SPECCG: SPECTRALLY ENHANCED JOINT CELL–GENE REPRESENTATION LEARNING FOR SPATIAL TRANSCRIPTOMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial transcriptomics (ST) provides spatially resolved gene expression profiles for characterizing tissue organization and cellular heterogeneity. However, existing representation learning methods largely overlook global gene expression patterns obscured by noise and are predominantly cell-centric, thereby failing to adequately model tissue-scale expression organization and gene-level dependency structures. To overcome these limitations, we propose **SpecCG**, a **Spec**trally enhanced joint **C**ell–**G**ene representation learning framework for spatial transcriptomics. Specifically, SpecCG first performs adaptive spectral modulation on spatial maps of gene expression to enhance informative global expression patterns obscured by noisy observations. Further, it employs a multi-view architecture to integrate local spatial-transcriptional context with spectrally enhanced global expression patterns for discriminative cell representations, while modeling gene co-expression dependencies as complementary biological signals. Finally, SpecCG introduces a gene–cell joint decoding module to couple cell and gene embeddings through joint reconstruction, enabling reciprocal supervision and coordinated cell–gene representation learning. Extensive experiments on eight public ST datasets demonstrate that SpecCG consistently outperforms state-of-the-art baselines and yields spatially coherent, biologically interpretable domains.

1 INTRODUCTION

The spatial structure of tissues fundamentally determines their biological functions, and its formation and alteration are closely related to organ development (Van Den Brink et al., 2020) and disease pathogenesis (Greenwald et al., 2024). Spatial transcriptomics (ST) enables simultaneous capture of transcriptional states and their spatial contexts by measuring gene expression profiles in situ (Miao et al., 2025; Hu et al., 2026), providing high-resolution insights into tissue architecture and underlying biological mechanisms (Williams et al., 2022). A core task in ST analysis is spatial domain identification, which partitions a tissue into functional subregions with spatially coherent gene expression patterns and is essentially a clustering process (Bai et al., 2025). Accurately resolving these regions is crucial for downstream biological research, including tumor microenvironment characterization (Zhang et al., 2019), intercellular communication network inference (Dong & Zhang, 2022), and differential gene expression analysis (Jin et al., 2024).

In recent years, deep learning methods have developed rapidly, and computational frameworks based on graph neural networks (GNNs) (Zhu et al., 2024) have emerged as the dominant paradigm for identifying spatial domains. By representing tissue cells as graph nodes to explicitly model transcriptional similarity and spatial proximity, GNN-based methods can integrate complementary spatial-transcriptional context to learn more expressive cell representations for unsupervised cell clustering (Zhu et al., 2024). Building on this foundation, a series of spatial clustering methods (Xu et al., 2024; Ren et al., 2022; Long et al., 2023) have been proposed to leverage the architectural advantages of GNNs in capturing potential cell-cell interactions, thereby enabling accurate spatial domain identification.

Despite their success, two fundamental limitations persist. 1) The global gene expression patterns encoded in the transcriptome remain underutilized, and existing methods fail to structurally exploit

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

the inherent global expression patterns. These patterns (e.g., smooth gradients along the anatomical axis or focal enrichment) contain developmental and functional partitioning information and provide reliable domain discriminative signals (Chang et al., 2024), yet are confounded by noise and missing data in ST. Some methods reconstruct the gene expression distributions to alleviate signal degradation (Wang et al., 2023), but never structurally benefit from clarified and enhanced global patterns. 2) Existing methods treat genes as isolated feature dimensions and only adopt a cell-centric modeling perspective, neglecting both crucial gene-level dependency structures and their inherent biological significance. However, genes collaborate through co-expression networks to participate in biological processes (Sun et al., 2025). Genes within the same module tend to exhibit related spatial patterns and shared functional annotations, providing a valuable level complementary to the cellular perspective. The critical neglect of gene-level information inevitably hinders the model from learning biologically interpretable tissue representation and achieving a deep understanding of biological mechanisms.

To address these challenges, we propose **SpecCG**, a **Spec**trally enhanced joint **C**ell–**G**ene representation learning framework for spatial transcriptomics. SpecCG tightly integrates gene frequency domain enhancement (GFE), multi-view cell representation (MVCR), and gene-cell joint decoding (GCJD) modules to identify coherent and biologically interpretable spatial domains. The flowchart is shown in Figure 1. It aims to recover and enhance global gene expression patterns end-to-end, simultaneously modeling ST data from both cell and gene perspectives within a unified framework. Specifically, by transforming the gridded spatial map of each gene into the frequency domain via a 2D Discrete Fourier Transform (2D DFT), the GFE module adaptively enhances global spatial expression patterns using a shared learnable spectral modulation weight tensor. Both the enhanced and original expression are then fed into the MVCR module, where complementary cell embeddings are generated by a spatial and feature-based GCN encoder. An attention mechanism is introduced to fuse multi-view cell embeddings into a consistent representation, and spatially aware contrastive learning aligns it with view-specific ones, yielding more discriminative fused embeddings and more coherent cluster structures. Concurrently, a gene co-expression GCN encodes gene embeddings from enhanced expression, explicitly capturing gene-level structures invisible to the cell-centric encoder. The GCJD module integrates gene- and cell-level representations to reconstruct gene expression profiles, enabling mutual guidance between cell and gene representations and constraining spectral modulation through joint reconstruction. The main contributions of SpecCG are summarized as follows:

- We propose SpecCG, a novel deep learning framework for clustering-oriented representation learning in spatial transcriptomics. It achieves end-to-end learnable spectral modulation, enabling models to structurally benefit from enhanced global gene expression patterns, providing reliable domain discriminative signals invisible to local methods.
- SpecCG models ST data at both gene and cell levels, effectively capturing the spatial context of cellular tissues and the co-expression structure at the gene level. The GCJD module facilitates mutual guidance and constraints between their respective representations, leveraging valuable biological information neglected by traditional cell-centric perspectives.
- Extensive experiments conducted on eight public ST datasets demonstrate that the proposed SpecCG method significantly outperforms other state-of-the-art baselines.

2 Related Work

GNN-Based Spatial Domain Identification. Driven by the inherent correspondence between ST data and graph structures, GNNs have been widely applied to spatial domain identification. Early studies focused on integrating gene expression and spatial information to fully exploit the complementary signals in ST data (Wang et al., 2023; Hu et al., 2021; Li et al., 2022). SpaGCN (Hu et al., 2021) and STAGATE (Dong & Zhang, 2022) achieved this through graph convolution and adaptive graph attention, respectively. SEDR (Xu et al., 2024) employs a variational graph autoencoder to integrate complementary information and learn spatially informed latent embeddings, effectively supporting downstream applications. However, the spatial and gene expression context information exhibits obvious semantic differences and redundant interference (He et al., 2024). To mitigate this, GraphST (Long et al., 2023) adopts a contrastive learning paradigm, enhancing representation discrimination by pulling spatial neighbors closer and pushing non-neighbors apart. MAFN (Zhu et al.,

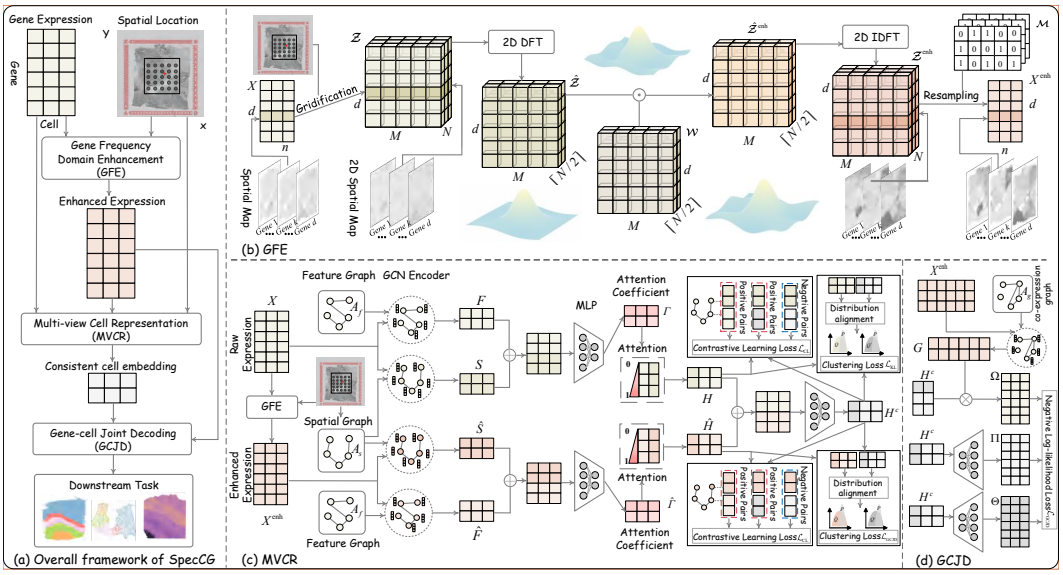

Figure 1: Flowchart of the proposed SpecCG.

2024) deploys two independent GCNs to extract spatial and transcriptional features respectively, and filters cross-view redundant information via a correlation reduction strategy. Recently, information from genes themselves as important biological symbols has been explored to promote interpretable representation learning. SAGE (He et al., 2026) combines NMF topic modeling and contrastive graph learning to screen genes with high spatial information and fuse local-non-local graphs to achieve accurate clustering and interpretable expression pattern mining. SemST (Long et al., 2026) introduces biological semantics derived from large language models to guide GNN-based spatial features by selecting highly expressed genes.

Frequency Domain Methods for ST Analysis. The application of frequency domain and spectral methods in spatial omics data is a promising emerging direction, motivated by the observation that biological spatial organization and technical noise exhibit distinct spectral signatures (Sun et al., 2025). For example, SpaGFT (Chang et al., 2024) introduces graph signal processing into spatial omics data analysis, transforming ST data from the spatial domain to the frequency domain through graph Fourier transform. SpaGFT generates interpretable graph signal representations that effectively support multiple downstream tasks. SGFN (Dai et al., 2026) uses Laplacian eigenbases to perform multi-scale spectral decomposition of spatial graphs, combining graph attention and contrastive learning to jointly model spatial and molecular features. SPLISOSM (Su et al., 2026) performs spectral decomposition of spatial patterns in ST data based on graph Fourier transform, and utilizes kernel independence tests to discover spatial variation patterns at the isoform level without prior annotation.

3 Methodology

3.1 Gene Frequency Domain Enhancement (GFE)

Gridification of ST data. The raw ST data are first preprocessed into an expression matrix $X \in \mathbb{R}^{n \times d}$ comprising n cells and d genes. To enable efficient frequency analysis, we convert X into a regular grid that preserves the relative positions of cells. Let B be the boundary rectangle of all cell coordinates, divided into an $M \times N$ uniform grid ($M \cdot N \geq n$). Each original cell is assigned to the nearest grid unit, and empty units are filled via KNN interpolation. After gridding, the expression of gene k is represented as a 2D spatial map $Z_k \in \mathbb{R}^{M \times N}$, forming a tensor $\boldsymbol{\mathcal{Z}} \in \mathbb{R}^{M \times N \times d}$ for the entire dataset. A binary mask tensor $\boldsymbol{\mathcal{M}} \in \{0,1\}^{M \times N \times d}$ is also constructed: $\boldsymbol{\mathcal{M}}_{p,q,k} = 1$ if grid unit (p,q) corresponds to an original observed cell, and 0 otherwise. This mask extracts expression values only at biological measurement locations after frequency domain processing. See Appendices A.1.1 and A.1.2 for preprocessing and more gridding details, respectively.

Frequency domain enhancement based on 2D DFT. The DFT (Li et al., 2020; Sundararajan, 2001) decomposes discrete signals into sinusoidal frequency components. For a 1D sequence $y = (y[0], y[1], \ldots, y[T-1]) \in \mathbb{R}^T$, the 1D DFT and its inverse IDFT are defined as:

$$\hat{y}[u] = \sum_{t=0}^{T-1} y[t]\, e^{-2\pi i\, ut/T}, \quad u = 0, 1, \ldots, T-1, \tag{1}$$

$$y[t] = \frac{1}{T} \sum_{u=0}^{T-1} \hat{y}[u]\, e^{2\pi i\, ut/T}, \quad t = 0, 1, \ldots, T-1. \tag{2}$$

where $\hat{y}[u] \in \mathbb{C}$ is the Fourier coefficient at frequency u.

The gene map of ST data is a 2D signal. 1D sorting along a single spatial axis collapses the true 2D tissue topology and severely loses spatial relationships, whereas 2D DFT (Ghouse, 1993) can process two spatial dimensions simultaneously and preserves their geometric relationships. For a 2D signal $Z \in \mathbb{R}^{M \times N}$ consisting of the gridded expression of a single gene, the 2D DFT is:

$$\hat{Z}[u, v] = \sum_{p=0}^{M-1} \sum_{q=0}^{N-1} Z[p, q]\, e^{-2\pi i \left(\frac{up}{M} + \frac{vq}{N}\right)}, \tag{3}$$

where $(u, v) \in \{0, \ldots, M-1\} \times \{0, \ldots, N-1\}$ is the discrete spatial frequency index. The 2D DFT can be viewed as performing 1D DFT along the two spatial dimensions alternately. For real signals, the 2D DFT exhibits conjugate symmetry: $\hat{Z}[M-u, N-v] = \overline{\hat{Z}[u, v]}$ for all (u, v), where $\bar{\cdot}$ denotes complex conjugation. Therefore, only about half the coefficients ($M \times \lceil N/2 \rceil$) are needed to completely determine the spectrum.

GFE aims to adaptively preserve and enhance potentially beneficial and biologically meaningful frequency components (such as low-frequency global anatomical patterns and mid-to-high-frequency local tissue textures). This is achieved via a learnable complex weight matrix $W \in \mathbb{C}^{M \times \lceil N/2 \rceil}$, which is shared across genes. This design reduces parameters and encourages discovery of broadly beneficial frequency modulation patterns. For implementation, W is broadcast across the gene dimension to form a tensor $\boldsymbol{\mathcal{W}} \in \mathbb{C}^{M \times \lceil N/2 \rceil \times d}$. The frequency-domain enhancement is performed via element-wise (Hadamard) multiplication:

$$\hat{\boldsymbol{\mathcal{Z}}}^{\text{enh}} = \boldsymbol{\mathcal{W}} \odot \hat{\boldsymbol{\mathcal{Z}}}, \tag{4}$$

where $\hat{\boldsymbol{\mathcal{Z}}} = \mathcal{F}_{2\text{D}}[\boldsymbol{\mathcal{Z}}] \in \mathbb{C}^{M \times \lceil N/2 \rceil \times d}$ is the frequency coefficients tensor computed from $\boldsymbol{\mathcal{Z}}$. The enhanced gridded tensor is recovered by 2D IDFT: $\boldsymbol{\mathcal{Z}}^{\text{enh}} = \mathcal{F}_{2\text{D}}^{-1}[\hat{\boldsymbol{\mathcal{Z}}}^{\text{enh}}] \in \mathbb{R}^{M \times N \times d}$.

The meshing step is crucial for applying the FFT (Ouyang et al., 2025) to improve computational efficiency and capture global spatial context. The binary mask tensor $\boldsymbol{\mathcal{M}}$ yields the final globally enhanced expression from the gridded enhanced map:

$$X^{\text{enh}} = \text{Mask}(\boldsymbol{\mathcal{Z}}^{\text{enh}}, \boldsymbol{\mathcal{M}}) \tag{5}$$

3.2 Multi-view Cell Representation (MVCR)

Dual cell graph. Two undirected graphs are first constructed to encode complementary prior information. Specifically, the spatial graph $G_s = (V, A_s)$, where the adjacency matrix A_s encodes physical proximity based on Euclidean distance. The feature graph $G_f = (V, A_f)$, where the adjacency matrix A_f uses k-nearest neighbors based on cosine similarity to capture transcriptional similarity. See Appendix A.2.1 for details.

Multi-view GCN encoding. To retain key biological signals in the original expression while exploiting globally enhanced expression patterns, we apply spatial and feature GCN encoders to both original expression X and enhanced expression X^{enh}. For the original expression, let S^l, $F^l \in \mathbb{R}^{n \times d_l}$ be view-specific embeddings of the l-th layer ($l = 0, 1, \ldots, L-1$), initialized as $S^0 = F^0 = X$. Layer-wise propagation is defined as:

$$S^{l+1} = \text{GCN}_s^{(l)}\left(S^l, \tilde{A}_s\right) = \sigma\left(\tilde{D}_s^{-\frac{1}{2}} \tilde{A}_s \tilde{D}_s^{-\frac{1}{2}}\, S^l\, \Theta_s^l\right), \tag{6}$$

$$F^{l+1} = \text{GCN}_f^{(l)}\left(F^l, \tilde{A}_f\right) = \sigma\left(\tilde{D}_f^{-\frac{1}{2}} \tilde{A}_f \tilde{D}_f^{-\frac{1}{2}}\, F^l\, \Theta_f^l\right), \tag{7}$$

where Θ^l are trainable weights, $\tilde{A} = A + I$ is the adjacency matrix with self-loops, $\tilde{D}$ is the degree matrix, and $\sigma(\cdot)$ is ReLU. The same architecture is applied to X^{enh}, yielding enhanced embeddings $\hat{S}^l$ and $\hat{F}^l$. After L layers, four view-specific cell embeddings S, F, $\hat{S}$ and $\hat{F} \in \mathbb{R}^{n \times d_L}$ are obtained.

Cross-view fusion and alignment. An attention mechanism (Peng et al., 2022) is first employed to integrate complementary signals across views. For the original view, attention coefficients are computed as:

$$\Gamma = [\gamma_s, \gamma_f] = \ell_2(\text{softmax}(\sigma([S \,\|\, F]\, U))), \tag{8}$$

where $[\cdot \,\|\, \cdot]$ denotes column concatenation, U is a learnable weight matrix, σ is LeakyReLU, and ℓ_2 denotes L2 normalization. The fused embedding of the original view is:

$$H = \left(\gamma_s \mathbf{1}_{d_L}^\top\right) \odot S + \left(\gamma_f \mathbf{1}_{d_L}^\top\right) \odot F, \tag{9}$$

where $\mathbf{1}_{d_L} \in \mathbb{R}^{d_L}$ is an all-ones vector and $\odot$ is element-wise broadcast multiplication. The attention scores $\gamma_{s,i}$ and $\gamma_{f,i}$ adaptively balance each view's contribution for cell i.

The enhanced view is fused similarly from $\hat{S}$ and $\hat{F}$, yielding $\hat{H} \in \mathbb{R}^{n \times d_L}$. A consistent cell representation is then obtained by integrating information from both views:

$$H^c = \text{MLP}_\text{h}\Big([H \,\|\, \hat{H}]\Big) \in \mathbb{R}^{n \times d_L} \tag{10}$$

To further improve the quality of the learned representations, a spatially aware contrastive learning framework is introduced. It aligns the semantics of cross-view cell representations and encodes tissue continuity priors into the embedding topology. Specifically, fused embeddings are aligned with the consistent embedding. The contrastive loss between the original view and the consistent embedding is:

$$\ell_i = -\log \frac{e^{\psi(H_i, H_i^c)/\tau} + \sum_{j \in \mathcal{N}_s(i)} e^{\psi(H_i, H_j)/\tau}}{\sum_{j \neq i} e^{\psi(H_i, H_j)/\tau} + \sum_{j \notin \mathcal{N}_s(i)} e^{\psi(H_i, H_j^c)/\tau}}. \tag{11}$$

where ψ is cosine similarity, τ is the temperature coefficient, and $\mathcal{N}_s(i)$ is the set of spatial neighbors of cell i in G_s. The contrastive loss ℓ_i^{enh} of enhanced view is defined similarly. The total contrastive loss is:

$$\mathcal{L}_{\text{CL}} = \frac{1}{2n} \sum_{i=1}^{n} \Big(\ell_i + \ell_i^{\text{enh}}\Big). \tag{12}$$

3.3 Gene-cell joint decoding (GCJD)

Gene co-expression modeling. Compared to cell-level representation learning, gene-level co-expression modules (Song et al., 2022) provide inherent biological guidance. To model this complementary information source, a gene co-expression graph $G_g = (V_g, A_g)$ is constructed, where nodes are genes and edges represent co-expression relationships. The edge weight between genes i and j is defined as the Pearson correlation coefficient (Dufera et al., 2023) between their expression profiles across all cells. See Appendix A.2.2 for co-expression graph details.

A GCN is then used to embed co-expression relationships into each gene. The enhanced expression matrix X^{enh} is transposed as the initial node feature: $G^0 = (X^{\text{enh}})^\top \in \mathbb{R}^{d \times n}$. Information propagates through an L-layer GCN:

$$G^{l+1} = \text{GCN}_g^{(l)}\big(G^l, \tilde{A}_g\big), \tag{13}$$

where $\tilde{A}_g$ is the adjacency matrix with self-loops. The final output $G = G^L \in \mathbb{R}^{d \times d_L}$ maps each gene into the same d_L-dimensional latent space as the cell embedding, which facilitates direct interaction between cell and gene representations and joint decoding of gene expression.

Gene-cell joint decoding. The zero-inflated negative binomial (ZINB) (Yu et al., 2022) distribution effectively models the overdispersion and sparsity of ST data. For cell i and gene j, the observed expression count x_{ij} is modeled as:

$$\text{ZINB}(x_{ij}|\pi, \mu, \theta) = \pi_{ij} \cdot \delta_0(x_{ij}) + (1 - \pi_{ij}) \cdot \text{NB}(x_{ij}|\mu_{ij}, \theta_{ij}) \tag{14}$$

where $\pi_{ij} \in [0, 1]$ is the zero-inflation probability (missing-event-induced zeros) and $\text{NB}(x_{ij} \mid \mu_{ij}, \theta_{ij})$ is a negative binomial probability mass function with mean μ_{ij} and dispersion θ_{ij}.

Unlike conventional cell-only ZINB decoding, the mean matrix $\Omega = [\mu_{ij}]$ is integrated from the consistent cell embeddings H^c and gene embeddings G, rather than decoded solely from the cell embeddings:

$$\mu_{ij} = \mathrm{softplus}(H_i^c G_j^{\mathrm{T}}) = \log(1 + \exp(H_i^c G_j^{\mathrm{T}})) \tag{15}$$

The softplus transformation ensures non-negativity. The zero-inflation parameter $\Pi = [\pi_{ij}]$ and the dispersion parameter $\Theta = [\theta_{ij}]$ are decoded separately from H^c:

$$\Pi = \mathrm{sigmoid}\big(\mathrm{MLP}_\pi(H^c)\big),\ \Theta = \mathrm{softplus}\big(\mathrm{MLP}_\theta(H^c)\big). \tag{16}$$

The final joint reconstruction loss is the negative log-likelihood:

$$\mathcal{L}_{\mathrm{GCJD}} = -\frac{1}{nd}\sum_{i=1}^{n}\sum_{j=1}^{d}\log \mathrm{ZINB}\big(x_{ij} \mid \pi_{ij}, \mu_{ij}, \theta_{ij}\big). \tag{17}$$

3.4 Model Optimization

Self-optimizing clustering strategy. To encourage cluster-oriented final representations in an unsupervised manner, a self-optimizing training strategy based on KL divergence is adopted. First, k-means is applied to the consistent embedding H^c to obtain K cluster centers c_j. The soft cluster assignments are calculated using the Student-t distribution (van der Maaten & Hinton, 2008) with 1 degree of freedom:

$$q_{ij}^c = \frac{\big(1 + \|H_i^c - c_j\|_2^2\big)^{-1}}{\sum_{k=1}^{K}\big(1 + \|H_i^c - c_k\|_2^2\big)^{-1}}, \tag{18}$$

Similarly, Q and $\hat{Q}$ are calculated from H and $\hat{H}$. The target distribution P amplifies high-confidence assignments and suppresses low-confidence assignments through squaring operations:

$$p_{ij} = \frac{(q_{ij}^c)^2 \big/ \sum_{i=1}^{n} q_{ij}^c}{\sum_{k=1}^{K}\big((q_{ik}^c)^2 \big/ \sum_{i=1}^{n} q_{ik}^c\big)}. \tag{19}$$

The final self-optimizing clustering loss is:

$$\mathcal{L}_{\mathrm{KL}} = \mathrm{KL}(P \,\|\, Q^c) + \mathrm{KL}(P \,\|\, Q) + \mathrm{KL}(P \,\|\, \hat{Q}), \tag{20}$$

Therefore, the cluster distributions of different cell embeddings can be well aligned, while promoting the acquisition of consistent discriminative representations oriented towards clustering.

Overall loss function. The total objective loss of SpecCG consists of joint reconstruction loss, contrastive learning loss, and self-optimizing clustering loss; therefore, the total loss is:

$$\mathcal{L} = \mathcal{L}_{\mathrm{GCJD}} + \beta\,\mathcal{L}_{\mathrm{CL}} + \gamma\,\mathcal{L}_{\mathrm{KL}}, \tag{21}$$

where $\beta, \gamma > 0$ are hyperparameters used to balance the effects of different components.

SpecCG can capture global spatial interaction information and enhance expression information, see Appendix A.3 for a detailed theoretical proof. And see Appendix A.4 for the SpecCG algorithmic procedure.

4 Experiments

4.1 Experimental Settings

Datasets. Eight public ST benchmark datasets are used to evaluate the proposed method. These include the human dorsolateral prefrontal cortex (DLPFC) (Huuki-Myers et al., 2024), human breast cancer (HBC) (Wang et al., 2025), and mouse brain anterior tissue (MBA) datasets (Zhu et al., 2024) from the 10× Visium platform (Ji et al., 2020), as well as the mouse embryo (ME9.5) dataset (Chen et al., 2022) from the Stereo-seq platform (Long et al., 2024). The DLPFC dataset contains 12 slices, each manually annotated into 5 or 7 cortical layers. Five of these slices (151510, 151669, 151670, 151671, and 151674) are selected for evaluation.

Table 1: Clustering performance on eight spatial transcriptomics datasets, with **bold** indicating the best and underline indicating the second best.

Datasets	151510		151669		151670		151671		151674		HBC		MBA		ME9.5		Average	
Methods	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI
GraphST (Nat. Commun.'23)	50.11	60.37	53.07	60.24	44.99	56.74	57.05	71.64	45.70	58.96	50.94	64.78	35.68	69.48	30.14	54.47	45.96	62.09
Spatial-MGCN (BIB'23)	52.73	67.06	39.17	59.63	46.16	55.90	60.37	73.05	53.96	65.08	62.76	65.99	45.84	67.65	42.28	55.83	50.41	63.77
SEDR (Genome Med.'24)	37.08	54.46	37.45	54.06	40.02	56.16	50.25	61.69	43.33	59.85	50.02	66.56	30.44	68.17	27.98	54.18	39.57	59.39
MAFN (TKDE'24)	67.33	69.38	59.50	63.56	52.80	61.11	84.52	79.01	53.61	63.85	62.26	66.20	41.77	69.72	40.44	54.11	57.78	65.87
DUSTED (AAAI'25)	49.84	60.24	46.11	59.93	42.26	57.71	59.01	69.23	40.70	55.63	44.08	62.76	39.52	71.94	31.41	55.67	44.12	61.64
SpaSEG (Genome Biol.'25)	49.24	59.87	37.88	59.66	39.24	55.01	55.35	65.80	48.32	63.01	56.99	64.77	43.41	66.01	27.89	53.68	44.79	60.98
stDCL (Adv. Sci.'25)	46.69	61.72	52.21	61.23	34.33	49.70	63.32	73.98	37.51	53.55	53.26	68.07	40.24	69.61	35.27	**59.43**	45.35	62.16
SAGE (Adv. Sci.'26)	49.38	63.64	47.63	61.01	46.55	59.92	57.03	68.70	54.90	68.43	55.79	68.50	39.69	71.76	38.28	55.07	48.66	64.63
STMCCL (JBHI'26)	47.74	62.42	66.02	57.77	53.78	55.25	77.03	68.33	57.98	68.83	64.17	69.57	41.63	71.86	33.11	47.31	55.18	62.67
STCF (TPAMI'26)	59.99	65.07	45.25	55.63	40.80	49.06	73.51	68.94	53.48	61.58	64.99	67.12	47.77	71.54	41.04	55.44	53.35	61.80
SpecCG (Ours)	**69.21**	**69.59**	**84.27**	**70.88**	**87.15**	**74.79**	**88.28**	**82.73**	**65.87**	**71.51**	**67.24**	**73.21**	**49.04**	**71.95**	**43.82**	57.71	**69.36**	**71.55**

Datasets	151510		151669		151670		151671		151674		HBC		MBA		ME9.5		Average	
Methods	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
GraphST (Nat. Commun.'23)	66.81	54.77	70.57	59.14	69.60	61.76	68.14	66.34	61.93	56.34	53.29	44.48	44.16	33.63	46.22	42.71	60.09	52.40
Spatial-MGCN (BIB'23)	60.11	52.52	58.91	55.91	66.79	48.33	72.76	71.02	63.49	56.06	61.43	46.30	46.72	37.87	53.37	46.26	60.45	51.78
SEDR (Genome Med.'24)	58.13	49.45	54.90	50.01	57.29	56.26	67.65	63.73	56.78	54.01	55.74	49.39	40.33	36.92	45.17	42.94	54.50	50.34
MAFN (TKDE'24)	73.36	63.95	72.22	58.55	70.84	55.50	88.27	80.15	67.59	58.71	63.59	49.47	46.98	38.84	55.18	46.88	67.25	56.51
DUSTED (AAAI'25)	63.35	51.60	67.77	56.41	64.78	53.47	73.83	67.53	54.80	53.30	48.13	41.17	47.42	41.68	46.32	37.45	58.30	50.33
SpaSEG (Genome Biol.'25)	68.10	51.12	57.43	52.57	59.76	43.31	70.17	64.14	60.88	56.42	61.11	36.08	42.34	20.32	43.65	39.97	57.93	45.49
stDCL (Adv. Sci.'25)	69.40	62.49	69.72	47.91	58.04	38.81	77.01	70.27	53.51	49.08	54.84	42.79	45.01	32.15	52.21	48.59	59.97	49.01
SAGE (Adv. Sci.'26)	69.29	52.11	68.40	55.83	67.19	52.99	71.95	66.67	67.24	**63.61**	55.90	46.66	46.12	35.94	55.03	**49.94**	62.64	52.97
STMCCL (JBHI'26)	68.77	58.49	76.79	50.72	71.33	59.80	81.16	64.38	71.09	61.53	63.98	49.48	48.91	39.49	48.25	35.37	66.29	52.41
STCF (TPAMI'26)	70.97	59.90	63.83	54.06	63.83	49.15	79.23	62.74	67.15	52.49	62.56	45.26	50.50	40.90	50.19	36.32	63.53	50.10
SpecCG (Ours)	**82.13**	**75.83**	**79.90**	**63.15**	**81.57**	**68.43**	**92.55**	**88.65**	**71.28**	56.73	**68.67**	**55.76**	**51.47**	**41.79**	**59.16**	45.68	**73.34**	**62.00**

Baselines. Ten state-of-the-art and representative methods are selected as baselines for comparison, including GraphST (Long et al., 2023), Spatial-MGCN (Wang et al., 2023), SEDR (Xu et al., 2024), MAFN (Zhu et al., 2024), DUSTED (Zhu et al., 2025), SpaSEG (Bai et al., 2025), stDCL (Yu et al., 2025), SAGE (He et al., 2026), STMCCL (Wang et al., 2025), and STCF (Zhu et al., 2026). All baselines are implemented with the default experimental settings from their original papers to achieve optimal performance. See Appendices A.5.1 and A.5.2 for details of the datasets and baselines, respectively.

Implementation Details. Experiments are conducted on a machine equipped with an AMD Ryzen 9 7945HX processor (2.50 GHz), 16 GB RAM, and an NVIDIA GeForce RTX 4060 GPU. Our framework is implemented based on PyTorch 2.4.0 and uses the Adam optimizer for model training. The learning rate and weight decay coefficient are set to 0.001 and 5×10^{-4} respectively, and the model is trained for 300 epochs. To comprehensively evaluate the proposed method, four clustering metrics are adopted. They are Accuracy (ACC) (Yu et al., 2024), Normalized Mutual Information (NMI) (Mei et al., 2024), Adjusted Rand Index (ARI) (Peng et al., 2023), and F1 score (Peng et al., 2023). Higher values for these four metrics generally indicate better clustering performance. See Appendix A.5.3 for detailed parameter settings.

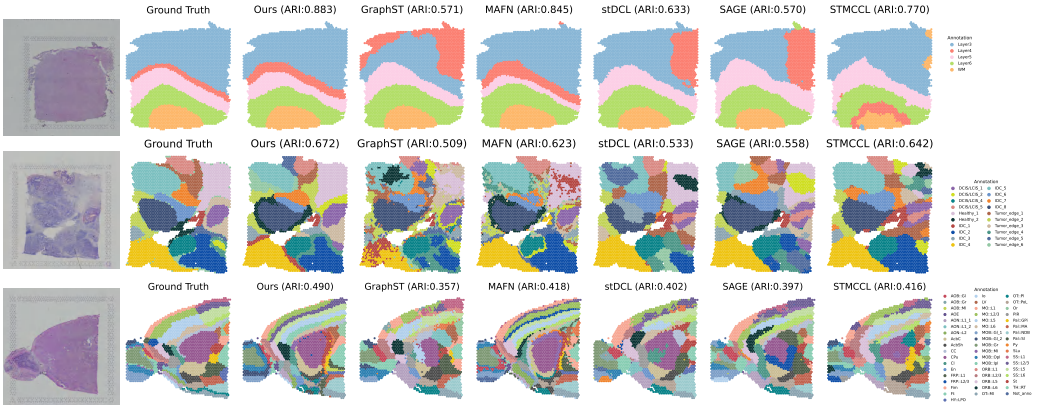

Figure 2: Visualization of spatial clustering across datasets, including histopathological images, manual annotations, and identified spatial domains. From top to bottom, the rows correspond to the 151671, HBC, and MBA datasets.

4.2 CLUSTERING RESULTS

To verify the performance advantages of the proposed SpecCG framework, Table 1 lists the numerical clustering results of all methods on different datasets. It is evident that SpecCG consistently outperforms all benchmarks in ARI and ACC across all eight datasets, and achieves the best NMI and F1 scores on the vast majority of datasets. Particularly on slice 151670, SpecCG outperforms the state-of-the-art benchmarks by 62.05%, 22.39%, 14.36%, and 10.80% in ARI, NMI, ACC, and F1 metrics respectively. Even on the HBC and MBA datasets with unclear domain boundaries and complex tissue distribution, SpecCG still achieves superior clustering performance. The quantitative results of average clustering performance further confirm the superiority of SpecCG. The visualization results shown in Figure 2 demonstrate that SpecCG produces domain partitions with clear boundaries that better resemble manual annotations, while other methods tend to deviate from the reference partition or produce fragmented spatial domains. This observation is consistent with the embedding results shown in Figure 7. This indicates that the proposed SpecCG can fully leverage global gene patterns and co-expression structures to produce clustering results that are more biologically meaningful and spatially coherent. More detailed analysis including statistical testing can be found in Appendices A.6.1 and A.6.2.

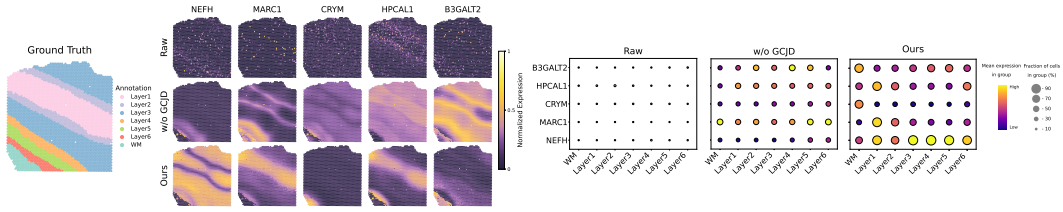

Figure 3: Visualization and quantitative results of the global spatial expression patterns of five cortical layer marker genes on 151510.

4.3 GLOBAL GENE EXPRESSION DISTRIBUTION

To verify that the proposed SpecCG can simultaneously leverage long-range interactions among cells and co-expression structures among genes to promote the reconstruction of global expression patterns, we visualized the spatial expression distributions of five layer marker genes on slice 151510. As shown in Figure 3, the raw expression patterns of these five marker genes are fragmented and difficult to discern, while SpecCG effectively reconstructs and enhances their global expression patterns. The quantitative results in Figure 3 clearly demonstrate that the marker genes reconstructed by SpecCG exhibit distinct differential expression patterns across cortical layers, and are consistent with the annotated cortical distribution. However, when the GCJD module is removed, the layer distribution differences of the marker genes decrease, and the spatial expression patterns of some marker genes (such as CRYM) are reconstructed poorly. This improvement may arise from GCJD's joint use of cell-level interactions and gene co-expression structures for expression reconstruction, as well as its provision of biologically informed guidance for spectral enhancement, thereby promoting the reconstruction and enhancement of the global expression pattern.

4.4 ABLATION STUDY

To evaluate the impact of different components on the overall performance, several variants are proposed for comprehensive comparative analysis. Table 2 presents the comparative results on four datasets. Specifically, w/o GFE represents the variant with the frequency domain enhancement module removed, and its performance degradation in different datasets can be observed to varying degrees. The visualization in Figure 8 more intuitively demonstrates that GFE facilitates more discriminative cell embeddings and more coherent cell cluster structures. This stems from the fact that the GFE module can automatically explore beneficial frequency modulation patterns through end-to-end design to enhance the global gene expression landscape, thereby providing reliable domain discrimination signals for the learning process. w/o GCJD denotes the variant that reconstructs gene expression solely from cell embeddings. The results show that neglecting co-expression structures

Table 2: Ablation experiments on different datasets.

Datasets	151671				HBC				MBA				ME9.5			
Variants	ARI	NMI	ACC	F1	ARI	NMI	ACC	F1	ARI	NMI	ACC	F1	ARI	NMI	ACC	F1
w/o GFE	83.47	77.38	86.73	80.65	61.24	68.09	62.35	51.69	39.90	71.13	48.42	41.04	36.24	51.18	45.85	37.71
w/o GCJD	78.59	71.03	78.92	66.26	64.70	69.06	64.74	52.29	41.18	69.81	46.46	38.14	37.48	54.21	47.23	41.85
w/o $\mathcal{L}_{\text{GCJD}}$	64.85	60.48	70.00	49.41	54.30	67.05	58.53	48.87	37.21	70.52	46.20	39.18	35.24	53.74	45.76	39.17
w/o $\mathcal{L}_{\text{CL}}$	85.13	81.01	91.52	87.76	53.82	62.83	56.40	46.03	37.41	67.41	42.97	34.58	39.18	51.90	48.55	39.25
w/o $\mathcal{L}_{\text{KL}}$	67.68	67.09	76.08	58.62	65.69	70.51	65.69	53.04	44.87	71.00	49.28	39.47	38.80	55.44	51.14	44.76
Ours	**88.28**	**82.73**	**92.55**	**88.65**	**67.24**	**73.21**	**68.67**	**55.76**	**49.04**	**71.95**	**51.47**	**41.79**	**43.82**	**57.71**	**59.16**	**45.68**

and functional associations at the gene level will yield suboptimal clustering results. This also reflects that the GCJD module can effectively couple complementary information from the cell and gene levels to promote biologically meaningful tissue representation. By removing different loss terms individually, the contributions of ZINB decoding, spatially aware contrastive learning, and self-supervised clustering are evaluated. The performance drop after removing the inter-view contrastive loss confirms that it also improves model performance to varying degrees by preserving semantic and spatial consistency. The results in Table 2 further indicate that gene distribution decoding and self-supervised clustering are also effective components that contribute to the model's performance.

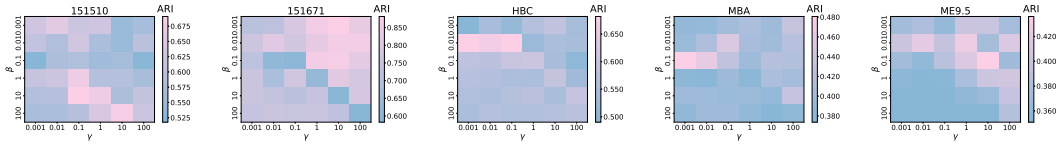

Figure 4: Heatmaps of ARI across combinations of β and γ.

4.5 Hyperparameter Analysis

The parameters to be determined include the regularization parameter of the loss term and the temperature coefficient of contrastive learning. To evaluate the impact of hyperparameters on clustering performance, we analyzed the effects of β and γ on five datasets (Figure 4) and the effect of τ on all eight datasets (Figure 9). Figure 4 shows the cross-analysis heatmap of β and γ, with both values set in the range $\{0.001, 0.01, 0.1, 1, 10, 100\}$. It can be observed that specific optimal values for β and γ need to be set for different datasets. An overly large or small β will lead to a decrease in clustering performance on HBC, MBA, and ME9.5. The temperature coefficient ranges over $\{0.1, 0.2, \ldots, 1.0\}$, as shown in Figure 9. It can be observed that for most datasets, variations in τ have a relatively pronounced effect on the overall clustering performance. Therefore, better parameter settings should be searched for different datasets. For example, τ can be set to 0.9 on 151510. The final detailed parameter search results can be found in Appendix A.5.3.

5 Conclusion

In this work, we propose SpecCG, a novel deep learning framework to address key challenges in spatial transcriptomics clustering. By jointly enhancing gene expression patterns in the spectral domain and capturing co-expression structures, SpecCG can restore the disrupted global gene expression patterns end-to-end. It enables joint analysis of spatial transcriptomics data at both gene and cell levels. The key biological signals from the original gene expression profiles are organically coupled with the reliable domain discriminative signals provided by enhanced expression through a multi-view architecture. This promotes more biologically meaningful and discriminative cell representations. Extensive experiments on eight spatial transcriptomics datasets demonstrate that the proposed SpecCG achieves significant performance improvements, providing new insights into biologically meaningful tissue characterization and revealing related biological mechanisms. Future work will further combine multi-omics information such as proteomics and explore the use of protein-protein interaction to guide the reconstruction of gene co-expression modules.

6 AI USAGE STATEMENT

During the preparation of this manuscript, Large Language Models (LLMs) were used to assist with writing and language refinement, including enhancing grammatical accuracy, clarity of expression, and coherence across sections. The core research contributions originated from the authors, who led the development of SpecCG, the experimental evaluation, and the interpretation of the results. The authors retained control over the manuscript and all editorial decisions, utilizing AI assistance to review and refine wording while ensuring the final text accurately conveyed the intended scientific meaning. We acknowledge this assistance to promote transparency and accept full responsibility for the accuracy and integrity of the final manuscript.

7 ETHICS STATEMENT

Throughout this study, we have been committed to upholding research integrity and transparency while respecting data privacy. Our experiments utilized publicly available spatial transcriptomics datasets and did not involve the recruitment of new participants or the collection of biological samples. We acknowledge the contributions of the original data providers and previous studies, and we strive to accurately describe our methods and findings. Through this work, we aim to improve computational tools for understanding tissue architecture and cellular heterogeneity, thereby supporting further biological research.

8 REPRODUCIBILITY STATEMENT

We have endeavored to ensure the reproducibility of our research by clearly describing the methodology and experimental procedures. The main text and Algorithm 1 outline the model construction and optimization processes, while the appendices (specifically Appendices A.1, A.2, A.4 and A.5) provide the specific details required to replicate the experiments. The source code for SpecCG is provided in the supplementary materials.

REFERENCES

Yong Bai, Xiangyu Guo, Keyin Liu, Bingjie Zheng, Yilin Wei, Yingyue Wang, Wenxi Zhang, Qiuhong Luo, Jianhua Yin, Liang Wu, et al. Spaseg: unsupervised deep learning for multi-task analysis of spatially resolved transcriptomics. *Genome Biology*, 26(1):230, 2025.

Yuzhou Chang, Jixin Liu, Yi Jiang, Anjun Ma, Yao Yu Yeo, Qi Guo, Megan McNutt, Jordan E Krull, Scott J Rodig, Dan H Barouch, et al. Graph fourier transform for spatial omics representation and analyses of complex organs. *Nature Communications*, 15(1):7467, 2024.

Ao Chen, Sha Liao, Mengnan Cheng, Kailong Ma, Liang Wu, Yiwei Lai, Xiaojie Qiu, Jin Yang, Jiangshan Xu, Shijie Hao, et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using dna nanoball-patterned arrays. *Cell*, 185(10):1777–1792, 2022.

Ruoyan Dai, Zhenghui Wang, Zhiwei Zhang, Lixin Lei, Mengqiu Wang, Zhenxing Li, Xingyu Liu, and Qianjin Guo. A multi-scale graph frequency network for structural and functional region analysis in spatial transcriptomics. *Functional & Integrative Genomics*, 26(1):176, 2026.

Kangning Dong and Shihua Zhang. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nature communications*, 13(1):1739, 2022.

Abdisa G Dufera, Tiantian Liu, and Jin Xu. Regression models of pearson correlation coefficient. *Statistical Theory and Related Fields*, 7(2):97–106, 2023.

Mohammed Ahmed Ghouse. 2d grid architectures for the dft and the 2d dft. *Journal of VLSI signal processing systems for signal, image and video technology*, 5(1):57–74, 1993.

Alissa C Greenwald, Noam Galili Darnell, Rouven Hoefflin, Dor Simkin, Christopher W Mount, L Nicolas Gonzalez Castro, Yotam Harnik, Sydney Dumont, Dana Hirsch, Masashi Nomura, et al. Integrative spatial analysis reveals a multi-layered organization of glioblastoma. *Cell*, 187 (10):2485–2501, 2024.

Xiao He, Chang Tang, Xinwang Liu, Chuankun Li, Shan An, and Zhenglai Li. Heterogeneous graph guided contrastive learning for spatially resolved transcriptomics data. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 8287–8295, 2024.

Yi He, Yunpei Xu, Liqing Ding, Hong-Dong Li, Yaohang Li, and Shaokai Wang. Sage: Spatially aware gene selection and dual-view embedding fusion for domain identification in spatial transcriptomics. *Advanced Science*, 13(16):e20333, 2026.

Chenlei Hu, Mehdi Borji, Giovanni J Marrero, Vipin Kumar, Jackson A Weir, Sachin V Kammula, Evan Z Macosko, and Fei Chen. Scalable spatial transcriptomics through computational array reconstruction. *Nature biotechnology*, 44(2):215–221, 2026.

Jian Hu, Xiangjie Li, Kyle Coleman, Amelia Schroeder, Nan Ma, David J Irwin, Edward B Lee, Russell T Shinohara, and Mingyao Li. Spagcn: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature methods*, 18(11):1342–1351, 2021.

Louise A Huuki-Myers, Abby Spangler, Nicholas J Eagles, Kelsey D Montgomery, Sang Ho Kwon, Boyi Guo, Melissa Grant-Peters, Heena R Divecha, Madhavi Tippani, Chaichontat Sriworarat, et al. A data-driven single-cell and spatial transcriptomic map of the human prefrontal cortex. *Science*, 384(6698):eadh1938, 2024.

Andrew L Ji, Adam J Rubin, Kim Thrane, Sizun Jiang, David L Reynolds, Robin M Meyers, Margaret G Guo, Benson M George, Annelie Mollbrink, Joseph Bergenstråhle, et al. Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *cell*, 182(2): 497–514, 2020.

Yang Jin, Yuanli Zuo, Gang Li, Wenrong Liu, Yitong Pan, Ting Fan, Xin Fu, Xiaojun Yao, and Yong Peng. Advances in spatial transcriptomics and its applications in cancer research. *Molecular Cancer*, 23(1):129, 2024.

Fae N Kronman, Josephine K Liwang, Rebecca Betty, Daniel J Vanselow, Yuan-Ting Wu, Nicholas J Tustison, Ashwin Bhandiwad, Steffy B Manjila, Jennifer A Minteer, Donghui Shin, et al. Developmental mouse brain common coordinate framework. *Nature communications*, 15(1):9072, 2024.

Jiachen Li, Siheng Chen, Xiaoyong Pan, Ye Yuan, and Hong-Bin Shen. Cell clustering for spatial transcriptomics data with graph neural networks. *Nature Computational Science*, 2(6):399–408, 2022.

Shaohua Li, Kaiping Xue, Bin Zhu, Chenkai Ding, Xindi Gao, David Wei, and Tao Wan. Falcon: A fourier transform based approach for fast and secure convolutional neural network predictions. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8702–8711. IEEE, 2020.

Jiangkai Long, Yanran Zhu, Chang Tang, Kun Sun, Yuanyuan Liu, and Xuesong Yan. When genes speak: A semantic-guided framework for spatially resolved transcriptomics data clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 800–808, 2026.

Yahui Long, Kok Siong Ang, Mengwei Li, Kian Long Kelvin Chong, Raman Sethi, Chengwei Zhong, Hang Xu, Zhiwei Ong, Karishma Sachaphibulkij, Ao Chen, et al. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with graphst. *Nature communications*, 14(1):1155, 2023.

Yahui Long, Kok Siong Ang, Raman Sethi, Sha Liao, Yang Heng, Lynn Van Olst, Shuchen Ye, Chengwei Zhong, Hang Xu, Di Zhang, et al. Deciphering spatial domains from spatial multiomics with spatialglue. *Nature methods*, 21(9):1658–1667, 2024.

Jianping Mei, Xiangli Li, and Yuanjian Mo. Dual semi-supervised hypergraph regular multi-view nmf with anchor graph embedding. *Knowledge-Based Systems*, 305:112662, 2024.

Jishuai Miao, Jinzhao Li, Jingxue Xin, Jiajuan Tu, Muyang Ge, Ji Qi, Xiaocheng Zhou, Ying Zhu, Can Yang, and Zhixiang Lin. Multigate: integrative analysis and regulatory inference in spatial multi-omics data via graph representation learning. *Nature Communications*, 16(1):9403, 2025.

Yi Ouyang, Yihong Zhu, Wenping Zhu, Bohan Yang, Zirui Zhang, Hanning Wang, Qichao Tao, Min Zhu, Shaojun Wei, and Leibo Liu. Falconsign: An efficient and high-throughput hardware architecture for falcon signature generation. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2025(1):203–226, 2025.

Siyuan Peng, Jingxing Yin, Zhijing Yang, Badong Chen, and Zhiping Lin. Multiview clustering via hypergraph induced semi-supervised symmetric nonnegative matrix factorization. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(10):5510–5524, 2023.

Zhihao Peng, Hui Liu, Yuheng Jia, and Junhui Hou. Deep attention-guided graph clustering with dual self-supervision. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(7): 3296–3307, 2022.

Honglei Ren, Benjamin L Walker, Zixuan Cang, and Qing Nie. Identifying multicellular spatiotemporal organization of cells with spaceflow. *Nature communications*, 13(1):4076, 2022.

Tianci Song, Kathleen K Markham, Zhuliu Li, Kristen E Muller, Kathleen Greenham, and Rui Kuang. Detecting spatially co-expressed gene clusters with functional coherence by graph-regularized convolutional neural network. *Bioinformatics*, 38(5):1344–1352, 2022.

Jiayu Su, Yiming Qu, Megan Schertzer, Haochen Yang, Jiahao Jiang, Tenzin Lhakhang, Theodore M Nelson, Stella Park, Qiliang Lai, Xi Fu, et al. Mapping isoforms and regulatory mechanisms from spatial transcriptomics data with splisosm. *Nature Biotechnology*, pp. 1–12, 2026.

Shuli Sun, Jixin Liu, Guojun Li, and Bingqiang Liu. Deepgft: identifying spatial domains in spatial transcriptomics of complex and 3d tissue using deep learning and graph fourier transform. *Genome Biology*, 26(1):153, 2025.

D Sundararajan. The discrete fourier transform: Theory. *Algorithms and Applications*, 2001.

Susanne C Van Den Brink, Anna Alemany, Vincent van Batenburg, Naomi Moris, Marloes Blotenburg, Judith Vivié, Peter Baillie-Johnson, Jennifer Nichols, Katharina F Sonnen, Alfonso Martinez Arias, et al. Single-cell and spatial transcriptomics reveal somitogenesis in gastruloids. *Nature*, 582(7812):405–409, 2020.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL `http://jmlr.org/papers/v9/vandermaaten08a.html`.

Bo Wang, Jiawei Luo, Ying Liu, Wanwan Shi, Zehao Xiong, Cong Shen, and Yahui Long. Spatialmgcn: a novel multi-view graph convolutional network for identifying spatial domains with attention mechanism. *Briefings in Bioinformatics*, 24(5):bbad262, 2023.

Juan Wang, Qi Gao, Shasha Yuan, and Junliang Shang. Cluster-guided contrastive learning with masked autoencoder for spatial domain identification based on spatial transcriptomics. *IEEE Journal of Biomedical and Health Informatics*, 30(2):1809–1820, 2025.

Cameron G Williams, Hyun Jae Lee, Takahiro Asatsuma, Roser Vento-Tormo, and Ashraful Haque. An introduction to spatial transcriptomics for biomedical research. *Genome medicine*, 14(1):68, 2022.

F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19(1):15, 2018.

Hang Xu, Huazhu Fu, Yahui Long, Kok Siong Ang, Raman Sethi, Kelvin Chong, Mengwei Li, Rom Uddamvathanak, Hong Kai Lee, Jingjing Ling, et al. Unsupervised spatially embedded deep representation of spatial transcriptomics. *Genome medicine*, 16(1):12, 2024.

Xuejiao Yu, Yi Jiang, Guoqing Chao, and Dianhui Chu. Deep contrastive multi-view subspace clustering with representation and cluster interactive learning. *IEEE transactions on knowledge and data engineering*, 37(1):188–199, 2024.

Zhuohan Yu, Yifu Lu, Yunhe Wang, Fan Tang, Ka-Chun Wong, and Xiangtao Li. Zinb-based graph embedding autoencoder for single-cell rna-seq interpretations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 4671–4679, 2022.

Zhuohan Yu, Yuning Yang, Xingjian Chen, Ka-Chun Wong, Zhaolei Zhang, Yuming Zhao, and Xiangtao Li. Accurate spatial heterogeneity dissection and gene regulation interpretation for spatial transcriptomics using dual graph contrastive learning. *Advanced Science*, 12(3):2410081, 2025.

Allen W Zhang, Ciara O'Flanagan, Elizabeth A Chavez, Jamie LP Lim, Nicholas Ceglia, Andrew McPherson, Matt Wiens, Pascale Walters, Tim Chan, Brittany Hewitson, et al. Probabilistic cell-type assignment of single-cell rna-seq for tumor microenvironment profiling. *Nature methods*, 16 (10):1007–1015, 2019.

Ruzhang Zhao, Jiuyao Lu, Yuzi Li, Weiqiang Zhou, Ni Zhao, and Hongkai Ji. A systematic evaluation of highly variable gene selection methods for single-cell rna-sequencing. *Genome biology*, 26(1):424, 2025.

Jun Zhu, Yifu Li, Zhenchao Tang, and Cheng Chang. Dusted: Dual-attention enhanced spatial transcriptomics denoiser. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1219–1227, 2025.

Yanran Zhu, Xiao He, Chang Tang, Xinwang Liu, Yuanyuan Liu, and Kunlun He. Multi-view adaptive fusion network for spatially resolved transcriptomics data clustering. *IEEE Transactions on Knowledge and Data Engineering*, 36(12):8889–8900, 2024.

Zeyu Zhu, Ke Liang, Lingyuan Meng, Wanwei Liu, and Xinwang Liu. Stcf: Multi-view clustering for spatial transcriptomics based on cross-view fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

A Appendix

Contents

A.1 Data Processing 14
A.1.1 Data Preprocessing 14
A.1.2 Gridification of ST Data 14
A.2 Graph Construction 14
A.2.1 Cell Graph 14
A.2.2 Gene Co-Expression Graph 15
A.3 Theoretical Proof 15
A.3.1 Global Spatial Interaction 16
A.3.2 Adaptive Spectral Denoising 17
A.4 Algorithm Overview 18
A.5 Detailed Experimental Settings 18
A.5.1 Dataset Details 19
A.5.2 Baseline Details 19
A.5.3 Detailed Parameter Settings 20
A.6 More Details and Analysis of Experimental Results 21
A.6.1 Visualization and Clustering Analysis 21

A.6.2 Statistical Test of Clustering Performance 21
A.6.3 More Details on Ablation Experiments 24
A.6.4 Hyperparameter Sensitivity Analysis 24

A.1 Data Processing

A.1.1 Data Preprocessing

Before applying SpecCG to clustering ST data, we performed standard preprocessing on the original dataset. First, genes detected in fewer than 5 cells were filtered to reduce technical noise. Then, the filtered data were normalized by library size using the SCANPY package (Wolf et al., 2018), and the top 3000 highly variable genes (HVGs) (Zhao et al., 2025) were retained as model input to obtain the expression matrix $X \in \mathbb{R}^{n\times d}$, where n and d are the cell and gene counts, respectively. The normalization function is defined as:

$$X_{ij} = \frac{\tilde{X}_{ij}}{\sum_j \tilde{X}_{ij}} \times 10000, \tag{22}$$

where $\tilde{X}_{ij}$ denotes the raw expression count of cell i on gene j before normalization.

A.1.2 Gridification of ST Data

To support efficient frequency analysis, we first convert the ST data into a regular grid that preserves the relative positions of cells. Let $B = [o_{\min}, o_{\max}] \times [g_{\min}, g_{\max}]$ represent the boundary rectangle of all cell coordinates. B is divided into an $M \times N$ uniform grid ($M \cdot N \geq n$), with each original cell assigned to the nearest grid unit. For grid units without observed cells, KNN interpolation is used to fill in the expression values. Specifically, for data from the $10\times$ Visium platform (Ji et al., 2020), only the cells in odd-numbered rows need to be translated to obtain a regular grid, while ST data from other platforms are gridded using the scaling factor described in Bai et al. (2025). After gridding, the expression of gene k is represented as a 2D spatial map $Z_k \in \mathbb{R}^{M\times N}$, forming a tensor $\boldsymbol{\mathcal{Z}} \in \mathbb{R}^{M\times N\times d}$ for the entire dataset. Simultaneously, a binary mask tensor $\boldsymbol{\mathcal{M}} \in \{0,1\}^{M\times N\times d}$ is constructed: if the grid unit (p, q) corresponds to an original observed cell, $\boldsymbol{\mathcal{M}}_{p,q,k} = 1$; otherwise, it is 0. This mask will be used to extract expression values only at biological measurement locations after frequency domain processing.

The gridding process is crucial: it transforms the irregular spatial sampling of ST technology into a regular grid, while maintaining the spatial continuity of gene expression through KNN interpolation. The interpolated values provide context for global frequency domain operations but are subsequently discarded through masking, ensuring that only true measurements participate in downstream analysis.

A.2 Graph Construction

A.2.1 Cell Graph

Since physically adjacent cells tend to belong to the same functional domain, and transcriptional similarity can connect spatially distant but functionally related cells, two undirected graphs are first constructed to encode this prior. Specifically, the spatial graph $G_s = (V, A_s)$, where the adjacency matrix A_s encodes physical proximity based on Euclidean distance, i.e.,

$$A_s(i,j) = \begin{cases} 1, & \|p_i - p_j\|_2 \leq r, \\ 0, & \text{otherwise}, \end{cases} \tag{23}$$

where p represents the spatial location of the cell, and r represents the distance threshold (e.g., $r = 560\,\mu\text{m}$ for the DLPFC dataset).

The feature graph $G_f = (V, A_f)$, where the adjacency matrix A_f uses k-nearest neighbors based on cosine similarity to capture transcriptional similarity, i.e.,

$$A_f(i,j) = \begin{cases} 1, & j \in \mathcal{N}_k(i) \text{ or } i \in \mathcal{N}_k(j), \\ 0, & \text{otherwise.} \end{cases} \tag{24}$$

where $\mathcal{N}_k(i)$ denotes the indices of the k cells with the highest cosine similarity to cell i. The two graphs share the representation matrix X as initial node attributes, providing a consistent foundation for cross-view representation learning.

A.2.2 Gene Co-Expression Graph

Compared to cell-level representation learning, gene-level co-expression modules (Song et al., 2022) provide inherent biological guidance, offering biologically meaningful contextual information for independent gene representations. To model this complementary information source, a gene co-expression graph $G_g = (V_g, A_g)$ is first constructed, where each node corresponds to a different gene, and edges capture co-expression relationships derived from the data.

We first compute the absolute Pearson correlation coefficient (Dufera et al., 2023) between the expression profiles of genes i and j across all cells:

$$\rho_{ij} = \left| \frac{\sum_{k=1}^{n}(X_{k,i} - \bar{X}_i)(X_{k,j} - \bar{X}_j)}{\sqrt{\sum_{k=1}^{n}(X_{k,i} - \bar{X}_i)^2}\sqrt{\sum_{k=1}^{n}(X_{k,j} - \bar{X}_j)^2}} \right|, \tag{25}$$

where $\bar{X}_i = \frac{1}{n}\sum_{k=1}^{n} X_{k,i}$ is the mean expression of gene i across cells. We retain the k_g most strongly correlated neighbors of each gene when constructing A_g:

$$A_g(i,j) = \begin{cases} 1, & j \in \mathcal{N}_{k_g}(i) \text{ or } i \in \mathcal{N}_{k_g}(j), \\ 0, & \text{otherwise,} \end{cases} \tag{26}$$

where $\mathcal{N}_{k_g}(j)$ denotes the indices of the k_g genes with the highest absolute Pearson correlations with gene j, excluding gene j itself.

A.3 Theoretical Proof

We theoretically analyze the gene frequency-domain enhancement (GFE) module in terms of its capacity for global spatial interaction and adaptive spectral denoising.

Notation. For a single gene, let $Z \in \mathbb{R}^{M\times N}$ denote its gridified expression map, and let $\mathcal{G} = \{0,\dots,M-1\} \times \{0,\dots,N-1\}$. The two-dimensional discrete Fourier transform (2D-DFT) and its inverse are

$$\widehat{Z}[u,v] = \mathcal{F}(Z)[u,v] = \sum_{(p,q)\in\mathcal{G}} Z[p,q]\, e^{-2\pi \mathrm{i}\left(\frac{up}{M}+\frac{vq}{N}\right)}, \tag{27}$$

$$Z[p,q] = \mathcal{F}^{-1}(\widehat{Z})[p,q] = \frac{1}{MN}\sum_{(u,v)\in\mathcal{G}} \widehat{Z}[u,v]\, e^{2\pi \mathrm{i}\left(\frac{up}{M}+\frac{vq}{N}\right)}. \tag{28}$$

Here, $\mathrm{i}^2 = -1$. We use $\overline{z}$ for complex conjugation, $|z|$ for the complex modulus, and $\|\cdot\|_F$ for the Frobenius norm.

GFE operates on the stored half-spectrum using element-wise modulation followed by a real inverse 2D-DFT. For analysis, we use its equivalent full-spectrum representation,

$$\mathcal{E}_W(Z) = \mathcal{F}^{-1}\big(W \odot \mathcal{F}(Z)\big), \tag{29}$$

where $\odot$ denotes element-wise multiplication and $W \in \mathbb{C}^{M\times N}$ is the effective frequency response, accounting for the conjugate-pair handling of the real inverse transform. It satisfies

$$W[(-u) \bmod M, (-v) \bmod N] = \overline{W[u,v]}. \tag{30}$$

Any full response satisfying equation 30 can be realized by choosing its stored half as the modulation weights.

A.3.1 GLOBAL SPATIAL INTERACTION

Lemma 1. For $f, g \in \mathbb{C}^{M\times N}$, define

$$(f \circledast g)[p,q] = \sum_{(a,b)\in\mathcal{G}} f[a,b]\, g[(p-a) \bmod M, (q-b) \bmod N]. \tag{31}$$

Then

$$\mathcal{F}(f \circledast g) = \mathcal{F}(f) \odot \mathcal{F}(g). \tag{32}$$

Proof. Substituting equation 31 into the DFT and exchanging the finite sums yields

$$\begin{aligned}&\mathcal{F}(f \circledast g)[u,v]\\ &\quad= \sum_{(a,b)\in\mathcal{G}} f[a,b] \sum_{(p,q)\in\mathcal{G}} g[(p-a) \bmod M, (q-b) \bmod N]\, e^{-2\pi\mathrm{i}\left(\frac{up}{M}+\frac{vq}{N}\right)}.\end{aligned} \tag{33}$$

For fixed (a,b), reindex the inner sum by $r = (p-a) \bmod M$ and $t = (q-b) \bmod N$. This is a bijection of $\mathcal{G}$, and the periodicity of the Fourier exponential gives

$$\begin{aligned}&\sum_{(p,q)\in\mathcal{G}} g[(p-a) \bmod M, (q-b) \bmod N]\, e^{-2\pi\mathrm{i}\left(\frac{up}{M}+\frac{vq}{N}\right)}\\ &\quad= e^{-2\pi\mathrm{i}\left(\frac{ua}{M}+\frac{vb}{N}\right)} \sum_{(r,t)\in\mathcal{G}} g[r,t]\, e^{-2\pi\mathrm{i}\left(\frac{ur}{M}+\frac{vt}{N}\right)}\\ &\quad= e^{-2\pi\mathrm{i}\left(\frac{ua}{M}+\frac{vb}{N}\right)} \mathcal{F}(g)[u,v].\end{aligned} \tag{34}$$

Substituting into equation 33, we obtain

$$\begin{aligned}\mathcal{F}(f \circledast g)[u,v] &= \left(\sum_{(a,b)\in\mathcal{G}} f[a,b]\, e^{-2\pi\mathrm{i}\left(\frac{ua}{M}+\frac{vb}{N}\right)}\right) \mathcal{F}(g)[u,v]\\ &= \mathcal{F}(f)[u,v]\, \mathcal{F}(g)[u,v],\end{aligned} \tag{35}$$

which proves the identity. □

Theorem 1. Let $K = \mathcal{F}^{-1}(W)$. Then K is real-valued and

$$\mathcal{E}_W(Z) = K \circledast Z \qquad \text{for every } Z \in \mathbb{R}^{M\times N}. \tag{36}$$

Equivalently,

$$\mathcal{E}_W(Z)[p,q] = \sum_{(a,b)\in\mathcal{G}} K[(p-a) \bmod M, (q-b) \bmod N]\, Z[a,b]. \tag{37}$$

The receptive field is determined by the support of K. In particular, a full-support kernel gives each output a global receptive field in a single application.

Proof. The symmetry in equation 30 ensures that $K = \mathcal{F}^{-1}(W)$ is real-valued: conjugate frequency terms pair in the inverse DFT, while self-conjugate terms are real. Since $\mathcal{F}(K) = W$, Lemma 1 gives

$$\mathcal{F}(K \circledast Z) = \mathcal{F}(K) \odot \mathcal{F}(Z) = W \odot \mathcal{F}(Z). \tag{38}$$

Applying the inverse transform,

$$K \circledast Z = \mathcal{F}^{-1}\big(W \odot \mathcal{F}(Z)\big) = \mathcal{E}_W(Z). \tag{39}$$

Reindexing the circular convolution so that (a,b) indexes the input position gives equation 37.

For fixed weights, the dependence on an input position is characterized by

$$\frac{\partial \mathcal{E}_W(Z)[p,q]}{\partial Z[a,b]} = K[(p-a) \bmod M, (q-b) \bmod N]. \tag{40}$$

Thus every input influences every output when K is nonzero at all spatial offsets. Such kernels are representable by the spectral parameterization: any real full-support kernel has a conjugate-symmetric DFT, whose stored half can be used as the modulation weights. This establishes the capacity for global spatial interaction in one application. □

The above result establishes that GFE is equivalent to spatial circular convolution and can achieve a global receptive field through a full-support kernel, enabling tissue-wide spatial interaction in a single operation.

A.3.2 Adaptive Spectral Denoising

Consider the observation model $Z = S + \varepsilon$, where $S \in \mathbb{R}^{M\times N}$ is a deterministic biological signal and $\varepsilon \in \mathbb{R}^{M\times N}$ is zero-mean, spatially uncorrelated noise with variance $\sigma^2 > 0$:

$$\mathbb{E}[\varepsilon[p,q]] = 0, \qquad \mathbb{E}[\varepsilon[p,q]\varepsilon[a,b]] = \sigma^2\delta_{p,a}\delta_{q,b}. \tag{41}$$

Here, δ denotes the Kronecker symbol. Expectations are taken over ε.

Theorem 2. Under equation 41, the expected reconstruction error

$$\mathcal{R}(W) = \mathbb{E}\Big[\|\mathcal{E}_W(S+\varepsilon) - S\|_F^2\Big] \tag{42}$$

has a unique minimizer over deterministic effective full-spectrum responses satisfying equation 30:

$$W^\star[u,v] = \frac{|\widehat{S}[u,v]|^2}{|\widehat{S}[u,v]|^2 + MN\sigma^2}, \tag{43}$$

where $\widehat{S} = \mathcal{F}(S)$. This response can be realized by half-spectrum modulation and strictly reduces the expected reconstruction error relative to the identity filter $W \equiv 1$.

Proof. Let $\widehat{\varepsilon} = \mathcal{F}(\varepsilon)$. Linearity gives $\mathbb{E}[\widehat{\varepsilon}[u,v]] = 0$. Expanding its squared modulus and applying equation 41 yields

$$\begin{aligned}
&\mathbb{E}\big[|\widehat{\varepsilon}[u,v]|^2\big] \\
&\quad= \sum_{(p,q)\in\mathcal{G}} \sum_{(a,b)\in\mathcal{G}} \mathbb{E}[\varepsilon[p,q]\varepsilon[a,b]]\, e^{-2\pi\mathrm{i}\left(\frac{u(p-a)}{M} + \frac{v(q-b)}{N}\right)} \\
&\quad= \sigma^2 \sum_{(p,q)\in\mathcal{G}} 1 = MN\sigma^2.
\end{aligned} \tag{44}$$

Thus the expected noise power is constant across frequencies. By Parseval's identity,

$$\begin{aligned}
\mathcal{R}(W) &= \frac{1}{MN}\mathbb{E}\left[\left\|W \odot (\widehat{S} + \widehat{\varepsilon}) - \widehat{S}\right\|_F^2\right] \\
&= \frac{1}{MN} \sum_{(u,v)\in\mathcal{G}} \mathbb{E}\left[\left|(W[u,v]-1)\widehat{S}[u,v] + W[u,v]\widehat{\varepsilon}[u,v]\right|^2\right].
\end{aligned} \tag{45}$$

Fix a frequency and write $w = W[u,v]$, $s = \widehat{S}[u,v]$, $\eta = \widehat{\varepsilon}[u,v]$, $P = |s|^2$, and $Q = MN\sigma^2 > 0$. Since $\mathbb{E}[\eta] = 0$, the cross term in the expected squared error vanishes:

$$\begin{aligned}
J(w) &:= \mathbb{E}\big[|(w-1)s + w\eta|^2\big] \\
&= P|w-1|^2 + Q|w|^2 \\
&= (P+Q)|w|^2 - 2P\,\mathrm{Re}(w) + P \\
&= (P+Q)\left|w - \frac{P}{P+Q}\right|^2 + \frac{PQ}{P+Q}.
\end{aligned} \tag{46}$$

Because $P+Q > 0$, the unique minimizer is $w^\star = P/(P+Q) \in [0,1)$. Applying this result at every frequency gives equation 43. Since S is real, its power spectrum is equal at conjugate frequency pairs; the real-valued optimal response therefore satisfies equation 30. It is consequently the unique optimal effective response and can be realized by choosing its stored half as the modulation weights.

Finally, equation 46 gives $J(w^\star) = PQ/(P+Q)$, whereas the identity filter has error $J(1) = Q$. Hence

$$J(1) - J(w^\star) = \frac{Q^2}{P+Q} > 0. \tag{47}$$

Algorithm 1: **Joint Cell–Gene Representation Learning with SpecCG**

Input: $\mathcal{A}$: ST AnnData object with raw gene counts and spatial coordinates
d: number of selected genes; K: number of spatial domains
r, k, k_g: graph construction parameters; τ: contrastive temperature
β, γ: loss hyperparameters; T: number of training epochs
Output: H^c: consistent cell embeddings; $\mathbf{y}$: spatial domain labels

Preprocess $\mathcal{A}$ to obtain $X \in \mathbb{R}^{n \times d}$ for n spots;
Construct spatial graph G_s, feature graph G_f, and gene co-expression graph G_g;
Gridify X to obtain spatial gene expression maps $\boldsymbol{\mathcal{Z}}$ and observation mask $\boldsymbol{\mathcal{M}}$;
$\hat{\boldsymbol{\mathcal{Z}}} \leftarrow \mathcal{F}_{2D}[\boldsymbol{\mathcal{Z}}]$;
Initialize spectral modulation weights W and all network parameters;
for $t \leftarrow 1$ **to** T **do**
  `// Gene frequency-domain enhancement`
  $\boldsymbol{\mathcal{Z}}^{\mathrm{enh}} \leftarrow \mathcal{F}_{2D}^{-1}[\boldsymbol{\mathcal{W}} \odot \hat{\boldsymbol{\mathcal{Z}}}]$, with $\boldsymbol{\mathcal{W}}$ obtained by broadcasting W across genes;
  $X^{\mathrm{enh}} \leftarrow \mathrm{Mask}(\boldsymbol{\mathcal{Z}}^{\mathrm{enh}}, \boldsymbol{\mathcal{M}})$; `// Extract observed spots`
  `// Multi-view cell representation learning`
  Encode X and X^{enh} over G_s and G_f into $(S, F), (\hat{S}, \hat{F})$ using equation 6–equation 7;
  Fuse (S, F) and $(\hat{S}, \hat{F})$ by attention to obtain H and $\hat{H}$ using equation 8–equation 9;
  $H^c \leftarrow \mathrm{MLP}_h([H \| \hat{H}])$;
  Compute spatially aware contrastive loss $\mathcal{L}_{\mathrm{CL}}$ using equation 12;
  `// Gene-cell joint decoding`
  Encode $(X^{\mathrm{enh}})^{\mathrm{T}}$ over G_g to obtain gene embeddings G using equation 13;
  $\Omega \leftarrow \mathrm{softplus}(H^c G^{\mathrm{T}})$; `// Joint reconstruction mean`
  Decode Π, Θ from H^c and compute $\mathcal{L}_{\mathrm{GCJD}}$ using equation 16–equation 17;
  `// Clustering guidance and joint optimization`
  Obtain K cluster centers from H^c using k-means;
  Compute soft assignments $Q^c, Q, \hat{Q}$ and target P using equation 18–equation 19;
  Compute clustering loss $\mathcal{L}_{\mathrm{KL}}$ using equation 20;
  $\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{GCJD}} + \beta \mathcal{L}_{\mathrm{CL}} + \gamma \mathcal{L}_{\mathrm{KL}}$;
  Update W and all network parameters by backpropagation with Adam;
end
Recompute H^c using the trained model;
$\mathbf{y} \leftarrow \mathrm{KMeans}(H^c, K)$; `// Identify spatial domains`
return $H^c, \mathbf{y}$;

Summing over frequencies in equation 45 establishes

$$\mathcal{R}(W^\star) < \mathcal{R}(\mathbf{1}) = \mathbb{E}\|\varepsilon\|_F^2 = MN\sigma^2, \tag{48}$$

where $\mathbf{1}$ denotes the all-ones full-spectrum response. □

Under the stated signal-plus-noise model, GFE can represent a signal-adaptive spectral response that reduces the expected reconstruction error, providing a theoretical basis for enhancing informative expression patterns obscured by noise.

A.4 Algorithm Overview

This section details the algorithmic procedure of the proposed SpecCG, as shown in Algorithm 1. The source code for SpecCG is provided in the supplementary material.

A.5 Detailed Experimental Settings

Table 3: The statistics of the datasets.

Datasets	Spots	Genes	Domains	Platforms
151510	4595	33538	7	10x Visium
151669	3636	33538	5	10x Visium
151670	3484	33538	5	10x Visium
151671	4093	33538	5	10x Visium
151674	3635	33538	7	10x Visium
HBC	3798	36601	20	10x Visium
MBA	2695	32285	52	10x Visium
ME9.5	5913	25568	12	Stereo-Seq

A.5.1 Dataset Details

Eight public ST benchmark datasets are used to evaluate the proposed method. These include the LIBD human dorsolateral prefrontal cortex (DLPFC) (Huuki-Myers et al., 2024), human breast cancer (HBC) (Wang et al., 2025), and mouse brain anterior tissue (MBA) (Zhu et al., 2024) datasets from the 10× Genomics Visium platform (Ji et al., 2020), as well as the mouse embryo (ME9.5) dataset (Chen et al., 2022) from the Stereo-seq platform (Long et al., 2024). The DLPFC dataset contains 12 slices, each manually annotated into five or seven spatial domains, including cortical layers and white matter. These contain 3,460–4,789 spots and 33,538 genes per slice. Five of these slices (151510, 151669, 151670, 151671, and 151674) are selected for evaluation. Among them, slices 151510 and 151674 are annotated into 7 cortical layers, while the remaining three are annotated into 5 cortical layers. The HBC dataset comprises 3,798 spots, each capturing 36,601 genes. Pathologists manually annotated it into four major regions and 20 subregions in total. These include ductal carcinoma in situ/lobular carcinoma in situ (DCIS/LCIS), invasive ductal carcinoma (IDC), healthy regions (Healthy), and tumor edge. The MBA dataset is annotated into 52 spatial domains according to the Allen Mouse Brain Reference Atlas (Kronman et al., 2024), including an unlabeled region designated as Not_annotation (Not_anno). The ME9.5 dataset contains mouse embryo samples at the E9.5 developmental stage and is annotated into 12 tissue regions. Detailed information on the above datasets is provided in Table 3.

A.5.2 Baseline Details

To comprehensively evaluate the performance advantages of the proposed SpecCG method, we compare it against ten representative state-of-the-art baselines for spatial transcriptomics analysis: GraphST, Spatial-MGCN, SEDR, MAFN, DUSTED, SpaSEG, stDCL, SAGE, STMCCL, and STCF. These methods integrate gene expression with spatial information to facilitate spatial domain identification. Detailed descriptions of these methods are provided below:

- **GraphST** (Long et al., 2023): GraphST combines a GCN encoder with self-supervised contrastive learning to derive discriminative feature representations. Its contrastive objective aims to bring spatially neighboring positive spots within a local context closer together while preserving the original spatial topology, and to push non-adjacent negative spots apart. GraphST further introduces an expression reconstruction objective to learn spatially aware features for subsequent spatial clustering.
- **Spatial-MGCN** (Wang et al., 2023): By constructing gene expression and spatial neighborhood graphs, Spatial-MGCN simultaneously captures transcriptional similarity and spatial local topology. It utilizes a multi-view graph convolutional network with an attention mechanism to simultaneously extract view-specific representations and cross-view shared representations. Furthermore, Spatial-MGCN introduces the Zero-Inflated Negative Binomial (ZINB) distribution and spatial regularization to reconstruct gene expression distributions and maintain coherent domain distributions.
- **SEDR** (Xu et al., 2024): SEDR combines deep expression autoencoder with variational graph autoencoder to learn low-dimensional embeddings of gene expression while preserving spatial information. The method recovers masked expression profiles through masked

self-supervised learning and reconstructs transcriptional expression by concatenating expression embeddings with spatial embeddings, effectively supporting tasks such as spatial domain identification and gene imputation.

- **MAFN** (Zhu et al., 2024): MAFN employs a multi-view adaptive fusion architecture to integrate spatial and transcriptomic information. This method extracts view-specific features using two GCN encoders and filters out irrelevant information by leveraging a cross-view correlation reduction strategy during complementary feature learning. Subsequently, a cross-view attention module is introduced to adaptively fuse multi-view features, enabling precise spatial domain identification.
- **DUSTED** (Zhu et al., 2025): DUSTED is a spatial expression denoising framework based on graph autoencoders. It combines gene channel attention with graph attention to account for gene-specific noise variations and interrelationships among genes, while adaptively aggregating information from spatial neighborhoods. Furthermore, it employs a negative binomial distribution or ZINB likelihood to estimate gene expression distribution parameters, providing high-quality expression profiles for downstream analyses including clustering.
- **SpaSEG** (Bai et al., 2025): SpaSEG is an unsupervised deep learning framework based on convolutional neural networks. It arranges dimensionality-reduced expression features into image-like tensors according to spatial coordinate information, thereby modeling spatial domain identification as an unsupervised segmentation problem. The method first pretrains a convolutional neural network with reconstruction objective, and further optimizes it using pseudo-label segmentation loss and edge strength penalty, jointly modeling transcriptional features and local spatial coherence.
- **stDCL** (Yu et al., 2025): stDCL integrates gene expression profiles and spatial graph structures using graph convolutional autoencoders, and reconstructs gene expression. The method combines dual contrastive learning. On one hand, it promotes the consistency of representations of spatially adjacent spots and aligns the corresponding spot features across views. On the other hand, it reduces the correlation between different features to improve the feature representation used for spatial clustering.
- **SAGE** (He et al., 2026): SAGE combines topic modeling based on non-negative matrix factorization (NMF) with importance scores derived from classifiers to select spatially informative genes. Subsequently, through consensus refinement and graph contrastive representation learning, it integrates local expression graph with topic-driven non-local graph. These two complementary views jointly capture local tissue organization and long-range functional relationships for spatial domain identification.
- **STMCCL** (Wang et al., 2025): STMCCL jointly learns feature representations through masked auto-encoding and cluster-guided contrastive learning. It constructs augmented views and generates a feature view using a masked encoder to learn view-specific embeddings. Its multi-clustering perspective module considers both geometric and structural relationships between clusters, while sample confidence is estimated from pre-clustering to guide the construction of positive and negative samples for contrastive learning.
- **STCF** (Zhu et al., 2026): STCF treats highly variable genes and lowly variable genes as two complementary expression views for spatial clustering. It integrates information from both views through a plug-and-play cross-view fusion strategy and introduces a reverse scaling cosine error (R-SCE) loss to balance embedding alignment and separation. This design leverages information from gene subsets with different variability levels to support spatial domain identification.

A.5.3 Detailed Parameter Settings

For graph construction, we set the distance threshold r for the spatial graph, and the neighborhood sizes k and k_g for the feature graph and gene co-expression graph, respectively. For the ME9.5 dataset from the Stereo-Seq platform, we instead construct the spatial graph by selecting the k nearest spatial neighbors of each spot based on Euclidean distances between spatial coordinates. These graph parameters are fixed during training. A grid search is performed over the contrastive temperature parameter τ and the loss hyperparameters β and γ, where β and γ control the contributions of contrastive learning and clustering guidance, respectively. The candidate values for both β and γ are

Table 4: Detailed parameter settings for each dataset.

Datasets	r (μm)	k	k_g	τ	β	γ
151510	560	14	14	0.9	10	0.1
151669	560	14	14	0.9	0.01	100
151670	560	14	14	0.9	0.001	10
151671	560	14	14	0.5	0.001	10
151674	560	14	14	1	0.1	100
HBC	560	15	14	0.2	0.01	0.1
MBA	560	15	14	0.8	0.1	0.001
ME9.5	-	15	14	0.9	0.1	10

$\{0.001, 0.01, 0.1, 1, 10, 100\}$, and τ is searched over $\{0.1, 0.2, \ldots, 1.0\}$. Table 4 reports the graph construction parameters and optimal hyperparameters.

A.6 More Details and Analysis of Experimental Results

A.6.1 Visualization and Clustering Analysis

The clustering visualizations in Figures 5 and 6 demonstrate that the proposed SpecCG produces domain partitions that better align with the pathological gold standard, whereas other baselines result in fragmented domain boundaries or significantly deviate from manual annotations. Figure 7 visualizes the low-dimensional embeddings learned by six different methods across four datasets. It is evident that the embeddings learned by SpecCG correspond to the clustering results, with both aspects highlighting SpecCG's performance superiority over the other baselines.

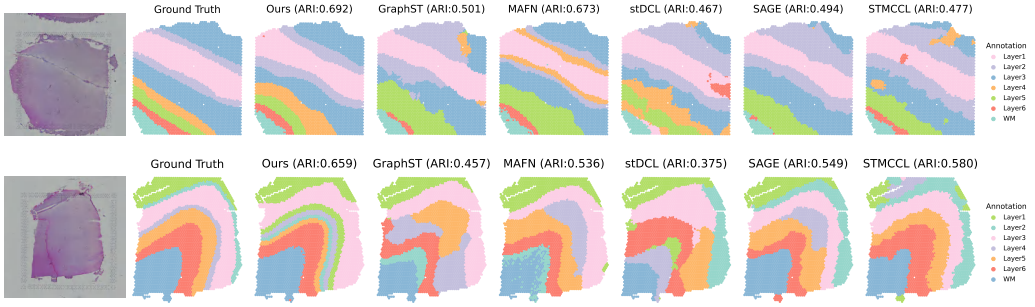

Figure 5: Visualization of spatial clustering across datasets, including histopathological images, manual annotations, and identified spatial domains. From top to bottom, the rows correspond to the 151510 and 151674 datasets.

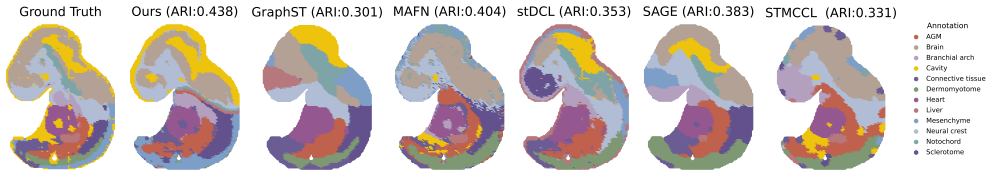

Figure 6: Visualization of manual annotations and spatial clustering on the ME9.5 dataset.

A.6.2 Statistical Test of Clustering Performance

To evaluate the statistical significance of the clustering performance improvements achieved by SpecCG, we employed the Friedman test and the Wilcoxon signed-rank test to analyze the ARI, NMI, ACC, and F1 metrics across eight datasets. First, the Friedman test is used to assess whether

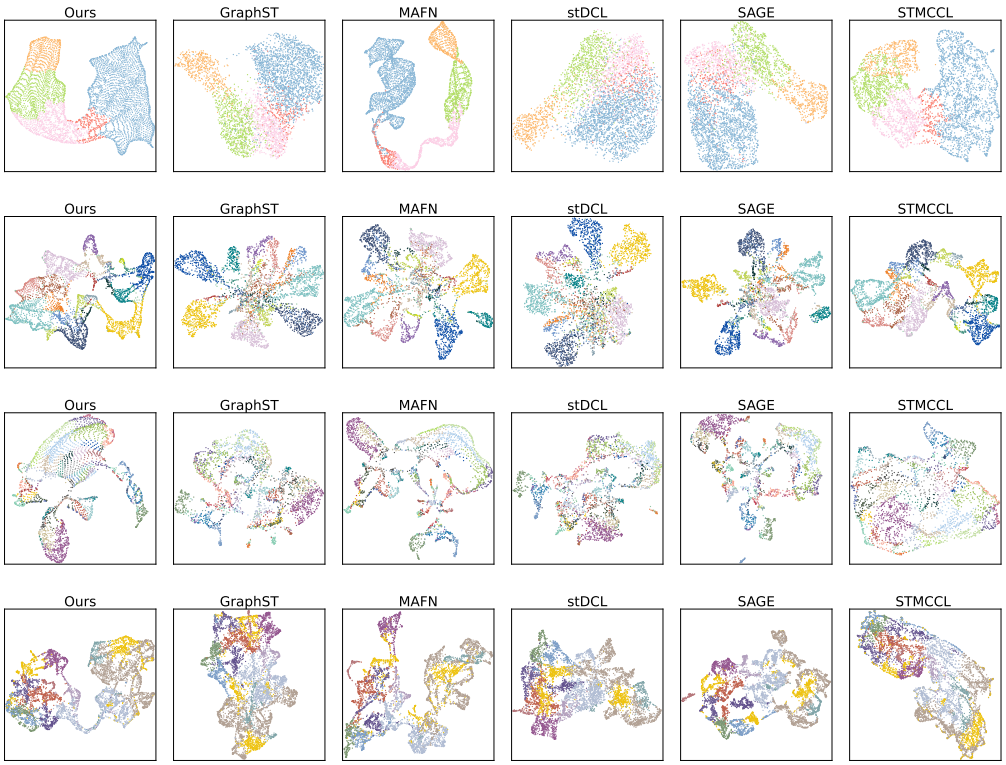

Figure 7: Visualization of UMAP embeddings across datasets. From top to bottom, the rows correspond to the 151671, HBC, MBA and ME9.5 datasets.

there is an overall performance difference among different methods, and the average ranks of each method across the eight datasets are reported, where a smaller rank indicates better performance. SpecCG achieved average ranks of 1.000 (ARI), 1.125 (NMI), 1.000 (ACC), and 1.875 (F1) across the four metrics, the average ranking results derived from the Friedman test are shown in Table 5. Furthermore, we utilized the Wilcoxon signed-rank test to conduct pairwise comparisons between SpecCG and each baseline on the same datasets, the results are shown in Tables 6–9. Based on the performance differences between SpecCG and the baselines, R^+ and R^- represent the rank sums corresponding to positive and negative differences, respectively. A larger R^+ indicates performance improvement (i.e., the results favor SpecCG's performance advantage), whereas a larger R^- indicates performance degradation. Additionally, a smaller p-value signifies greater statistical significance. The test results demonstrate that, at a significance level of $p < 0.05$, SpecCG significantly outperforms all baselines on ARI, NMI, and ACC, and significantly outperforms all baselines except SAGE on F1, providing statistical support for its superior clustering performance.

Table 5: Average ranks of different methods on four metrics based on the Friedman test. Smaller rank indicates better performance.

Metric	SpecCG	MAFN	STMCCL	STCF	Spatial-MGCN	SAGE	GraphST	stDCL	DUSTED	SpaSEG	SEDR
ARI	**1.000**	3.625	4.250	4.500	4.625	6.125	7.375	7.875	8.125	8.250	10.250
NMI	**1.125**	4.375	6.250	6.750	5.875	4.500	7.000	5.750	6.500	9.125	8.750
ACC	**1.000**	2.875	3.375	5.000	6.625	5.625	7.625	7.125	7.750	8.750	10.250
F1	**1.875**	3.250	5.500	7.500	5.875	5.250	5.875	7.625	6.750	9.000	7.500

Table 6: Wilcoxon signed-rank test results (ARI).

Comparison	R^+	R^-	p-value*
SpecCG vs. GraphST	36	0	0.00781
SpecCG vs. Spatial-MGCN	36	0	0.00781
SpecCG vs. SEDR	36	0	0.00781
SpecCG vs. MAFN	36	0	0.00781
SpecCG vs. DUSTED	36	0	0.00781
SpecCG vs. SpaSEG	36	0	0.00781
SpecCG vs. stDCL	36	0	0.00781
SpecCG vs. SAGE	36	0	0.00781
SpecCG vs. STMCCL	36	0	0.00781
SpecCG vs. STCF	36	0	0.00781

Table 7: Wilcoxon signed-rank test results (NMI).

Comparison	R^+	R^-	p-value*
SpecCG vs. GraphST	36	0	0.00781
SpecCG vs. Spatial-MGCN	36	0	0.00781
SpecCG vs. SEDR	36	0	0.00781
SpecCG vs. MAFN	36	0	0.00781
SpecCG vs. DUSTED	36	0	0.00781
SpecCG vs. SpaSEG	36	0	0.00781
SpecCG vs. stDCL	35	1	0.01563
SpecCG vs. SAGE	36	0	0.00781
SpecCG vs. STMCCL	36	0	0.00781
SpecCG vs. STCF	36	0	0.00781

Table 8: Wilcoxon signed-rank test results (ACC).

Comparison	R^+	R^-	p-value*
SpecCG vs. GraphST	36	0	0.00781
SpecCG vs. Spatial-MGCN	36	0	0.00781
SpecCG vs. SEDR	36	0	0.00781
SpecCG vs. MAFN	36	0	0.00781
SpecCG vs. DUSTED	36	0	0.00781
SpecCG vs. SpaSEG	36	0	0.00781
SpecCG vs. stDCL	36	0	0.00781
SpecCG vs. SAGE	36	0	0.00781
SpecCG vs. STMCCL	36	0	0.00781
SpecCG vs. STCF	36	0	0.00781

Table 9: Wilcoxon signed-rank test results (F1).

Comparison	R^+	R^-	p-value*
SpecCG vs. GraphST	36	0	0.00781
SpecCG vs. Spatial-MGCN	35	1	0.01563
SpecCG vs. SEDR	36	0	0.00781
SpecCG vs. MAFN	33	3	0.03906
SpecCG vs. DUSTED	36	0	0.00781
SpecCG vs. SpaSEG	36	0	0.00781
SpecCG vs. stDCL	35	1	0.01563
SpecCG vs. SAGE	32	4	0.05469
SpecCG vs. STMCCL	34	2	0.02344
SpecCG vs. STCF	36	0	0.00781

A.6.3 More Details on Ablation Experiments

To intuitively demonstrate the effectiveness of the GFE module, we visualized the low-dimensional embeddings generated by SpecCG and its variant w/o GFE, as shown in Figure 8. It is evident that GFE facilitates more coherent embeddings with clearer boundaries, particularly on the 151671 dataset, as it enhances beneficial global patterns to guide the learning of more discriminative feature representations.

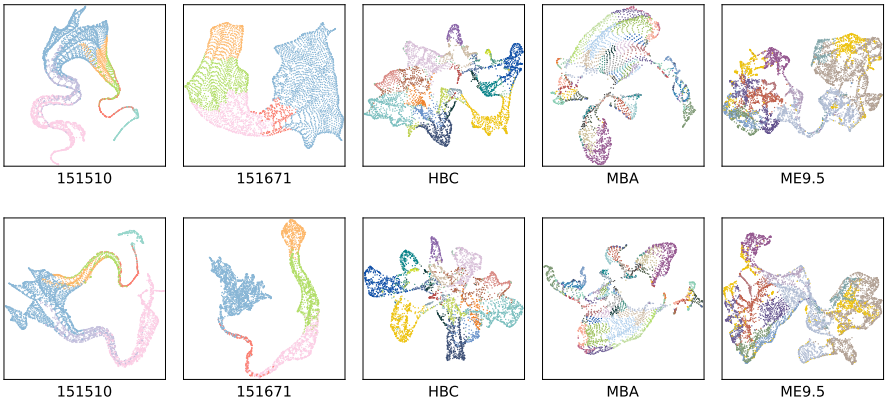

Figure 8: Visualizations of UMAP embeddings learned by SpecCG and its variant w/o GFE on different datasets. From top to bottom, the rows correspond to SpecCG and its variant w/o GFE.

A.6.4 Hyperparameter Sensitivity Analysis

This section analyzes the impact of the parameter τ on clustering performance across all datasets. As shown in Figure 9, even minor changes in τ resulted in significant fluctuations in overall performance for some datasets, indicating the need to identify optimal parameter settings for different datasets, specific parameter search results are presented in Table 4.

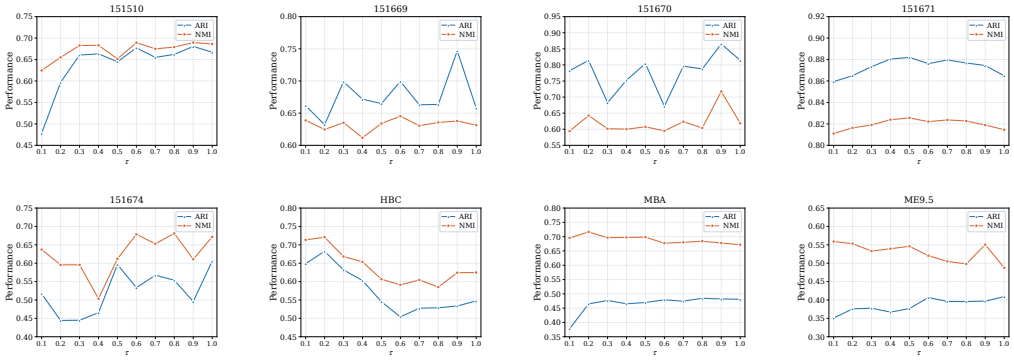

Figure 9: Sensitivity analysis of the contrastive temperature τ on eight ST datasets.