}$. Under Assumption A.3, a margin $\epsilon > 0$ separates the intrasubspace similarities from the intersubspace similarities:

$$\min_{j:\mathbf{z}_j^{(l)} \in \mathcal{S}_{y_i}} \tilde{s}_{ij}^{(l)} > \max_{j:\mathbf{z}_j^{(l)} \notin \mathcal{S}_{y_i}} \tilde{s}_{ij}^{(l)} + \epsilon. \tag{29}$$

Step 1: Intersubspace sparsity. If the separation margin ϵ is sufficiently large (combined with the sufficient intrasubspace sampling of $\mathcal{S}_{y_i}$ from Assumption A.2), then the threshold $\tau_i^{(l)}$ strictly exceeds the values of all intersubspace similarity terms $(\alpha-1)\tilde{s}_{ij}^{(l)}$ for $j : \mathbf{z}_j^{(l)} \notin \mathcal{S}_{y_i}$; i.e.,

$$\tau_i^{(l)} \geq (\alpha-1) \max_{j:\mathbf{z}_j^{(l)} \notin \mathcal{S}_{y_i}} \tilde{s}_{ij}^{(l)}. \tag{30}$$

For any cross-subspace feature $\mathbf{z}_j^{(l)} \notin \mathcal{S}_{y_i}$, we have

$$(\alpha-1)\tilde{s}_{ij}^{(l)} - \tau_i^{(l)} \leq 0, \ \forall j : \mathbf{z}_j^{(l)} \notin \mathcal{S}_{y_i}. \tag{31}$$

This implies that $\widetilde{a}_{ij}^{(l)} = 0, \ \forall j : \mathbf{z}_j^{(l)} \notin \mathcal{S}_{y_i}$.

Step 2: Intrasubspace sparsity. To satisfy the unit sum constraint $\sum_j a_{ij}^{(l)} = 1$, the threshold $\tau_i^{(l)}$ must be bounded from above by the similarity of the active intrasubspace elements. Thus, there exists a threshold $\tau_i^{(l)}$ such that

$$(\alpha-1) \max_{j:\mathbf{z}_j^{(l)} \notin \mathcal{S}_{y_i}} \tilde{s}_{ij}^{(l)} \leq \tau_i^{(l)} < (\alpha-1) \max_{j:\mathbf{z}_j^{(l)} \in \mathcal{S}_{y_i}} \tilde{s}_{ij}^{(l)}. \tag{32}$$

In $\mathcal{S}_{y_i}$, the threshold $\tau_i^{(l)}$ is not required to be below all intrasubspace similarities. On the contrary, $\tau_i^{(l)}$ is uniquely determined by the simplex constraint and the distribution of intrasubspace similarities. Specifically, for $j : \mathbf{z}_j^{(l)} \in \mathcal{S}_{y_i}$:

$$\widetilde{a}_{ij}^{(l)} > 0 \iff (\alpha-1)\tilde{s}_{ij}^{(l)} > \tau_i^{(l)}. \tag{33}$$

By refining $\tilde{s}_{ij}^{(l)}$ to amplify the similarity gap within $\mathcal{S}_{y_i}$, the dynamic sparsity gating mechanism ensures that only the most spatially proximate intrasubspace neighbors satisfy the connection condition, resulting in a sparse attention graph.

Therefore, the support set $\mathcal{A}_i = \{j : \widetilde{a}_{ij}^{(l)} > 0\}$ contains only a specific subset of intrasubspace features, i.e., $|\mathcal{A}_i| \ll |\mathcal{S}_{y_i}|$, which successfully preserves the subspace structures while enforcing the sparse attention property of the ℓ_0-norm. □

A.4 Theoretical Bridge: Investigating Block-Diagonal Structures in Directed Similarity Graphs

Assumption A.1 (Independent Subspaces)**.** A collection of K latent subspaces $\{\mathcal{S}_k\}_{k=1}^{K}$ is linearly independent if for any $k \in \{1, \ldots, K\}$, the following holds:

$$\mathcal{S}_k \cap \sum_{j \neq k} \mathcal{S}_j = \{\mathbf{0}\}. \tag{34}$$

Assumption A.2 (Sufficient Intrasubspace Sampling)**.** For any feature $\mathbf{z}_i^{(l)} \in \mathcal{S}_k$, its neighborhood set $\mathcal{N}_i$ contains at least d_k linearly independent samples from $\mathcal{S}_k$, such that

$$\text{span}\left(\{\mathbf{z}_j^{(l)}\}_{j \in \mathcal{N}_i \cap \mathcal{S}_k}\right) = \mathcal{S}_k, \tag{35}$$

where $d_k = \dim(\mathcal{S}_k)$.

Assumption A.3 (Similarity Separation)**.** For any feature $\mathbf{z}_i^{(l)} \in \mathcal{S}_k$, the directed similarities satisfy a margin condition: there exists a sufficiently large margin $\epsilon > 0$ such that

$$\max_{j:\mathbf{z}_j^{(l)} \notin \mathcal{S}_k} \tilde{s}_{ij}^{(l)} < \min_{j:\mathbf{z}_j^{(l)} \in \mathcal{S}_k} \tilde{s}_{ij}^{(l)} - \epsilon, \tag{36}$$

where $\tilde{s}_{ij}^{(l)}$ denotes the directed similarity between $\mathbf{z}_i^{(l)}$ and $\mathbf{z}_j^{(l)}$.

Theorem A.4 (Sparse Attention Recovers Block-Diagonal Structures)**.** *Under Assumption A.1, let $\mathbf{A}^{(l)} \in \mathbb{R}^{n \times n}$ be the attention graph produced via the α-entmax projection method with $\alpha = 1.5$. If the bilinear similarity $\tilde{\mathbf{s}}_i^{(l)}$ satisfies the separation condition (Assumption A.3), there exists a permutation matrix $\mathbf{P}$ such that the permuted matrix*

$$\widetilde{\mathbf{A}}^{(l)} = \mathbf{P}\mathbf{A}^{(l)}\mathbf{P}^\top \tag{37}$$

has a block-diagonal structure, i.e.,

$$\widetilde{\mathbf{A}}^{(l)} = \begin{pmatrix} \mathbf{B}_1^{(l)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_2^{(l)} & \cdots & \mathbf{0} \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{B}_K^{(l)} \end{pmatrix} \tag{38}$$

where $\mathbf{B}_k^{(l)} \in \mathbb{R}^{n_k \times n_k}$ denotes the intrasubspace attention submatrix corresponding to $\mathcal{S}_k$.

Proof. Under Assumption A.1, each feature $\mathbf{z}_i^{(l)}$ belongs to a unique subspace $\mathcal{S}_k$. According to Theorem 3.3, the entmax operator with $\alpha > 1$ produces a subspace-preserving sparse representation. Specifically, for each $\mathbf{z}_i^{(l)} \in \mathcal{S}_k$, we have

$$\widetilde{a}_{ij}^{(l)} = 0, \quad \forall j \text{ such that } \mathbf{z}_j^{(l)} \notin \mathcal{S}_k. \tag{39}$$

Therefore, the nonzero entries of $\mathbf{A}^{(l)}$ are confined to pairs (i, j) with $\mathbf{z}_i^{(l)}, \mathbf{z}_j^{(l)} \in \mathcal{S}_k$ for some k.

We consider a permutation matrix $\mathbf{P}$ that reorders the indices such that all samples acquired from the same subspace are grouped together. Let

$$\widetilde{\mathbf{A}}^{(l)} = \mathbf{P}\mathbf{A}^{(l)}\mathbf{P}^\top. \tag{40}$$

Since the cross-subspace entries are zero, $\widetilde{\mathbf{A}}^{(l)}$ has a block-diagonal structure, i.e.,

$$\widetilde{\mathbf{A}}^{(l)} = \begin{pmatrix} \mathbf{B}_1^{(l)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_2^{(l)} & \cdots & \mathbf{0} \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{B}_K^{(l)} \end{pmatrix}, \tag{41}$$

where $\mathbf{B}_k^{(l)} \in \mathbb{R}^{n_k \times n_k}$ corresponds to the intrasubspace attention structure contained in $\mathcal{S}_k$. Thus, $\mathbf{A}^{(l)}$ is block-diagonal up to a permutation of the indices. □

Theorem A.4 suggests that the sparse attention graph $\mathbf{A}^{(l)}$ can recover a complete block-diagonal structure if two conditions are satisfied: independent subspaces (Assumption A.1) and the similarity separation condition (Assumption A.3). Under these conditions, the α-entmax projection maps all cross-subspace attention weights to exactly zero regardless of feature positions. Because the memberships of features in different subspaces are determined by relative similarities rather than absolute locations, the block-diagonal structure is preserved for unseen test samples. This guarantees consistent clustering performance across both training and test sets.

A.5 Theoretical Bridge: From the ℓ_0-Norm to Simplex Projection

The sparse self-representation lies in the discrete ℓ_0-norm constraint in Eq. (4), which strictly limits the number of nonzero reconstruction coefficients, i.e., $\|\mathbf{a}_i^{(l)}\|_0 \ll n$. However, directly optimizing the ℓ_0-norm over high-dimensional data is an NP-hard and strictly nondifferentiable problem. The traditional continuous relaxation schemes, such as the ℓ_1-norm regularization, often require computationally prohibitive iterative solvers, e.g., an alternating direction method of multipliers framework [36]. To overcome this limitation, SAGL formulates the construction of the sparse attention graph as a Tsallis entropy-regularized projection onto the probability simplex Δ^{n-1}. We provide the following theorem to establish the equivalence between the continuous simplex projection and the discrete ℓ_0-norm constraint.

Theorem A.5 (Differentiable Relaxation and Support Sparsity of the ℓ_0-Norm Constraint)**.** *Let $\tilde{\mathbf{s}}_i^{(l)} \in \mathbb{R}^n$ be the directed similarity vector for the ith feature. The structured sparse projection imposed on the probability simplex Δ^{n-1} via the α-entmax transformation acts as a continuous and differentiable relaxation of the discrete ℓ_0-norm constraint. Specifically, it guarantees a strictly sparse support set supp$(\mathbf{a}_i^{(l)}) = \{j \mid a_{ij}^{(l)} > 0\}$, where the discrete cardinality $\|\mathbf{a}_i^{(l)}\|_0$ is exactly determined by the continuous topological uncertainty factor $\tau_i^{(l)}$.*

Proof. The α-entmax transformation computes the attention distribution by solving a constrained variational optimization problem over the probability simplex Δ^{n-1}:

$$\mathbf{a}_i^{(l)} = \underset{\mathbf{a}_i^{(l)} \in \Delta^{n-1}}{\arg\max} \left(\left(\mathbf{a}_i^{(l)}\right)^\top \tilde{\mathbf{s}}_i^{(l)} + \mathcal{H}_\alpha^{\mathrm{T}}\left(\mathbf{a}_i^{(l)}\right) \right) \tag{42}$$

where the Tsallis α-entropy for $\alpha > 1$ is defined as

$$\mathcal{H}_\alpha^{\mathrm{T}}(\mathbf{a}_i^{(l)}) = \frac{1}{\alpha(\alpha-1)} \sum_{j=1}^{n} \left(a_{ij}^{(l)} - \left(a_{ij}^{(l)}\right)^\alpha \right). \tag{43}$$

To reveal its connection with the ℓ_0-norm, we formulate the Lagrangian of this optimization problem by incorporating the probability simplex constraints ($\mathbf{1}^\top \mathbf{a}_i^{(l)} = 1$ and $\mathbf{a}_i^{(l)} \geq \mathbf{0}$):

$$\mathcal{L}(\mathbf{a}_i^{(l)}, \rho, \boldsymbol{\mu}) = \sum_{j=1}^{n} a_{ij}^{(l)} \tilde{s}_{ij}^{(l)} + \frac{1}{\alpha(\alpha-1)} \sum_{j=1}^{n} \left(a_{ij}^{(l)} - \left(a_{ij}^{(l)}\right)^\alpha \right) - \rho \left(\sum_{j=1}^{n} a_{ij}^{(l)} - 1 \right) + \sum_{j=1}^{n} \mu_j a_{ij}^{(l)} \tag{44}$$

where $\rho \in \mathbb{R}$ and $\boldsymbol{\mu} \in \mathbb{R}_+^n$ are the Lagrange multipliers for the unit-sum and non-negativity constraints, respectively.

Table 6: Statistics of the datasets.

NAME	TRAIN	TEST	CLASSES
PETS	3,680	3,669	37
KITTI	5,985	1,496	4
FLOWERS	2,040	6,149	102
CALTECH101	3,060	6,084	102
EUROSAT	10,000	5,000	10
SUN397	19,850	19,850	397
FOOD101	75,750	25,250	101
IMAGENET-1K	1,281,167	50, 000	1,000

Setting the partial derivative with respect to $a_{ij}^{(l)}$ to zero gives the Karush-Kuhn-Tucker (KKT) stationarity condition:

$$\frac{\partial \mathcal{L}}{\partial a_{ij}^{(l)}} = \tilde{s}_{ij}^{(l)} + \frac{1 - \alpha \left(a_{ij}^{(l)}\right)^{\alpha-1}}{\alpha(\alpha-1)} - \rho + \mu_j = 0. \tag{45}$$

By the KKT complementary slackness condition, $\mu_j = 0$ whenever $a_{ij}^{(l)} > 0$. Upon solving for $a_{ij}^{(l)}$ under this active support condition, we obtain

$$a_{ij}^{(l)} = \left[(\alpha-1)\left(\tilde{s}_{ij}^{(l)} - \rho\right) + \frac{1}{\alpha}\right]^{\frac{1}{\alpha-1}}. \tag{46}$$

Letting $\rho^* = (\alpha-1)\rho - \frac{1}{\alpha}$ and applying the truncation operator $[x]_+ = \max(x, 0)$ to enforce nonnegativity, we have the exact closed-form solution:

$$a_{ij}^{(l)} = \left[(\alpha-1)\left(\tilde{s}_{ij}^{(l)}\right) - \rho^*\right]_+^{\frac{1}{\alpha-1}}. \tag{47}$$

By substituting the similarity $\tilde{s}_{ij}^{(l)} = s_{ij}^{(l)} \cdot \left(1 - \omega_i^{(l)}\right)$ in Eq. (9) into the above equation, we obtain

$$a_{ij}^{(l)} = \left[(\alpha-1)\left(s_{ij}^{(l)} \cdot \left(1 - \omega_i^{(l)}\right)\right) - \rho^*\right]_+^{\frac{1}{\alpha-1}}. \tag{48}$$

The truncation operator $[\cdot]_+$ sets any similarity that is below the threshold ρ^* to zero, which directly bounds the ℓ_0-norm such that $\|\mathbf{a}_i^{(l)}\|_0 \ll n$. Moreover, as the feature-specific compression factor $\omega_i^{(l)}$ increases toward 1, the variance of the similarities increases. This causes the threshold operator to prune more connections. Therefore, the number of nonzero entries in the sparse attention vector is adaptively controlled through the differentiable simplex projection. □

Theorem A.5 shows that the α-entmax projection onto the probability simplex Δ^{n-1} provides an exact differentiable relaxation of the discrete ℓ_0-norm constraint. This eliminates the need for computationally expensive iterative solvers, enabling efficient end-to-end optimization of sparse attention graphs.

A.6 Experimental Details

The statistics of the datasets are summarized in Table 6. All experiments are performed on a Linux workstation equipped with a GeForce RTX 4090 GPU (24 GB memory), an Intel Xeon Platinum 8336C CPU, and 128 GB of RAM. SAGL is implemented in PyTorch [30].

A.6.1 Parameter Settings

During both training and testing, the learning rate for the proposed SAGL model is empirically set to 5×10^{-4} for the KITTI, Flowers, Food101 and ImageNet-1K datasets and to 1×10^{-3} for all other datasets. The batch size for both training and testing is selected from the set

Table 7: CKA scores for various pretrained model pairs.

DINOv3	SigLIP 2	CLIP ViT-L/14	ConvNeXt V2	Pets	KITTI	Flowers	Caltech101	EuroSAT	SUN397	Food101	ImageNet-1K
✓	✓			0.63	0.84	0.65	0.62	0.70	0.49	0.67	0.46
✓		✓		0.59	0.81	0.61	0.57	0.70	0.48	0.64	0.41
✓			✓	0.06	0.19	0.15	0.07	0.16	0.07	0.04	0.06
	✓	✓		0.85	0.86	0.76	0.79	0.70	0.75	0.78	0.79
	✓		✓	0.05	0.18	0.25	0.09	0.18	0.08	0.07	0.09

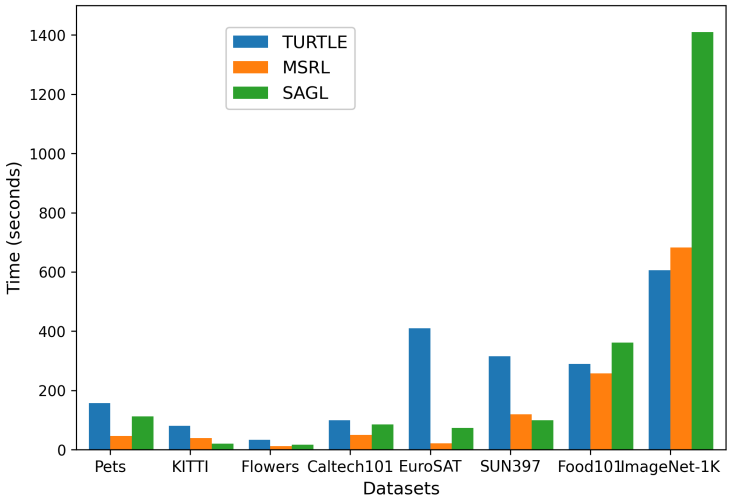


Figure 2: Training times (in seconds) of TURTLE, MSRL, and SAGL on eight vision datasets.

$\{100, 500, 1{,}000, 5{,}000, 10{,}000\}$. Specifically, the batch size is determined by scaling the number of clusters by an approximate multiplier $a \in \{5, 10, 20, 50, 100\}$. For datasets with a relatively small number of clusters, we apply smaller multipliers (e.g., 5, 10, or 20), whereas larger multipliers are used otherwise. The proposed SAGL model is trained for 600 epochs. The dropout rate is selected from the set $\{0.0, 0.1, 0.2, 0.3, 0.4\}$, with lower values applied to smaller datasets and higher values applied to larger datasets. We employ a grid search strategy to determine the optimal combination of hyperparameter values, selecting γ from the set $\{5.0, 10.0, 20.0\}$, and β from the set $\{0.2, 0.5, 1.0\}$. We report the best-tuned performance for all competing methods to ensure a fair comparison.

A.6.2 Measuring Similarity Across Heterogeneous Views

Different backbones exhibit different representation levels: transformer-based models (e.g., DINOv3, SigLIP 2 and CLIP ViT-L/14) typically produce global semantic representations, while convolutional models (e.g., ConvNeXt V2) capture more localized spatial features. We adopt centered kernel alignment (CKA) [37] to measure the similarity between feature distributions produced by different pretrained model pairs. CKA summarizes the similarity between two feature distributions as a scalar value in $[0, 1]$. A higher CKA score indicates that the feature distributions of a pair of sets tend to be identical. Table 7 shows the CKA scores for various pretrained model pairs. As expected, the CKA scores between transformer-based and convolutional architectures are relatively low, reflecting their fundamentally different feature distributions.

A.6.3 Comparison of Training Times for Self-Supervised Learning

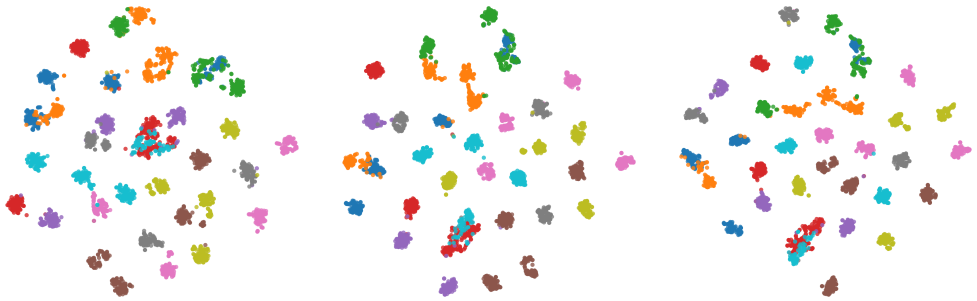
To evaluate the training efficiency of the proposed SAGL method, we compare the computational costs of TURTLE, MSRL, and SAGL on the training sets of all eight datasets. For fair comparison, we report the computational cost of the competing methods that utilize two pretrained backbones. Fig. 2 shows the training times of TURTLE, MSRL, and SAGL on eight datasets. MSRL runs faster than TURTLE and SAGL on most datasets. This is because it directly employs standard dense attention without considering intrinsic subspace structures embedded in the high-dimensional multiview data. As observed, SAGL achieves higher training efficiency than TURTLE on small and medium-scale datasets while consistently maintaining superior clustering performance. However, SAGL requires more training time on large-scale datasets such as ImageNet-1K. This is due to the

structured sparse projection based on the α-entmax transformation. This represents a reasonable computational trade-off for the significant performance gains achieved.



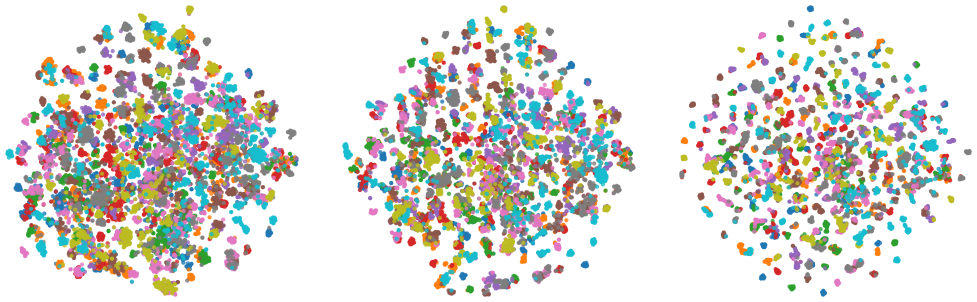
(a) The original features (b) The linear features (c) The corresponding representations

Figure 3: The t-SNE visualization of three levels of features generated by SAGL on the Pets dataset, where the original features are extracted using SigLIP 2.



(a) The original features (b) The linear features (c) The corresponding representations

Figure 4: The t-SNE visualization of three levels of features generated by SAGL on the Pets dataset, where the original features are extracted using DINOv3.

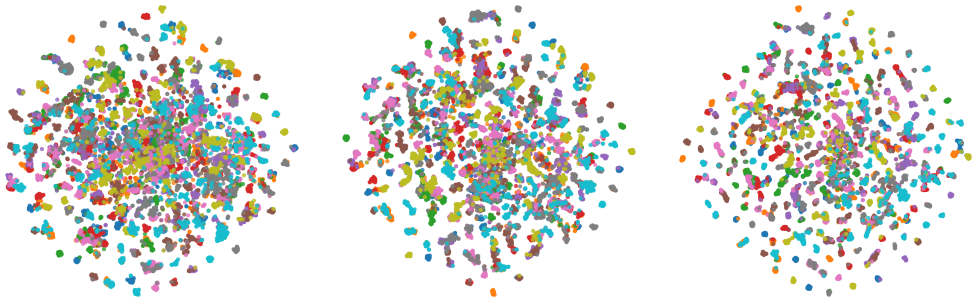


(a) The original features (b) The linear features (c) The corresponding representations

Figure 5: The t-SNE visualization of three levels of features generated by SAGL on the SUN397 dataset, where the original features are extracted using SigLIP 2.

A.6.4 Visualizations

To evaluate the learned representations, we employ t-SNE [38] to visualize three levels of features on three representative datasets of varying scales: Pets, SUN397, and Food101. Specifically, the three levels are: (1) the original features extracted from the two pretrained backbones, (2) the projected



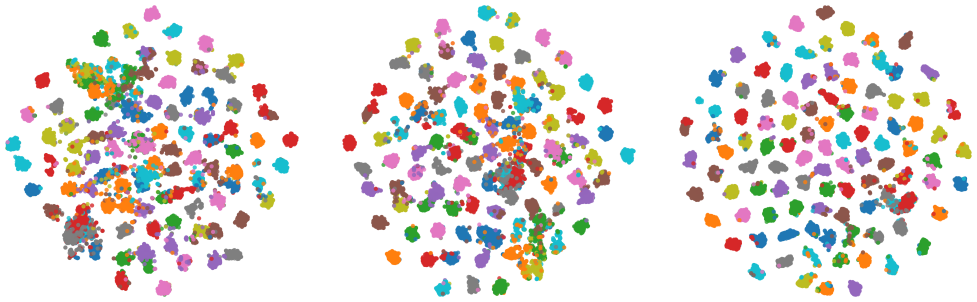
(a) The original features (b) The linear features (c) The reconstructed representations

Figure 6: The t-SNE visualization of three levels of features generated by SAGL on the SUN397 dataset, where the original features are extracted using DINOv3.



(a) The original features (b) The linear features (c) The reconstructed representations

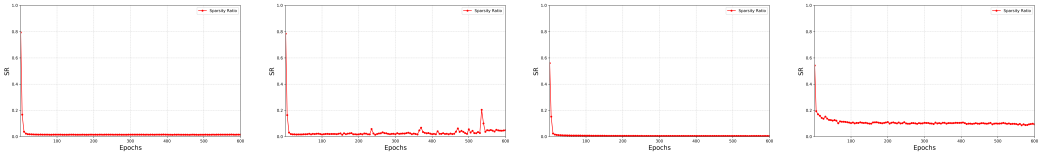
Figure 7: The t-SNE visualization of three levels of features generated by SAGL on the Food101 dataset, where the original features are extracted using SigLIP 2.



(a) The original features (b) The linear features (c) The reconstructed representations

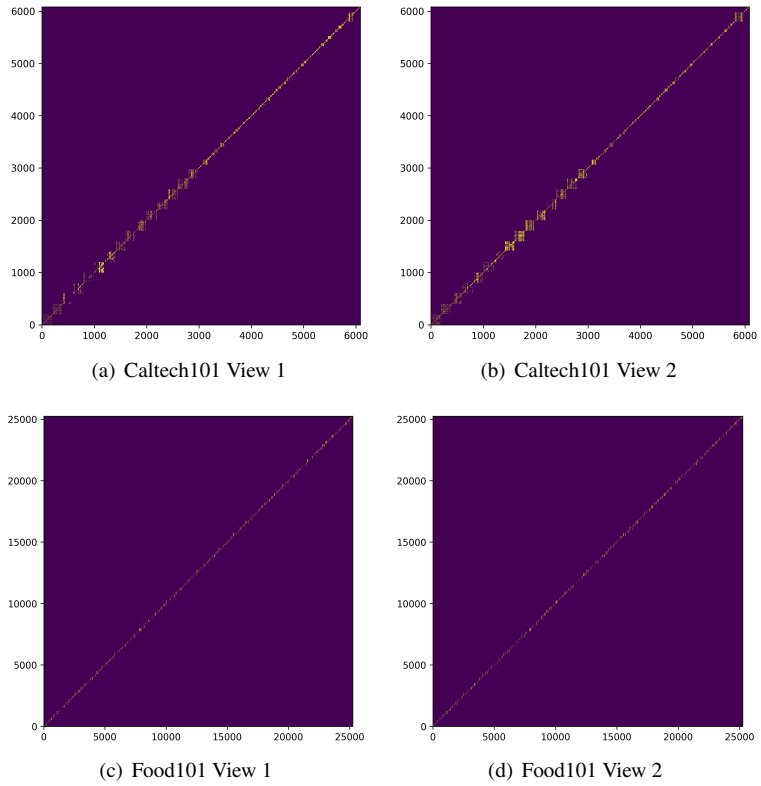
Figure 8: The t-SNE visualization of three levels of features generated by SAGL on the Food101 dataset, where the original features are extracted using DINOv3.

features after the linear transformation layers, and (3) the final representations constructed by SAGL via sparse information aggregation. Figs. 3-8 show the t-SNE visualizations across these three levels. We can observe that the learned representations exhibit significantly better class separability than those in the earlier stages. These results indicate that SAGL produces more discriminative representations for improved clustering performance.



(a) Caltech101 View 1 (b) Caltech101 View 2 (c) Food101 View 1 (d) Food101 View 2

Figure 9: Sparsity ratio evolution of sparse attention graphs during training.



(a) Caltech101 View 1 (b) Caltech101 View 2

(c) Food101 View 1 (d) Food101 View 2

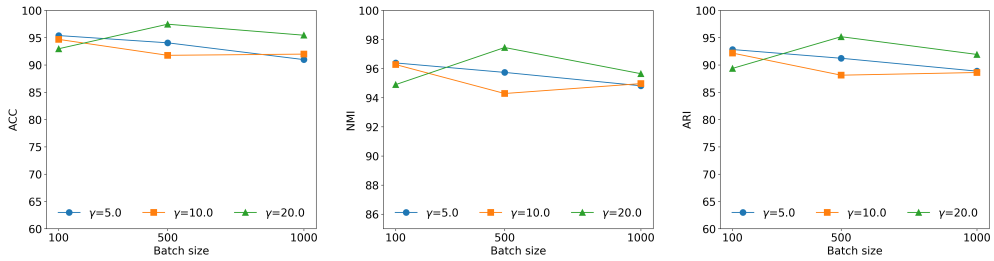
Figure 10: Block-diagonal structures of the learned sparse attention graphs on the Caltech101 and Food101 datasets.

A.6.5 Sparsity Analysis on Sparse Attention Graphs

We first analyze the sparsity of the learned attention graphs during training. The sparsity ratio (SR) is defined as the number of nonzero elements in $\mathbf{A}^{(l)}$ divided by the total number of elements. Specifically, we examine the sparsity ratios of sparse attention graphs on the two representative datasets, Caltech101 and Food101, during training. Fig. 9 illustrates the evolution of sparsity ratios of the learned sparse attention graphs during training, where View 1 and View 2 correspond to features extracted from SigLIP 2 and DINOv3, respectively. We can observe that the sparsity ratio drops rapidly in the first few epochs and fluctuates slightly during the remaining epochs.

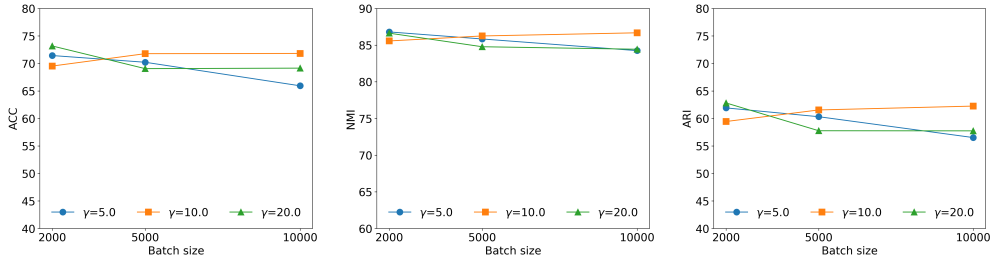
To investigate the subspace-preserving property of the proposed SAGL model, we further examine subspace structures of the learned sparse attention graphs during testing. Fig. 10 illustrates that the

learned sparse attention graphs exhibit clear block-diagonal structures, where each diagonal block corresponds to a distinct latent semantic category. This block-diagonal pattern directly reflects the intrinsic subspace structures embedded in the high-dimensional multiview data. These results provide intuitive evidence that SAGL faithfully recovers the underlying subspace structures.



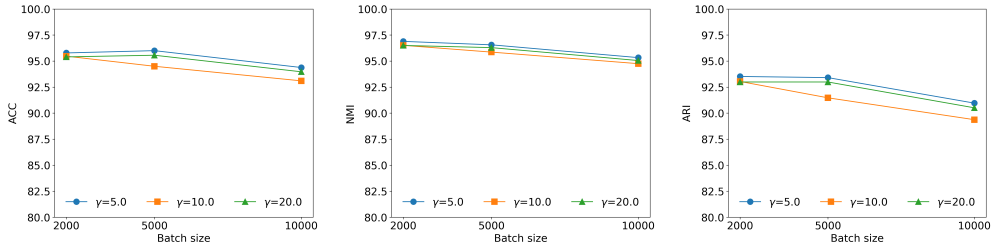
(a) ACC (b) NMI (c) ARI

Figure 11: Clustering results of SAGL on the Pets dataset across different batch sizes.



(a) ACC (b) NMI (c) ARI

Figure 12: Clustering results of SAGL on the SUN397 dataset across different batch sizes.



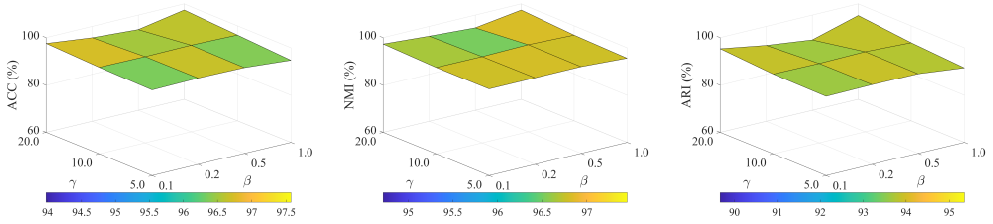
(a) ACC (b) NMI (c) ARI

Figure 13: Clustering results of SAGL on the Food101 dataset across different batch sizes.

A.6.6 Parameter Sensitivity Analysis

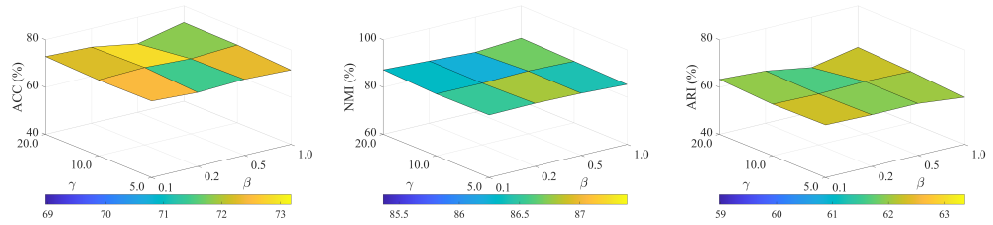
By exploiting the sparse self-representation property of features, each representation is constructed as a sparse linear combination of spatially proximate neighbors. Consequently, the batch size plays an important role in determining the quality of the learned sparse attention graphs. To investigate the sensitivity of SAGL to batch size, we evaluate the clustering performance on three representative datasets of varying scales: Pets, SUN397, and Food101. Following the batch size selection strategy provided in Section A.6.1, the candidate batch sizes are set to $\{100, 500, 1{,}000\}$ for Pets and $\{2{,}000, 5{,}000, 10{,}000\}$ for SUN397 and Food101. The parameter γ is selected from $\{5.0, 10.0, 20.0\}$, and β is fixed at 1.0. Figs. 11-13 show the clustering performance of SAGL under different batch sizes on the Pets, SUN397, and Food101 datasets, respectively. The clustering performance of SAGL remains relatively stable across different batch sizes, particularly on large-scale datasets such as

SUN397 and Food101. This indicates that SAGL yields competitive clustering results across a wide range of batch sizes.



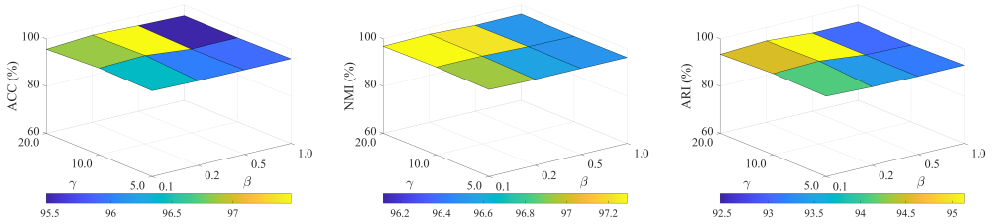
(a) ACC (b) NMI (c) ARI

Figure 14: The clustering results of SAGL under different γ and β combinations on the Pets dataset.



(a) ACC (b) NMI (c) ARI

Figure 15: The clustering results of SAGL under different γ and β combinations on the SUN397 dataset.



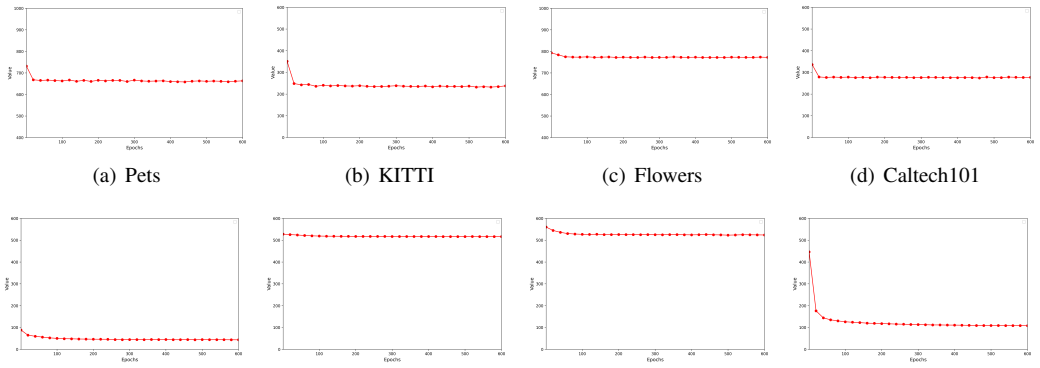
(a) ACC (b) NMI (c) ARI

Figure 16: The clustering results of SAGL under different γ and β combinations on the Food101 dataset.

We further evaluate the sensitivity of the proposed SAGL model to hyperparameters γ and β. Specifically, we investigate the clustering performance of SAGL under different combinations of hyperparameters $\gamma \in \{5.0, 10.0, 20.0\}$ and $\beta \in \{0.1, 0.2, 0.5, 1.0\}$. Figs. 14-16 illustrate the clustering performance of SAGL in terms of ACC, NMI, and ARI on the Pets, SUN397, and Food101 datasets. The proposed SAGL method performs steadily with relatively wide ranges of the parameters γ and β. According to the sensitivity experiments, we empirically observe that setting $\gamma \in \{5.0, 10.0\}$ with $\beta = 1.0$ often yields satisfactory clustering results for SAGL.

A.6.7 Convergence Analysis

We empirically evaluate the convergence property of the proposed method across all eight datasets. Fig. 17 shows the convergence curves of Algorithm 1, where the x-axis corresponds to iterations, and the y-axis represents the objective loss defined in Eq. (22). A positive constant is added to the y-axis values for better readability. As observed in the figures, the objective value drops dramatically



(a) Pets (b) KITTI (c) Flowers (d) Caltech101

(e) EuroSAT (f) SUN397 (g) Food101 (h) ImageNet-1K

Figure 17: Convergence results of Algorithm 1 on all the datasets.

within the first few dozen iterations and quickly converges to a steady state. These results empirically validate the rapid and stable convergence properties of the proposed method in practice.