

GROVER: Graph-guided Representation of Omics and Vision with Expert Regulation for Adaptive Spatial Multi-omics Fusion

Yongjun Xiao^{1,2*}, Dian Meng^{3*},

Xinlei Huang¹, Yanran Liu¹, Shiwei Ruan¹, Ziyue Qiao¹, Xubin Zheng^{1†}

¹Dongguan Key Laboratory for Intelligence and Information Technology, School of Computing and Information Technology, Great Bay University, Guangdong, China

²College of Information Engineering, Sichuan Agricultural University, Ya'an, China

³Department of Biomedical Informatics, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
xbzheng@gbu.edu.cn

Abstract

Effectively modeling multimodal spatial omics data is critical for understanding tissue complexity and underlying biological mechanisms. While spatial transcriptomics, proteomics, and epigenomics capture molecular features, they lack pathological morphological context. Integrating these omics with histopathological images is thus critical for comprehensive disease tissue analysis. However, substantial heterogeneity across omics, imaging, and spatial modalities poses significant challenges. Naive fusion of semantically distinct sources often leads to ambiguous representations. Additionally, the resolution mismatch between high-resolution histology images and lower-resolution sequencing spots complicates spatial alignment. Biological perturbations during sample preparation further distort modality-specific signals, hindering accurate integration. To address these challenges, we propose **Graph-guided Representation of Omics and Vision with Expert Regulation for Adaptive Spatial Multi-omics Fusion (GROVER)**, a novel framework for adaptive integration of spatial multi-omics data. GROVER leverages a Graph Convolutional Network encoder based on Kolmogorov–Arnold Networks to capture the nonlinear dependencies between each modality and its associated spatial structure, thereby producing expressive, modality-specific embeddings. To align these representations, we introduce a spot-feature-pair contrastive learning strategy that explicitly optimizes the correspondence across modalities at each spot. Furthermore, we design a dynamic expert routing mechanism that adaptively selects informative modalities for each spot while suppressing noisy or low-quality inputs. Experiments on real-world spatial omics datasets demonstrate that GROVER outperforms state-of-the-art baselines, providing a robust and reliable solution for multimodal integration.

Code — <https://github.com/Xubin-s-Lab/GROVER>

Introduction

Spatially resolved transcriptomics and spatial proteomics were recognized as the Methods of the Year by Nature

*These authors contributed equally.

†Corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

in 2021 (Marx 2021) and 2024 (Karimi et al. 2024), respectively. These advancements extend single-cell analyses of gene expression and surface protein abundance into the spatial domain, offering unprecedented insight into tissue organization. More recently, spatially resolved multimodal omics—encompassing transcriptomics, proteomics, and high-resolution histological imaging—have emerged as a powerful paradigm for integrative analysis, enabling a comprehensive understanding of gene regulation and the tissue microenvironment within their native spatial context (Coleman et al. 2025). The main challenge in multimodal spatial omics analysis is the effective integration of features from diverse modalities to generate coherent low-dimensional representations that facilitate downstream tasks, such as spatial domain identification via clustering. Most recent approaches primarily focus on integrating transcriptomic and proteomic data (Meng et al. 2024), while overlooking the valuable structural context provided by histological images. For instance, SpatialGlue (Long et al. 2024) employs cross-modal attention to fuse transcriptomic and proteomic features, and PRAGA (Huang et al. 2025) integrates spatial multi-omics data with adaptive graph structures and dynamic prototype contrastive learning. MISO (Coleman et al. 2025) represents a recent advancement by incorporating histology images into a multimodal integration pipeline through outer-product interactions. Moreover, these methods tend to treat all modalities equally across spatial locations, ignoring substantial variations in data quality. In practice, spatial omics data are often compromised by noise from both technical limitations (e.g., dropout events in single-cell sequencing) (Ge et al. 2025) and biological or experimental artifacts (e.g., tissue sectioning errors in histology) (Totty, Hicks, and Guo 2025). Such heterogeneity makes it critical to assess the reliability of each modality per spot. Yet, current integration frameworks are unable to adaptively weigh or filter unreliable features based on local signal quality, which may limit the accuracy and robustness of spatial analysis. To address this, we propose a multi-expert learning framework that performs modality-aware integration at the single-spot level, selectively emphasizing high-confidence signals and mitigating the impact of noise. Fur-

thermore, due to the substantial semantic gap between omics data and histological images, as well as the many-to-many mapping between image patches and spatial spots, achieving accurate cross-modal alignment remains highly challenging. To overcome this, we introduce a contrastive learning strategy that encourages consistency between spatial and morphological representations in a weakly paired setting.

In this paper, we propose a novel spatially resolved multi-omics framework, **Graph-guided Representation of Omics and Vision with Expert Regulation for Adaptive Spatial Multi-omics Fusion (GROVER)**. GROVER employs a spot-feature-pair based contrastive learning approach to integrate features from three modalities, guided by graph-based spatial relationships and structural information derived from histological images. Inspired by the Mixture-of-Experts (MoE) paradigm, we design a multi-expert model to adaptively filter and fuse modality-specific features at the spot level. Extensive qualitative and quantitative results demonstrate that GROVER excels at aggregating spatial multi-omics information into spot-type-resolvable representations by effectively integrating histological image features and performing adaptive feature fusion through a multi-expert framework. Our contributions are summarized as follows:

- We propose a novel spatially resolved multi-omics framework, **GROVER**, which adaptively integrates transcriptomic, proteomic, and histological modalities at the single-spot level. We introduce a spot-feature-pair based contrastive learning strategy to bridge the semantic gap between omics data and histological images, enabling accurate cross-modal alignment.
- We design a modality-specific multi-expert architecture with gated routing to adaptively weigh heterogeneous signals, enhancing both robustness and interpretability.
- We develop a hybrid graph encoder, the Graph Convolutional Network based on Kolmogorov–Arnold Networks (KAN-GCN), which enhances message passing with kernel-based nonlinear transformations for expressive, structure-aware representation learning.

Related Works

Spatial Multi-omics Integration

Recently, spatial multi-omics technologies have emerged as powerful tools for linking spatial context with molecular profiling, offering new opportunities to dissect cellular heterogeneity within the tissue microenvironment. SpatialGlue (Long et al. 2024) integrates spatial and omics information using a dual-attention graph neural network that captures cross-modal correspondence and spatial structure. COSMOS (Zhou et al. 2025) combines graph convolutional networks (GCNs) with a weighted nearest neighbor (WNN) framework, and further employs Deep Graph Infomax (DGI) and spatial regularization to fuse complementary features for downstream analysis. However, these approaches primarily focus on integrating molecular modalities and often neglect histological context. Due to the semantic disparity and alignment difficulty between tissue images and omics measurements, image information is rarely incorporated into the integration process.

Integration of Image and Omics Data

Histological images offer crucial morphological context that complements molecular data, making them a valuable addition in spatial omics (Chelebian, Avenel, and Wählby 2025). Their integration enhances both biological insight and analytical depth. MISO encodes each modality independently using Multilayer Perceptron (MLP), models interactions via outer products, and concatenates results into a unified embedding. However, it treats all modalities equally across spatial spots, ignoring differences in quality, noise, or relevance. To address this, we propose a multi-expert framework inspired by mixture-of-experts (Jacobs et al. 1991). It assigns adaptive weights to modality-specific experts at each spot, allowing the model to prioritize informative signals while suppressing noisy or irrelevant inputs, leading to more robust and precise integration.

Method

Preliminaries

Spatial multi-modal omics integration aims to jointly model molecular measurements along with their spatial context, thereby obtaining unified spot-level representations suitable for downstream analyses. Formally, given a spatial dataset comprising N sequencing spots, we denote their spatial coordinates as:

$$S = \{(x_i, y_i)\}_{i=1}^N, \quad (x_i, y_i) \in \mathbb{R}^2. \quad (1)$$

Each spot is characterized by measurements from M different modalities, typically including transcriptomic, proteomic, and histological imaging modalities. For each modality $m \in \{1, \dots, M\}$, the features associated with the spots are defined as:

$$F^{(m)} = \{f_i^{(m)} \in \mathbb{R}^{D_m}\}_{i=1}^N, \quad (2)$$

where $f_i^{(m)}$ is the preprocessed D_m -dimensional feature vector corresponding to spot i .

For example, $F^{(m)}$ could represent gene expression features reduced by principal component analysis (PCA), antibody-derived tag (ADT) protein abundances, or visual features extracted from histology images using a pre-trained pathology foundation model such as OmiCLIP (Chen et al. 2025). Notably, our framework is highly modular and flexible, enabling the seamless incorporation of any state-of-the-art (SOTA) pathology foundation model, thus ensuring adaptability to future technological advances.

The goal of spatial multi-modal integration is to use a function Φ that adaptively fuses all modality features, informed by spatial topology, into a unified low-dimensional representation Z :

$$Z = \Phi(F^{(1)}, \dots, F^{(M)}, S). \quad (3)$$

The integrated representation Z is expected to preserve complementary biological signals across modalities while maintaining spatial coherence. This embedding can be used for various downstream tasks, such as spatial domain identification, cell type annotation, and tumor microenvironment characterization.

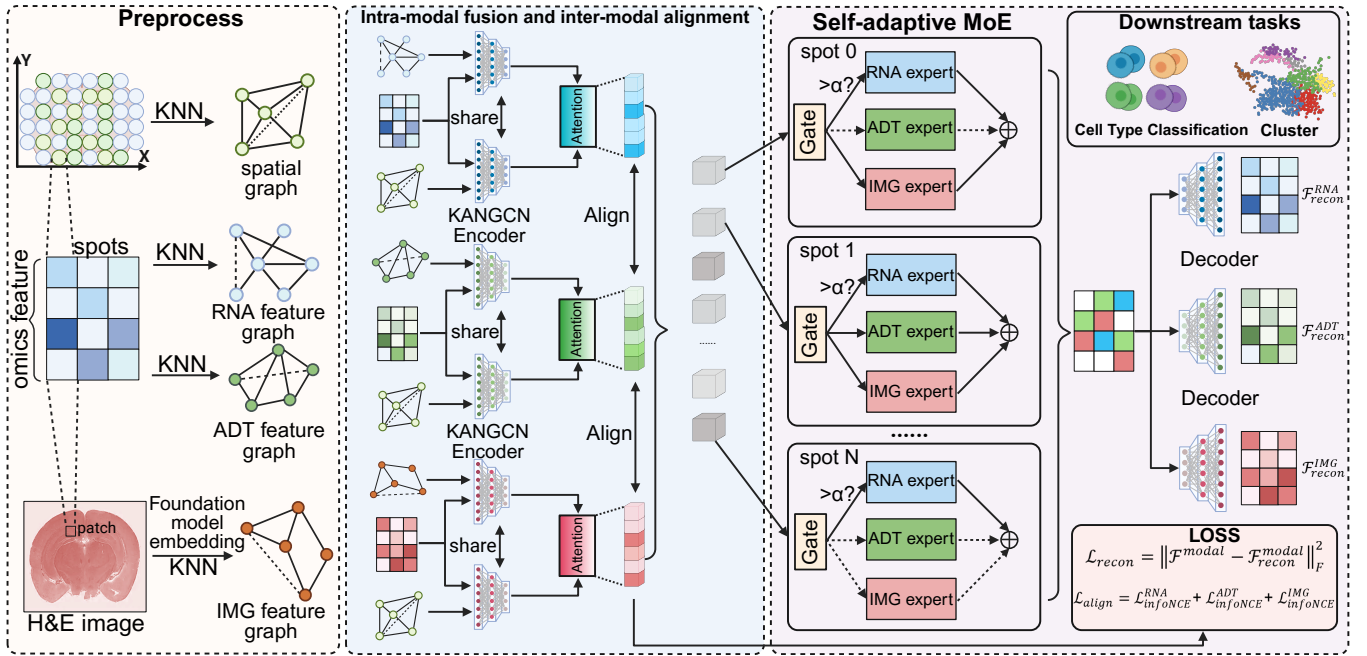


Figure 1: The framework of the proposed GROVER. GROVER encodes modality-specific feature graphs and spatial adjacency graphs using KAN-GCN, then applies attention-based weighted fusion to obtain integrated multimodal representations (RNA, protein, and image). A spot-feature-pair based contrastive learning aligns semantic information across modalities before feeding the embeddings into a self-adaptive Mixture-of-Experts model for fusion. The entire model is trained with modality-specific reconstruction losses and the spot-feature-pair contrastive loss.

GROVER

In this paper, we propose GROVER, a graph-guided architecture that models the function Φ in Eq. (1), adaptively integrating transcriptomic, proteomic, and histological modalities at single-spot resolution into a unified representation.

Given M modalities $\{F^{(m)}\}_{m=1}^M$, including omics features (e.g., RNA and protein) and histological image features, GROVER constructs two graphs per modality: a spatial graph $\mathcal{G}_S = (S, A_S)$ based on spot coordinates $S = \{(x_i, y_i)\}_{i=1}^N$, and a modality-specific feature graph $\mathcal{G}_F^{(m)} = (F^{(m)}, A_F^{(m)})$. Both A_S and $A_F^{(m)}$ are built via K-Nearest Neighbor (KNN), where A_S reflects physical proximity and $A_F^{(m)}$ captures feature similarity. These dual graphs jointly encode spatial and semantic structure.

For each modality, we fuse embeddings from the two graphs using a spatial-feature attention module:

$$\tilde{e}_i^{(m)} = \text{Attention}(e_i^S, e_i^{F, (m)}), \quad (4)$$

where e_i^S and $e_i^{F, (m)}$ are obtained via KAN-GCN on \mathcal{G}_S and $\mathcal{G}_F^{(m)}$, respectively.

To reduce modality discrepancy, we apply spot-level contrastive learning to align the fused embeddings $\tilde{e}_i^{(m)}$. To address varying modality quality, an adaptive multi-expert model with dynamic gating is introduced. The final fused representation Z is obtained via this gating mechanism. An overview of GROVER is shown in Figure 1, with the full pipeline summarized in Algorithm 1.

Encoder based on the KAN-GCN

To extract meaningful node representations from both spatial and modality-specific feature graphs, we adopt a graph convolutional encoder framework. Traditional Graph Convolutional Networks (GCNs) are widely used for learning node embeddings via message passing over graph structures (Kipf 2016).

Formally, given a graph $\mathcal{G} = (V, A)$ with N nodes, where V denotes the set of spatial spots and $A \in 0, 1^{N \times N}$ is the adjacency matrix representing the spatial or functional connectivity between spots, the graph is characterized by its node feature matrix $X \in \mathbb{R}^{N \times d_0}$, where d_0 is the input feature dimension. In our context, each node corresponds to a spatial spot, and its associated feature vector may include modality-specific measurements such as gene expression, protein abundance, or chromatin accessibility.

The layer-wise propagation rule of a standard GCN is defined as:

$$H^{(l+1)} = \sigma \left(\tilde{D}^{-1/2} \tilde{A} \tilde{D}^{-1/2} H^{(l)} W^{(l)} \right), \quad (5)$$

where $H^{(0)} = X$, $W^{(l)} \in \mathbb{R}^{d_l \times d_{l+1}}$ is the learnable weight matrix at layer l , and $\tilde{A} = A + I$ is the adjacency matrix with added self-loops, with \tilde{D} being the corresponding degree matrix. Here, $l = 0, 1, \dots, L - 1$ denotes the layer index, and L is the total number of encoder layers.

To enhance the expressive power of GCNs, we replace the fixed linear transformation with a nonlinear function modeled by a Kolmogorov–Arnold Network (KAN) (Liu et al.

2024). Each KAN layer consists of a matrix of trainable univariate functions $\varphi_{q,p}^{(l)} : [0, 1] \rightarrow \mathbb{R}$, where $p = 1, \dots, d_l$ and $q = 1, \dots, d_{l+1}$. The transformed features are computed as:

$$\mathcal{F}^{(l)}(H^{(l)})_{i,q} = \sum_{p=1}^{d_l} \varphi_{q,p}^{(l)}(H_{i,p}^{(l)}), \quad (6)$$

and the node update becomes:

$$H^{(l+1)} = \sigma(\hat{A} \cdot \mathcal{F}^{(l)}(H^{(l)})). \quad (7)$$

In GROVER, we apply a multi-layer KAN-GCN encoder to the spatial graph \mathcal{G}_S and each modality-specific graph $\mathcal{G}_F^{(m)}$, and denote the final-layer outputs as $H_S^{(L)} \in \mathbb{R}^{N \times d_L}$ and $H_{F^{(m)}}^{(L)} \in \mathbb{R}^{N \times d_L}$, respectively. The node embeddings for spot i are:

$$e_i^S = H_{S,i}^{(L)}, \quad e_i^{F^{(m)}} = H_{F^{(m)},i}^{(L)}. \quad (8)$$

To adaptively fuse these two types of embeddings, we use a within-modality attention aggregation layer. Let $\mathbf{W} \in \mathbb{R}^{d_L \times d_{\text{att}}}$ and $\mathbf{b} \in \mathbb{R}^{d_{\text{att}}}$ be the shared linear transformation parameters, and let $\mathbf{q} \in \mathbb{R}^{d_{\text{att}}}$ be a learnable attention vector. Here, d_L is the dimensionality of the encoder output, and d_{att} is a hidden projection dimension used specifically for computing attention scores.

The scalar compatibility score for each embedding source $t \in \{S, F\}$ is computed as:

$$e_i^{(t)} = \mathbf{q}^\top \tanh(\mathbf{W}e_i^{(t)} + \mathbf{b}). \quad (9)$$

We then apply a softmax to normalize the scores into attention weights:

$$\alpha_i^{(t)} = \frac{\exp(e_i^{(t)})}{\exp(e_i^{(S)}) + \exp(e_i^{(F)})}. \quad (10)$$

Finally, the fused modality-aware embedding is obtained by weighted summation:

$$\tilde{e}_i^{(m)} = \alpha_i^{(S)} e_i^S + \alpha_i^{(F)} e_i^{F^{(m)}}, \quad (11)$$

where $\tilde{e}_i^{(m)} \in \mathbb{R}^{d_L}$ serves as the unified representation for downstream tasks.

Spot-Feature-Pair Based Contrastive Learning

The transcriptomic (RNA), proteomic (ADT), and histological (image) modalities differ significantly in terms of data distributions and biological semantic characteristics. To address this challenge, we introduce a dual-alignment strategy that aligns both spatial and semantic information prior to modality fusion. For spatial alignment, we extract image patches centered at each spot's absolute coordinates on the tissue section and encode them using a pretrained vision foundation model to obtain location-aware image embeddings. For semantic alignment, we implement a bi-directional masked contrastive learning framework that enforces semantic consistency across pairs of modalities while mitigating the impact of spurious negatives caused by biologically similar spots.

Specifically, for each modality $m \in \{\text{RNA, ADT, Image}\}$, we obtain fused embeddings $\tilde{E}^{(m)} = \{\tilde{e}_i^{(m)}\}_{i=1}^N$. We first compute a cosine similarity matrix:

$$S_{i,j}^{(m)} = \text{sim}(\tilde{e}_i^{(m)}, \tilde{e}_j^{(m)}), \quad (12)$$

and define a binary similarity mask $M^{(m)} \in \{0, 1\}^{N \times N}$ as:

$$M_{i,j}^{(m)} = \begin{cases} 0, & \text{if } S_{i,j}^{(m)} > \delta \text{ and } i \neq j, \\ 1, & \text{otherwise,} \end{cases} \quad (13)$$

where δ is a threshold controlling the exclusion of highly similar negatives. This mask removes semantically similar spots from the contrastive denominator to mitigate false negatives during training.

Using the mask, we define the masked InfoNCE loss for a modality pair (m_1, m_2) as:

$$\ell_{\text{masked}}(\tilde{E}^{(m_1)}, \tilde{E}^{(m_2)}, M^{(m_1)}) = -\frac{1}{N} \sum_{i=1}^N \log s_i, \quad (14)$$

$$s_i = \frac{\exp(\text{sim}(\tilde{e}_i^{(m_1)}, \tilde{e}_i^{(m_2)})/\tau)}{\sum_{j=1}^N M_{i,j}^{(m_1)} \exp(\text{sim}(\tilde{e}_i^{(m_1)}, \tilde{e}_j^{(m_2)})/\tau)}, \quad (15)$$

where τ is a temperature parameter and $\text{sim}(\cdot, \cdot)$ denotes the cosine similarity between normalized vectors.

To ensure symmetric alignment between modalities, we compute the masked InfoNCE loss (Oord, Li, and Vinyals 2018) in both directions and average them:

$$\mathcal{L}_{\text{contrast}}^{m_1, m_2} = \frac{1}{2} \left(\ell_{\text{masked}}(\tilde{E}^{(m_1)}, \tilde{E}^{(m_2)}, M^{(m_1)}) + \ell_{\text{masked}}(\tilde{E}^{(m_2)}, \tilde{E}^{(m_1)}, M^{(m_2)}) \right). \quad (16)$$

We apply contrastive loss to three modality pairs (RNA-ADT, RNA-Image, ADT-Image). The aligned embeddings are denoted $\hat{e}_i^{(m)}$.

Self-adaptive Mixture of Experts

In practice, due to quality differences among modalities, naively integrating all modalities at the single-spot resolution can be suboptimal. To address this, we propose a self-adaptive Mixture of Experts (MoE) framework.

Given the aligned embeddings $\hat{e}_i^{(R)}$, $\hat{e}_i^{(A)}$, and $\hat{e}_i^{(I)}$ from RNA, protein, and image modalities respectively, we first compute an aggregated gating input by averaging:

$$x_i = \frac{1}{3} \left(\hat{e}_i^{(R)} + \hat{e}_i^{(A)} + \hat{e}_i^{(I)} \right). \quad (17)$$

A gating network parameterized by $W_{\text{gate}} \in \mathbb{R}^{D \times 3}$ produces raw confidence scores $\mathbf{g}_i = W_{\text{gate}} x_i \in \mathbb{R}^3$, which are converted via softmax to normalized weights:

$$\beta_i^{(m)} = \frac{\exp(g_i^{(m)})}{\sum_{m'} \exp(g_i^{(m')})}, \quad m \in \{R, A, I\}. \quad (18)$$

To filter out unreliable modalities, we apply a threshold γ to the gating scores. In our experiments, we set $\gamma = 0.3$.

$$\tilde{\beta}_i^{(m)} = \begin{cases} \beta_i^{(m)}, & \text{if } \beta_i^{(m)} \geq \gamma, \\ 0, & \text{otherwise} \end{cases}. \quad (19)$$

These filtered weights are renormalized to sum to one:

$$s_i^{(m)} = \frac{\tilde{\beta}_i^{(m)}}{\sum_{m'} \tilde{\beta}_i^{(m')} + \epsilon}, \quad (20)$$

where $\epsilon = 10^{-6}$ is a small constant for numerical stability.

Each modality has a dedicated expert implemented as a feed-forward network:

$$h_i^{(m)} = \text{FFN}^{(m)}(\hat{e}_i^{(m)}). \quad (21)$$

The final fused spot-level representation is a weighted sum over experts:

$$z_i = \sum_m s_i^{(m)} \cdot h_i^{(m)}. \quad (22)$$

In rare cases where all $\beta_i^{(m)} < \gamma$, the fused representation defaults to the expert output of the modality with the highest original confidence:

$$z_i = h_i^{(m^*)}, \quad m^* = \arg \max_m \beta_i^{(m)}. \quad (23)$$

Finally, we collect the fused representations z_i of all spots into a set, denoted as

$$Z = \{z_1, z_2, \dots, z_N\}, \quad (24)$$

which forms a unified representation under the spatial structure S , corresponding to the output of Equation (3).

This unified representation Z comprehensively captures the integrated features of multimodal spatial omics data, facilitating subsequent downstream analysis tasks.

To reconstruct modality-specific features from the fused embeddings, we employ a graph-based decoder that utilizes the spatial adjacency structure \hat{A} . Specifically, for each modality $m \in \{R, A, I\}$, we define the reconstructed features as:

$$\hat{F}^{(m)} = \sigma(\hat{A} \cdot \mathcal{F}^{(m)}(Z)), \quad (25)$$

This formulation ensures that reconstruction leverages both the fused representation and the spatial context encoded in \hat{A} .

The reconstruction loss for each modality is computed as:

$$\mathcal{L}_{\text{rec}}^{(m)} = \frac{1}{N} \sum_{i=1}^N \|f_i^{(m)} - \hat{f}_i^{(m)}\|_2^2, \quad (26)$$

where $\hat{f}_i^{(m)}$ is the reconstructed feature for spot i in modality m , obtained from $\hat{F}^{(m)}$.

The overall training objective is formulated as:

$$\mathcal{L}_{\text{total}} = \sum_{m \in \{R, A, I\}} \mathcal{L}_{\text{rec}}^{(m)} + \lambda \sum_{m_i \neq m_j} \mathcal{L}_{\text{contrast}}^{m_i, m_j}, \quad (27)$$

where λ is a hyperparameter balancing the contribution of the contrastive alignment loss. In our experiments, we set $\lambda = 2$.

Algorithm 1: GROVER

Input: Multi-modal features $\{F^{(m)}\}_{m=1}^M$, spatial coordinates $S = \{(x_i, y_i)\}_{i=1}^N$

Parameters: Epochs E , temperature τ , contrastive loss weight λ , confidence threshold γ

- 1: Construct graphs: \mathcal{G}_S from S and $\mathcal{G}_F^{(m)}$ from $F^{(m)}$ via KNN
- 2: **for** $e = 1$ to E **do**
- 3: **for** each modality m **do**
- 4: Extract e_i^S and $e_i^{F^{(m)}}$ via KAN-GCN (Eq. 8)
- 5: Fuse to obtain $\tilde{e}_i^{(m)}$ via attention (Eq. 11)
- 6: **end for**
- 7: Contrastive loss: $\mathcal{L}_{\text{contrast}}$ across $\{\tilde{e}_i^{(m)}\}$ (Eq. 16)
- 8: MoE routing to get Z (Eq. 22, 24)
- 9: Reconstruction loss: $\mathcal{L}_{\text{recon}}$ from Z (Eq. 26)
- 10: Update model via total loss $\mathcal{L}_{\text{total}}$ (Eq. 27)
- 11: **end for**

Output: Unified comprehensive embeddings Z

The resulting embeddings Z serve as comprehensive representations for downstream analysis tasks, such as clustering and spatial domain detection.

Experiments

Experimental Setups

Dataset. We conduct quantitative and qualitative experiments on four public datasets to verify the effectiveness of the proposed method: 1) 10x Visium human breast Cancer gene and protein expression dataset (10x Genomics 2023a); 2) 10x Visium human glioblastoma gene and protein expression dataset (10x Genomics 2023b); 3) 10x Visium human tonsil gene and protein expression dataset (10x Genomics 2023c); 4) 10x Visium human tonsil with add-on antibodies gene and protein expression dataset (10x Genomics 2023d); These datasets are detailed in the Appendix¹.

Baselines. We compare our approach with three recent state-of-the-art multimodal omics methods: MISO, Spatial-Glue, and COSMOS. Notably, MISO supports image modality integration.

Metrics. We selected nine diverse metrics to comprehensively evaluate the model performance, including Adjusted Rand Index (ARI) (Steinley 2004), Normalized Mutual Information (NMI) (Vinh, Epps, and Bailey 2009), Fowlkes–Mallows Index (FMI) (Fowlkes and Mallows 1983), Silhouette Coefficient (SC) (Rousseeuw 1987), Adjusted Mutual Information (AMI) (Vinh, Epps, and Bailey 2009), Jaccard Index (Niwatanakul et al. 2013), Davies–Bouldin Index (DBI) (Davies and Bouldin 1979), Calinski–Harabasz Index (CHI) (Caliński and Harabasz 1974), and Purity. Detailed experimental settings are provided in the Appendix.

Resources. All experiments were run on a workstation with dual NVIDIA RTX A5000 GPUs (24 GB) and dual Intel Xeon Silver 4210R CPUs (2.40 GHz, 20 cores \times 2). GROVER converged within 300 epochs under this setup.

Method	ARI % \uparrow	NMI % \uparrow	FMI % \uparrow	SC % \uparrow	AMI % \uparrow	Jaccard % \uparrow	CHI \uparrow	Purity % \uparrow	DBI % \downarrow
Human Tonsil dataset									
GROVER	45.2\pm7.8	54.3\pm9.9	54.1\pm6.8	31.6\pm3.9	54.2\pm10.1	37.3\pm6.6	2494.4\pm285.5	69.4\pm5.4	139.8\pm10.5
MISO	41.3 \pm 6.7	51.2 \pm 4.6	52.5 \pm 4.3	7.0 \pm 1.6	51.2 \pm 4.6	35.4 \pm 3.8	244.4 \pm 14.6	64.2 \pm 5.5	203.4 \pm 14.8
Spatialglue	43.3 \pm 6.7	53.9 \pm 8.9	52.4 \pm 6.1	<u>23.8\pm3.2</u>	53.9 \pm 8.9	35.3 \pm 5.6	<u>1063.6\pm123.6</u>	<u>68.7\pm5.0</u>	159.6 \pm 7.0
COSMOS	19.8 \pm 6.7	27.9 \pm 6.0	32.3 \pm 6.6	20.0 \pm 0.7	27.6 \pm 5.6	19.3 \pm 4.9	937.4 \pm 99.6	49.9 \pm 9.0	<u>157.8\pm4.2</u>
Human Breast Cancer									
GROVER	44.1\pm10.7	52.4 \pm 8.7	53.9\pm8.6	36.3\pm7.7	52.3 \pm 8.6	37.3\pm8.1	2436.3\pm385.1	64.8 \pm 9.9	139.6\pm13.8
MISO	37.5 \pm 3.0	47.9 \pm 2.0	49.8 \pm 3.0	11.0 \pm 0.6	47.7 \pm 2.0	32.7 \pm 2.7	289.4 \pm 20.8	56.7 \pm 3.6	211.5 \pm 10.7
Spatialglue	43.0 \pm 6.9	53.0\pm5.1	<u>52.1\pm6.1</u>	20.2 \pm 0.8	53.5\pm4.8	<u>35.2\pm6.0</u>	1175.0 \pm 135.1	67.2\pm5.0	172.2 \pm 3.3
COSMOS	25.6 \pm 2.2	36.5 \pm 3.5	37.0 \pm 1.8	<u>24.8\pm0.8</u>	36.3 \pm 3.5	22.7 \pm 1.6	<u>1226.4\pm106.1</u>	54.5 \pm 2.9	<u>143.4\pm2.6</u>
Human Glioblastoma									
GROVER	40.8 \pm 6.6	53.9\pm4.1	51.6 \pm 4.6	22.6 \pm 1.1	53.8\pm3.8	34.1 \pm 4.8	<u>1412.9\pm110.7</u>	71.9\pm3.1	<u>157.0\pm3.8</u>
MISO	43.5\pm6.9	49.2 \pm 2.2	55.5\pm7.0	9.6 \pm 2.9	49.0 \pm 2.2	38.4\pm7.2	421.2 \pm 47.1	65.3 \pm 7.5	235.8 \pm 10.3
Spatialglue	40.1 \pm 7.6	<u>53.8\pm7.3</u>	50.9 \pm 5.5	<u>23.4\pm0.5</u>	<u>53.8\pm7.3</u>	33.4 \pm 5.4	1430.3\pm132.7	<u>72.3\pm3.6</u>	157.2 \pm 3.9
COSMOS	32.0 \pm 6.9	48.6 \pm 4.3	44.2 \pm 4.5	25.8\pm2.4	48.4 \pm 4.2	28.0 \pm 4.1	1325.2 \pm 92.3	67.8 \pm 3.9	137.4\pm9.0
Human Tonsil with Add-on Antibodies									
GROVER	46.5\pm5.6	59.0\pm4.8	55.3 \pm 6.0	38.2\pm1.2	58.8\pm4.7	38.0 \pm 5.7	3979.0\pm185.4	70.5\pm6.1	105.8\pm2.9
MISO	44.6 \pm 11.9	56.1 \pm 7.6	55.9\pm10.4	8.3 \pm 0.5	55.9 \pm 7.6	38.9\pm10.2	356.7 \pm 32.5	65.5 \pm 11.0	217.2 \pm 15.2
Spatialglue	45.3 \pm 7.3	<u>58.1\pm5.7</u>	54.1 \pm 7.3	21.4 \pm 1.1	58.0 \pm 5.8	36.9 \pm 6.6	<u>1331.3\pm133.9</u>	70.5\pm5.9	160.6 \pm 2.9
COSMOS	24.6 \pm 4.3	35.1 \pm 1.0	36.4 \pm 5.7	18.4 \pm 2.5	35.0 \pm 0.9	22.1 \pm 4.2	1194.0 \pm 139.2	51.5 \pm 6.8	168.8 \pm 5.7

Table 1: Performance comparison of GROVER and baselines on four spatial multi-omics datasets using nine clustering metrics. Bold indicates the best result, and underline denotes the second-best. All metrics are the higher the better, except for DBI, where a lower value indicates better performance.

Method	ARI % \uparrow	NMI % \uparrow	FMI % \uparrow	SC % \uparrow	CHI \uparrow	Purity % \uparrow	AMI % \uparrow	Jaccard % \uparrow	DBI % \downarrow
GROVER	46.5\pm5.6	59.0\pm4.8	55.3\pm6.0	38.2 \pm 1.2	3979.0 \pm 185.4	70.5\pm6.1	58.8\pm4.7	38.0\pm5.7	105.8 \pm 2.9
w/o MoE	42.5 \pm 4.3	56.8 \pm 3.0	51.6 \pm 4.8	21.8 \pm 1.2	1081.5 \pm 94.8	69.9 \pm 4.3	56.7 \pm 3.1	34.6 \pm 4.5	158.6 \pm 5.9
w/o $\mathcal{L}_{contrast}$	45.5 \pm 7.2	57.8 \pm 4.3	54.1 \pm 7.4	21.6 \pm 2.6	2201.3 \pm 306.1	68.4 \pm 6.1	57.7 \pm 4.3	37.3 \pm 7.2	168.8 \pm 10.5
w/o KAN-GCN	42.7 \pm 6.7	55.9 \pm 5.2	52.5 \pm 6.7	52.6\pm1.1	6708.4\pm247.3	68.7 \pm 6.7	55.9 \pm 5.2	35.6 \pm 6.3	89.6\pm4.7

Table 2: Ablation study results on the Human Tonsil with add-on antibodies (results on other datasets are provided in the Appendix). GROVER achieves strong performance on most clustering metrics. Removing the self-adaptive MoE module, contrastive loss, or the KAN-GCN leads to noticeable performance degradation, confirming the effectiveness of each component.

Quantitative Experimental Results

We conducted quantitative experiments on four real-world spatial multi-omics datasets. Due to the absence of ground-truth annotations spanning all modalities, we evaluated performance using cell-type clustering labels derived separately from RNA and ADT, and reported the mean and standard deviation across five clustering settings (10 to 6). Since RNA and ADT labels reflect related but non-identical biological groupings, certain methods show slightly higher variance on specific datasets. As shown in Table 1, GROVER consistently ranks among the top-performing methods across nine clustering metrics and all four datasets. For example, it improves ARI by 4.4% and SC by 32.8% on the Human Tonsil dataset, and boosts ARI and FMI by 2.6 and 3.3 percentage points on Human Breast Cancer. These results demonstrate GROVER’s strong ability to integrate heterogeneous modalities while preserving spatial and structural coherence. Notably, the bimodal method SpatialGlue often outperforms the trimodal method MISO, suggesting that uniform fusion of all modalities may not be ideal. In contrast, GROVER’s adaptive fusion strategy dynamically

adjusts modality weights per spot, leading to more effective integration tailored to local data characteristics.

Qualitative Experimental Results

We performed qualitative clustering on four spatial multi-omics datasets with histological images. Taking the Human Tonsil dataset as an example (Figure 2), GROVER’s fusion better reconstructs biological structures and clearly outlines germinal center (GC) regions and boundaries. Unlike MISO, which equally fuses all modalities, GROVER selectively emphasizes informative ones to identify regions requiring multi-modal synergy. While SpatialGlue also detects GC regions, it often splits the same type into subgroups, revealing the limitations of bi-modal integration.

Ablation Studies

To evaluate the effectiveness of each core component in the GROVER framework, we performed ablation studies on the Human Tonsil dataset with add-on antibodies. As shown in Table 2, removing the expert routing in the MoE module and replacing it with a simple summation of modality-

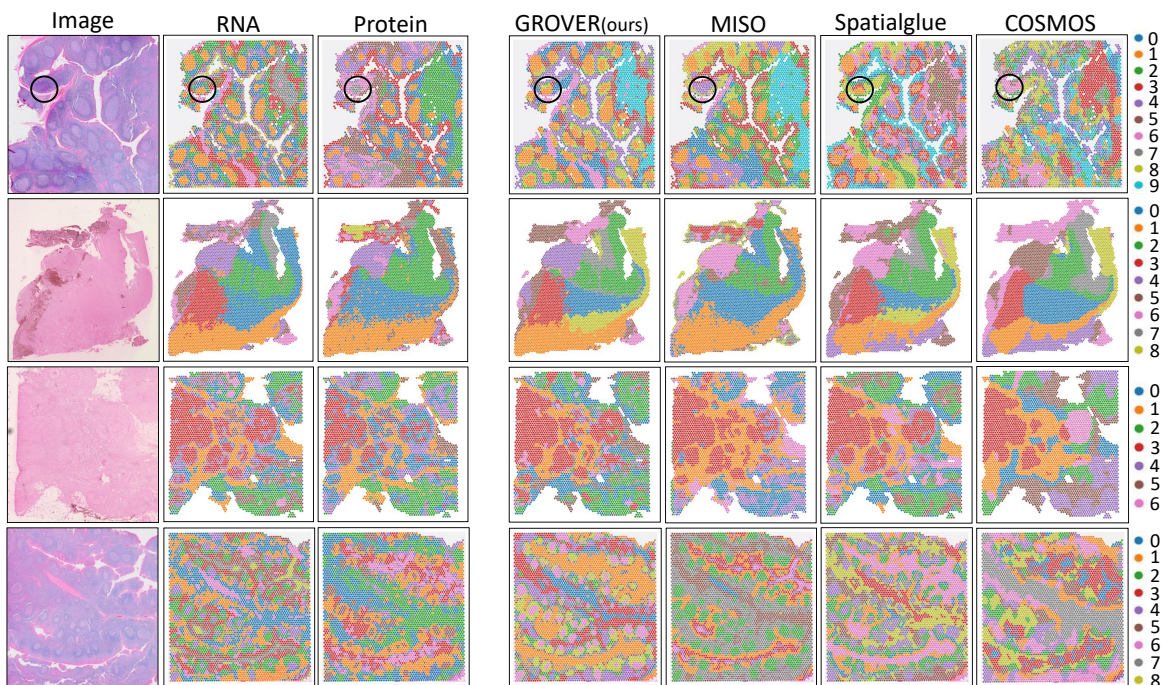


Figure 2: Visualization of clustering results by GROVER and baseline methods on four spatial multi-omics datasets. From top to bottom: (1) Human Tonsil, (2) Human Glioblastoma, (3) Human Breast Cancer, and (4) Human Tonsil with Add-on Antibodie.

specific features led to notable performance drops, with ARI, NMI, and FMI decreasing by 4.0%, 2.2%, and 3.7%, respectively. This highlights the importance of dynamically assigning experts to handle heterogeneous signal quality. Removing the spot-feature-pair contrastive loss reduced SC by 16.6%, showing its role in preserving spatial coherence. Replacing the KAN-GCN encoder with a standard GCN slightly improved unsupervised metrics SC and DBI, but clearly decreased supervised metrics, demonstrating the advantage of KAN’s nonlinear modeling in capturing complex spatial-feature interactions. Together, these results validate the necessity and effectiveness of the expert routing strategy, contrastive learning design, and KAN-GCN based encoder within the overall GROVER architecture.

Parameter Sensitivity Experiments

We conducted parameter sensitivity experiments to evaluate GROVER’s robustness to key hyperparameters, including the confidence threshold γ and contrastive loss weight λ . As shown in Figure 3, GROVER exhibits stable performance across a range of γ values in terms of ARI, NMI, and FMI, indicating that the model is not particularly sensitive to this parameter.

Specifically, extremely low values such as $\gamma = 0.1$ lead to suboptimal results, as the model fails to filter out unreliable modalities. Performance peaks around $\gamma = 0.2$ and 0.3 , where most spots with normal-quality signals are able to incorporate all modalities and benefit from richer information. Although performance remains relatively stable as γ increases, it slightly declines beyond this range due to over-reliance on a single modality, causing loss of complemen-

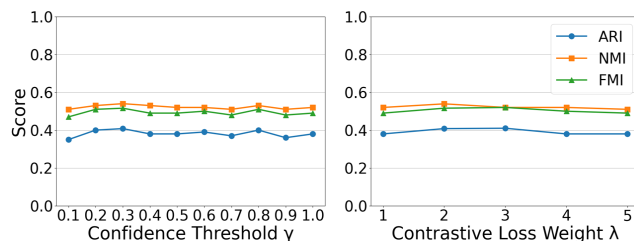


Figure 3: Parameter sensitivity analysis of GROVER on the Human Glioblastoma dataset.

tary features. Based on these observations, we set $\gamma = 0.3$ in our experiments, as it lies near the reciprocal of the number of experts and strikes a good balance between modality filtering and integration. In contrast, the model demonstrates even greater robustness to the contrastive loss weight λ , with minimal variation observed across metrics, confirming the stability of the contrastive learning component.

Conclusion

In this paper, we propose GROVER, an adaptive framework that integrates spatial multi-omics data at single-spot resolution using spatial and modality-specific graphs. With a multi-expert fusion mechanism and spot-feature-pair contrastive learning, GROVER dynamically routes information based on data quality, enabling robust integration under high heterogeneity or noise. It achieves superior performance across multiple metrics on four benchmark datasets.

Acknowledgements

This work was supported by National Natural Science Foundation of China (32300554). The computational resources are supported by SongShan Lake HPC Center (SSL-HPC) in Great Bay University.

References

- 10x Genomics. 2023a. Visium CytAssist Gene and Protein Expression Library of Human Breast Cancer, IF, 6.5mm (FFPE). <https://www.10xgenomics.com/datasets/gene-and-protein-expression-library-of-human-breast-cancer-cytassist-ffpe-2-standard>.
- 10x Genomics. 2023b. Visium CytAssist Gene and Protein Expression Library of Human Glioblastoma, IF, 11mm (FFPE). <https://www.10xgenomics.com/datasets/gene-and-protein-expression-library-of-human-glioblastoma-cytassist-ffpe-2-standard>.
- 10x Genomics. 2023c. Visium CytAssist Gene and Protein Expression Library of Human Tonsil, H&E, 6.5 mm (FFPE). <https://www.10xgenomics.com/datasets/gene-protein-expression-library-of-human-tonsil-cytassist-ffpe-2-standard>.
- 10x Genomics. 2023d. Visium CytAssist Gene and Protein Expression Library of Human Tonsil with Add-on Antibodies, H&E, 6.5 mm (FFPE). <https://www.10xgenomics.com/datasets/visium-cytassist-gene-and-protein-expression-library-of-human-tonsil-with-add-on-antibodies-h-e-6-5-mm-ffpe-2-standard>.
- Caliński, T.; and Harabasz, J. 1974. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1): 1–27.
- Chelebian, E.; Avenel, C.; and Wählby, C. 2025. Combining spatial transcriptomics with tissue morphology. *Nature Communications*, 16(1): 4452.
- Chen, W.; Zhang, P.; Tran, T. N.; Xiao, Y.; Li, S.; Shah, V. V.; Cheng, H.; Brannan, K. W.; Youker, K.; Lai, L.; et al. 2025. A visual-omics foundation model to bridge histopathology with spatial transcriptomics. *Nature Methods*, 1–15.
- Coleman, K.; Schroeder, A.; Loth, M.; Zhang, D.; Park, J. H.; Sung, J.-Y.; Blank, N.; Cowan, A. J.; Qian, X.; Chen, J.; et al. 2025. Resolving tissue complexity by multimodal spatial omics modeling with MISO. *Nature methods*, 22(3): 530–538.
- Davies, D.; and Bouldin, D. 1979. A cluster separation measure: IEEE transactions on pattern analysis and machine intelligence. itpidj 0162-8828, pami-1, 2 224–227. *Crossref Web of Science*, 21868852.
- Fowlkes, E. B.; and Mallows, C. L. 1983. A method for comparing two hierarchical clusterings. *Journal of the American statistical association*, 78(383): 553–569.
- Ge, S.; Sun, S.; Xu, H.; Cheng, Q.; and Ren, Z. 2025. Deep learning in single-cell and spatial transcriptomics data analysis: advances and challenges from a data science perspective. *Briefings in Bioinformatics*, 26(2): bbaf136.
- Huang, X.; Ma, Z.; Meng, D.; Liu, Y.; Ruan, S.; Sun, Q.; Zheng, X.; and Qiao, Z. 2025. PRAGA: prototype-aware graph adaptive aggregation for spatial multi-modal omics analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 326–333.
- Jacobs, R. A.; Jordan, M. I.; Nowlan, S. J.; and Hinton, G. E. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1): 79–87.
- Karimi, E.; Simo, N.; Milet, N.; TE, W.; ALSH, A.; QU, N.; AIL, L.; ABS, R.; ALIND, A.; GOODMA, N. M.; et al. 2024. Method of the Year 2024: spatial proteomics. *Nat Methods*, 21: 2195–2196.
- Kipf, T. 2016. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv preprint arXiv:1609.02907*.
- Liu, Z.; Wang, Y.; Vaidya, S.; Ruehle, F.; Halverson, J.; Soljačić, M.; Hou, T. Y.; and Tegmark, M. 2024. Kan: Kolmogorov-arnold networks. *arXiv preprint arXiv:2404.19756*.
- Long, Y.; Ang, K. S.; Sethi, R.; Liao, S.; Heng, Y.; van Olst, L.; Ye, S.; Zhong, C.; Xu, H.; Zhang, D.; et al. 2024. Deciphering spatial domains from spatial multi-omics with SpatialGlue. *Nature Methods*, 21(9): 1658–1667.
- Marx, V. 2021. Method of the Year: spatially resolved transcriptomics. *Nature methods*, 18(1): 9–14.
- Meng, D.; Feng, Y.; Yuan, K.; Yu, Z.; Cao, Q.; Cheng, L.; and Zheng, X. 2024. scMMAE: masked cross-attention network for single-cell multimodal omics fusion to enhance unimodal omics. *Briefings in Bioinformatics*, 26(1).
- Niwattanakul, S.; Singthongchai, J.; Naenudorn, E.; and Wanapu, S. 2013. Using of Jaccard coefficient for keywords similarity. In *Proceedings of the international multiconference of engineers and computer scientists*, volume 1, 380–384.
- Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.
- Rousseeuw, P. J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20: 53–65.
- Steinley, D. 2004. Properties of the hubert-arable adjusted rand index. *Psychological methods*, 9(3): 386.
- Totty, M.; Hicks, S. C.; and Guo, B. 2025. SpotSweeper: spatially aware quality control for spatial transcriptomics. *Nature Methods*, 1–11.
- Vinh, N. X.; Epps, J.; and Bailey, J. 2009. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th annual international conference on machine learning*, 1073–1080.
- Zhou, Y.; Xiao, X.; Dong, L.; Tang, C.; Xiao, G.; and Xu, L. 2025. Cooperative integration of spatially resolved multi-omics data with COSMOS. *Nature communications*, 16(1): 27.