Learning to Cluster Rare Cell Types: Implicit Semantic Data Augmentation for Spatial Multi-modal Omics Analysis

Daixian Liu[1*], Hau-Sing So[2*], Haoran Chen[3], Jiao Li[4], Shanshan Wang[5], Mengzhu Wang[6†], Jingcai Guo[7]

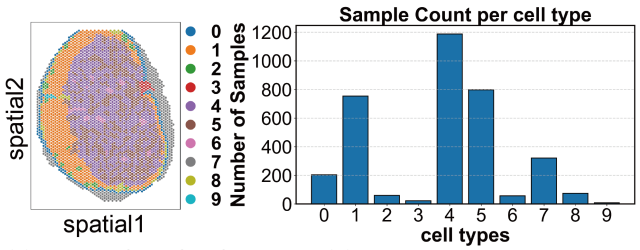
[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]University of Macau
[3]Sichuan Agricultural University
[4]University of Electronic Science and Technology of China
[5]Anhui University
[6]Hebei University of Technology
[7]The Hong Kong Polytechnic University
daixian2023@gmail.com, superhausing@gmail.com, 202308611@stu.sicau.edu.cn, jiaoli.research@outlook.com, wang.shanshan@ahu.edu.cn, dreamkily@gmail.com, jc-jingcai.guo@polyu.edu.hk

Abstract

Spatial multi-modal omics technologies have transformed biological research by enabling the simultaneous profiling of gene expression, protein abundance, and chromatin accessibility within their native spatial contexts. Despite these advances, accurately clustering rare cell types remains a major challenge due to data sparsity, high dimensionality, and limited annotated samples. While Graph Neural Networks (GNNs) have shown potential in modeling spatial omics data, their effectiveness is often constrained by the use of fixed K-nearest neighbor (KNN) graph structures, which fail to capture latent semantic relationships masked by sequencing noise. To overcome these limitations, we propose CRCT (Clustering Rare Cell Types): a novel framework that combines Implicit Semantic Data Augmentation (ISDA) with adaptive graph learning for spatial multi-modal omics analysis. Unlike traditional augmentation strategies that generate explicit synthetic samples, CRCT operates in the deep feature space by dynamically estimating intra-class covariance matrices and implicitly perturbing features along semantically meaningful directions. This enables effective augmentation for rare cell populations while preserving biological fidelity. Extensive experiments across four real-world datasets (HLN, MB, Stereo-CITE-seq, and SPOTS) and one synthetic benchmark demonstrate the state-of-the-art performance of CRCT, achieving improvements of up to +1.7 NMI and +7.8 ARI over strong baseline methods.

Introduction

Spatial transcriptomics has emerged as a transformative advancement in biological research, building on the success of single-cell RNA sequencing. By preserving the two-dimensional spatial coordinates of each captured transcript, it enables direct investigation into how gene expression programs are spatially organized within tissue architecture and how neighboring cells influence each other's states (Rao et al. 2021; Williams et al. 2022; Moses and Pachter 2022; Tian, Chen, and Macosko 2023). The field is now rapidly evolving toward spatial multi-omics, where multiple molecular modalities—such as RNA expression, chromatin accessibility, histone modifications, and surface protein abundance—are simultaneously profiled within the same tissue section using technologies including RNA-seq (Conesa et al. 2016), ATAC-seq (Chen et al. 2016), CUT&Tag-RNA-seq (Zhang et al. 2023), ADT-seq (Gravis et al. 2016), and Stereo-CITE-seq (Liao et al. 2023). These complementary layers of molecular information offer a holistic view of cellular identity, regulatory mechanisms, and cell–cell interactions in situ, holding the potential to revolutionize our understanding of development, immune responses, and disease pathogenesis.



(a) HLN Visualization **(b) HLN cell type counts**

Figure 1: (a) Spatial transcriptomic map of the Human Lymph Node (HLN) dataset. (b) Bar chart summarizing the number of cells per cell type in the HLN dataset.

Integrating such heterogeneous signals is intrinsically challenging. Each modality possesses distinct dimensionalities (for example, antibody-derived tag profiles are orders of magnitude shorter than RNA counts), modality-specific technical noise, and batch effects introduced by the diverse chemistries. Spatial alignment further compounds these difficulties because the data must be fused at matched coordinates without destroying local neighbourhood structure. Directly transplanting advances from natural language pro-

*These authors contributed equally.
†Corresponding author.

cessing or computer vision (Wang et al. 2025a), which have recently benefited from large and homogeneous training corpora, often fails because biological measurements violate the assumption of independent and identically distributed samples and exhibit complex data-generation biases.

Despite these challenges, current computational strategies for joint modeling of spatially-resolved multimodal data remain inadequate. The majority of existing approaches are constrained to either analyzing individual modalities independently (e.g., MOFA+ (Argelaguet et al. 2020), MultiVI (Ashuach et al. 2023), TotalVI (Gayoso et al. 2021)) or operating without spatial context (e.g., CiteFuse (Kim et al. 2020), STAGATE (Dong and Zhang 2022), PAST (Li et al. 2023)). Recently, emerging methodologies have begun to address this gap by explicitly incorporating spatial information with multi-omic data. Notable examples include SpatialGlue (Long et al. 2024), which implements graph neural networks with dual-attention mechanisms for modality integration, and PRAGA (Huang et al. 2025), which advances the field through dynamic graph architectures coupled with a Bayesian Gaussian mixture model for enhanced cross-modal denoising and representation learning.

However, two critical challenges persist in practical applications: (1) the scarcity of spatial multi-omics datasets due to prohibitive experimental costs limits the scale required for learning generalizable representations; and (2) the inherent long-tailed distribution of cellular composition in most tissues. As illustrated in Figure 1, dominant cell types disproportionately influence model training, while rare cell populations and boundary regions remain systematically underrepresented. This imbalance leads to three key limitations: models tend to overfit to prevalent classes, exhibit blurred decision boundaries, and critically fail to identify biologically important rare populations (Cui et al. 2020; Su et al. 2025). Compounding these issues, unlike computer vision data where geometric transformations can effectively augment datasets (Wang 2025), omics data lacks robust augmentation strategies, leaving few viable solutions to address these fundamental imbalances.

Building upon transfer-learning principles (Wang et al. 2021), we present CRCT (*Clustering Rare Cell Types*), a novel framework that pioneers implicit semantic augmentation for spatial multi-omics integration. Operating in a fully unsupervised paradigm, CRCT implements a three-phase iterative process: (1) generating pseudo-labels through initial clustering to approximate cellular heterogeneity; (2) computing class prototypes as centroids in the latent feature space; and (3) performing stochastic prototype-based augmentation during subsequent training epochs. Crucially, augmented samples preserve both cellular identity and multi-omic characteristics while expanding the representation of rare populations. This implicit data expansion strategy effectively addresses three fundamental challenges: mitigating overfitting to dominant cell types, refining cluster boundaries, and enhancing detection sensitivity for biologically significant rare populations. The main contributions of this paper are as follows:

- CRCT introduces the novel adaptation of semantic augmentation to spatial multi-omics data, enabling effective data expansion without requiring additional experimental samples or explicit geometric transformations.
- Our framework autonomously identifies and amplifies underrepresented cell populations through iterative prototype refinement and stochastic augmentation, overcoming the limitations of long-tailed cellular distributions without requiring predefined cell type labels.
- Comprehensive evaluations across multiple benchmarking datasets confirm CRCT's consistent superiority over existing state-of-the-art methods.

Related Work

Multi-modal Omics Aggregation

Multi-modal omics aggregation aims to consolidate data across different omic layers to comprehensively reveal the intricate mechanisms of biological systems (Subramanian et al. 2020). Existing methods are primarily divided into statistical and deep learning categories (Ballard et al. 2024; Wang et al. 2025b). Among statistical methods, Principal Component Analysis (PCA) and its variants are widely used to reduce data dimensionality while preserving as much variance as possible (Lock et al. 2013). Similarly, Canonical Correlation Analysis (CCA) seeks to find linear combinations of variables in two datasets that are maximally correlated (Singh et al. 2019). Non-negative Matrix Factorization (NMF) decomposes the original data matrix into the product of two non-negative matrices, resulting in a low-dimensional representation of the data (Kriebel and Welch 2022). In the realm of deep learning, multiDGD (Schuster et al. 2024) employs generative modeling to learn latent inter-omics representations for improved cell state characterization. PRAGA (Huang et al. 2025) constructs dynamic graph structures and combines a prototype-aware contrastive learning strategy to integrate spatial and feature semantics. scMMAE (Meng et al. 2025) employs masked autoencoders and cross-attention mechanisms to extract both shared and modality-specific features, while also supporting knowledge transfer to unimodal analysis.

Spatial Resolved Omics

Spatially resolved omics technologies preserve spatial information within tissue sections while combining molecular measurements to reveal the spatial organization and interactions of cells in situ (Dezem et al. 2024). These technologies can be broadly categorized into three main types: 1) Image-based methods, including in situ hybridization (ISH) and in situ sequencing (ISS); 2) mRNA capture-based methods, such as laser capture microdissection (LCM); 3)Deep learning-based methods (Lee, Yoo, and Choi 2022). Compared to the first two methods, deep learning-based methods provide a flexible and scalable way. The COSMOS model (Zhou et al. 2025) employs graph convolutional networks (GCNs) to encode each omic modality and utilizes the Weighted Nearest Neighbor (WNN) algorithm to integrate multi-modal representations into a unified embedding space for tissue segmentation and functional region identification. SpatialGlue (Long et al. 2024) introduces a dual-attention

graph neural network that performs intra-omic and cross-omic integration of spatial and molecular features to decode spatial domains. Additionally, the MISO algorithm (Coleman et al. 2025) is proposed to integrate multimodal spatial omics data by first constructing adjacency matrices for each modality and using multilayer perceptrons (MLPs) to learn modality-specific low-dimensional embeddings. MISO then computes interaction features between all modality pairs and concatenates the embeddings and interactions into a unified representation for feature extraction and clustering analysis.

Method

Learning to Cluster Rare Cell Types (CRCT)

Spatial multi-modal omics integration aims to jointly model spatial information and multiple omics modalities, such as RNA expression (transcriptomics), Assay for Transposase Accessible Chromatin (ATAC-seq), and Antibody Derived Tags (ADT), in order to obtain unified representations for downstream analysis. Formally, given spatial coordinates of N tissue spots $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{N}$ and their associated features from M modalities $\mathcal{F}_M = \{f_i^m\}_{i=1,m=1}^{N,M}$ with $f^m \in \mathbb{R}^{D_m}$ denoting the m-th modality's D_m-dimensional features, the goal is to learn a latent representation Φ:

$$\mathcal{Z} = \Phi(\mathcal{F}_M, \mathcal{S}), \tag{1}$$

where $\mathcal{Z} \in \mathbb{R}^{N \times D_z}$ is a comprehensive latent embedding capturing both spatial and multi-modal molecular information. This unified embedding $\mathcal{Z}$ supports downstream tasks including cell type identification, spatial domain discovery, and tumor microenvironment analysis. Spatial multi-modal omics datasets are inherently imbalanced: rare cell types (*i*) contribute only a few sequencing spots and (*ii*) often reside at tissue boundaries where expression profiles are ambiguous. Building on the graph–based fusion mechanism of PRAGA (Huang et al. 2025), we introduce CRCT to explicitly augment the separability of these under-represented cell types, as illustrated in Figure 2.

Graph Construction and Encoding

For each modality m ($1 \le m \le M$), we construct two undirected graphs: (1) the *spatial adjacency graph* $G^S = (\mathcal{S}, A^S)$, where the adjacency matrix A_{ij}^S is constructed using the K-Nearest Neighbors (kNN) of spatial coordinates (x_i, y_i) (with 0 otherwise), and (2) the *feature adjacency graph* $G_m^F = (F_m, A_m^F)$, where F_m represents the modality-specific feature matrix and A_m^F denotes its corresponding adjacency matrix. To integrate the spatial and feature information, we first stack the adjacency matrices A^S and A_m^F along a new channel dimension, forming a 3D tensor $A_{\text{cat}} \in \mathbb{R}^{N \times N \times 2}$. The fusion is then performed through channel-wise concatenation followed by a 1×1 convolutional layer:

$$\hat{A}_m^F = \text{Conv}_{1\times1}\big([\, A^S \,\|\, A_m^F \,]\big) \in \mathbb{R}^{N \times N}, \tag{2}$$

where $\|$ denotes concatenation along the channel dimension. This operation learns optimal weights to combine spatial proximity and feature similarity while preserving the graph structure. Building upon the fused adjacency $\hat{A}_m^F$, we employ a single-layer Graph Convolutional Network (GCN) to generate refined features for each modality. The encoding process is formulated as:

$$\hat{F}_m = \hat{A}_m^F F_m W_m^{\text{enc}}, \tag{3}$$

where $W_m^{\text{enc}} \in \mathbb{R}^{d_m \times d_h}$ is a learnable weight matrix that projects the input features F_m of dimension d_m into a hidden space of dimension d_h. The resulting *encoded graph* $\tilde{G}_m^F = (\hat{F}_m, \hat{A}_m^F)$ integrates both topological and feature information from modality m, serving as the input for downstream integration tasks. Encoded features from all modalities are concatenated and transformed via a multi-layer perceptron to yield the unified representation:

$$\begin{aligned} \mathcal{Z} &= \text{MLP}\big(\text{Concat}(\hat{F}_1, \ldots, \hat{F}_M)\big) \in \mathbb{R}^{N \times D_z}, \\ \mathcal{Z} &= \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}. \end{aligned} \tag{4}$$

Cluster Implicit Semantic Data Augmentation

Spatial multimodal datasets often exhibit a pronounced class imbalance, making conventional image-level augmentations ineffective. We therefore introduce a feature-space augmentation strategy that operates on the unified embedding $\mathcal{Z}$ and draws on the idea of implicit semantic data augmentation (Wang et al. 2021; Xie et al. 2023). This strategy increases the *intra*-class diversity while keeping *inter*-class semantics intact, and requires no additional data. Let $\mathbf{z}_i \in \mathbb{R}^d$ be the latent vector of spot i and $\hat{y}_i \in \{1, \ldots, C\}$ its pseudo-label obtained from the prototype assignment. During training step t we collect the batch mean feature $\boldsymbol{\mu}_c^{\prime(t)}$, covariance $\boldsymbol{\Sigma}_c^{\prime(t)}$, and sample count $m_c^{(t)}$ for every class c present in the mini-batch. Global statistics are updated by exponential moving average:

$$\boldsymbol{\mu}_c^{(t)} = \frac{n_c^{(t-1)}\, \boldsymbol{\mu}_c^{(t-1)} + m_c^{(t)}\, \boldsymbol{\mu}_c^{\prime(t)}}{n_c^{(t-1)} + m_c^{(t)}}, \tag{5}$$

$$\begin{aligned} \boldsymbol{\Sigma}_c^{(t)} = {} & \frac{n_c^{(t-1)}\, \boldsymbol{\Sigma}_c^{(t-1)} + m_c^{(t)}\, \boldsymbol{\Sigma}_c^{\prime(t)}}{n_c^{(t-1)} + m_c^{(t)}} \\ & + \frac{n_c^{(t-1)} m_c^{(t)}}{(n_c^{(t-1)} + m_c^{(t)})^2} \\ & \times (\boldsymbol{\mu}_c^{(t-1)} - \boldsymbol{\mu}_c^{\prime(t)})(\boldsymbol{\mu}_c^{(t-1)} - \boldsymbol{\mu}_c^{\prime(t)})^\top, \end{aligned} \tag{6}$$

with the cumulative counter $n_c^{(t)} = n_c^{(t-1)} + m_c^{(t)}$. All updates are $\mathrm{O}(d^2)$ but executed only for the classes present in the mini-batch, ensuring modest overhead. Let $\mathbf{z}i \in \mathcal{Z}$ denote the latent representation of spot i, and let $\hat{y}i$ be its corresponding pseudo-label. For each class, we maintain an online estimation of its feature covariance matrix using mini-batch statistics. Specifically, from the batch-wise estimator defined in Eq. 6, we retrieve the running covariance matrix $\boldsymbol{\Sigma}^{(t)}\hat{y}i \in \mathbb{R}^{d \times d}$ for the class associated with $\hat{y}_i$ at training epoch t. Directly sampling a perturbation vector from a full-rank Gaussian distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\hat{y}_i}^{(t)})$ is computationally expensive, requiring $\mathcal{O}(d^3)$ operations due to

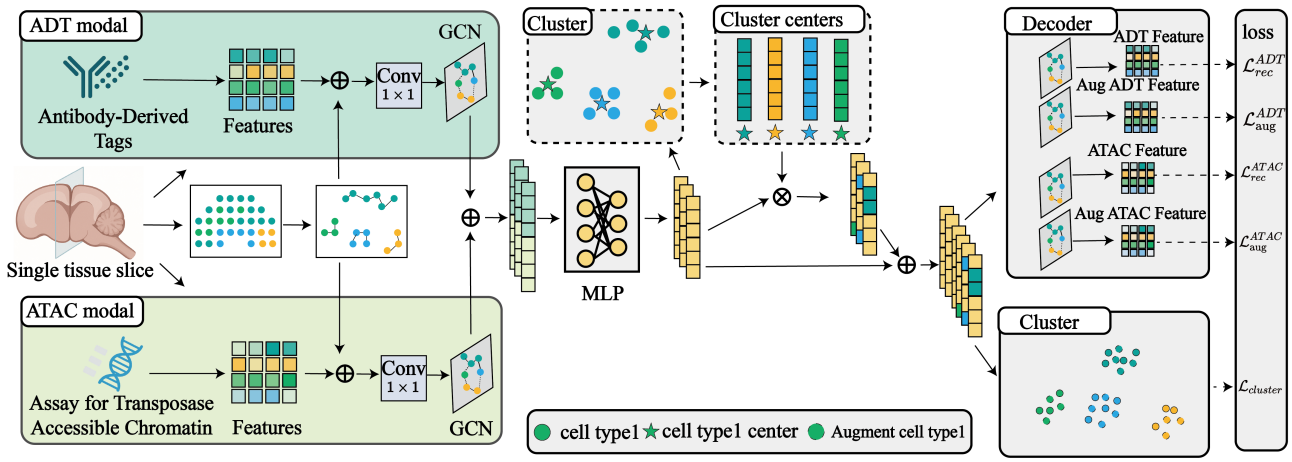


Figure 2: Learning to Cluster Rare Cell Types (CRCT) framework. For each modality, a graph convolutional network (GCN) encodes the modality-specific feature graph fused with the spatial KNN graph, producing latent node embeddings.An MLP concatenates and transforms the modality-specific embeddings into a unified representation.K-means clustering yields cluster assignments and prototypes; these prototypes drive the generation of latent augmentations that are added to the original embeddings. Modality-specific decoders reconstruct both the original and augmented features, giving reconstruction losses $\mathcal{L}_{\text{rec}}^{\text{ADT}}$ and $\mathcal{L}_{\text{rec}}^{\text{ATAC}}$, while a clustering loss $\mathcal{L}_{\text{cluster}}$ promotes compactness around prototypes.

matrix decomposition (e.g., Cholesky). To circumvent this issue, we adopt an efficient isotropic approximation following (Xie et al. 2023), which preserves the variance of each individual dimension while ignoring inter-dimensional correlations. Specifically, we construct the perturbation vector as:

$$\boldsymbol{\varepsilon}_i = \boldsymbol{\sigma}_{\hat{y}_i}^{(t)} \odot \boldsymbol{\xi}_i, \qquad \boldsymbol{\xi}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d), \tag{7}$$

where $\boldsymbol{\sigma}_{\hat{y}_i}^{(t)} = \sqrt{\operatorname{diag}\left(\boldsymbol{\Sigma}_{\hat{y}_i}^{(t)}\right)}$ stores the running per-dimension standard deviation. The augmented feature $\tilde{\mathbf{z}}_i$ for each sample is then obtained by translating $\mathbf{z}_i$ along a random direction sampled from $\mathcal{N}(\mathbf{z}_i, \rho^{(t)}\boldsymbol{\varepsilon}_i)$. The augmented embedding is:

$$\tilde{\mathbf{z}}_i = \mathbf{z}_i + \rho^{(t)}\,\boldsymbol{\varepsilon}_i, \qquad \rho^{(t)} = \frac{t}{T}\,\rho_0, \tag{8}$$

where ρ_0 is a base magnitude and the linear schedule $\rho^{(t)}$ suppresses noisy covariances in early epochs.

Multi-omics Integration and Reconstruction

After generating covariance-guided augmentations $\tilde{\mathcal{Z}} = \{\tilde{\mathbf{z}_1}, \ldots, \tilde{\mathbf{z}_N}\}$, we forward both the original and the augmented embeddings through the modality-specific decoders to recover the input omics features:

$$\begin{aligned} \hat{F}_m^{(i)} &= \text{Dec}_m(\mathbf{z}_i), \\ \hat{\tilde{F}}_m^{(i)} &= \text{Dec}_m(\tilde{\mathbf{z}}_i), \quad m = 1, \ldots, M. \end{aligned} \tag{9}$$

This yields two reconstruction terms:

$$\mathcal{L}_{\text{rec}} = \sum_{m=1}^{M} w_m \left\| F_m - \hat{F}_m \right\|_2^2, \tag{10}$$

$$\mathcal{L}_{\text{aug}} = \sum_{m=1}^{M} w_m \left\| F_m - \hat{\tilde{F}}_m \right\|_2^2, \tag{11}$$

where w_m is the weight of modality m. In practice, each omics layer carries a different signal-to-noise ratio. During training the learnable adjacency matrices $\hat{A}_m^F$ (Eq. 2) are refined by a light MLP. To prevent them from drifting too far away from the data-driven initialisation $A_{m,\text{init}}^F$, we introduce a Frobenius-norm regularizer:

$$\mathcal{L}_{\text{graph}} = \frac{1}{2M} \sum_{m=1}^{M} \left\| \left\| \hat{A}_{m,\text{MLP}}^F - A_{m,\text{init}}^F \right\|_F \right\|_F. \tag{12}$$

Contrastive Clustering of Rare Cell Types

Both the original embeddings $\{\mathbf{z}_i\}$ and their covariance-guided augmentations $\{\tilde{\mathbf{z}}_i\}$ participate in prototype alignment. Define the extended set $\mathcal{Z}^{\text{ext}} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N, \tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_N\}$ of size $2N$ and reuse the pseudo-labels $\hat{y}_i$ for both $\mathbf{z}_i$ and $\tilde{\mathbf{z}}_i$. For centroid $\mathbf{c}_k$ the positive pool is $\mathcal{P}_k = \{\, u \in \mathcal{Z}^{\text{ext}} \mid \hat{y}(u) = k \,\}$, and the negative pool is $\mathcal{N}_k = \mathcal{Z}^{\text{ext}} \setminus \mathcal{P}_k$. The temperature-scaled cluster-contrastive loss becomes:

$$\mathcal{L}_{\text{cluster}} = -\frac{1}{\hat{C}} \sum_{k=1}^{\hat{C}} \log \frac{\dfrac{1}{|\mathcal{P}_k|} \displaystyle\sum_{u \in \mathcal{P}_k} \exp\!\left(u^\top \mathbf{c}_k / \tau\right)}{\dfrac{1}{2N} \displaystyle\sum_{v \in \mathcal{Z}^{\text{ext}}} \exp\!\left(v^\top \mathbf{c}_k / \tau\right)}, \tag{13}$$

where $u, v \in \mathcal{Z}^{\text{ext}}$ are embedding vectors, and the numerator averages over both real and augmented positives and the denominator normalises over all $2N$ embeddings. Including $\tilde{\mathbf{z}}_i$ enlarges minority clusters in the positive set, yielding stronger prototype attraction and sharper decision boundaries without introducing additional raw data. In the early stage, due to the inaccurate features extracted by the model,

the clustering results in the early stage were not good. Therefore, we set λ and β to be very small in the early stage, usually 0.001 and 0, respectively, and then the gradients increased.

Overall Formulation

The the final training objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{graph}} + \lambda \mathcal{L}_{\text{aug}} + \beta \mathcal{L}_{\text{cluster}}, \quad (14)$$

where $\mathcal{L}_{\text{rec}}$ encourages faithful reconstruction of the input features, $\mathcal{L}_{\text{graph}}$ regularizes the learned adjacency matrices to remain close to the data-driven initial graphs, $\mathcal{L}_{\text{aug}}$ improves robustness by enforcing consistency under covariance-guided augmentations, and $\mathcal{L}_{\text{cluster}}$ promotes discriminative clustering via prototype-based contrastive learning. λ and β are scalar weights that balance the contribution of the augmentation and clustering terms.

Experiments

Experimental Settings

Datasets We evaluate our method on five publicly available benchmarks that cover real and synthetic spatial multi-modal scenarios, including HLN, MB, Simulation, Stereo-CITE-seq, and SPOTS datasets. **Human Lymph Node Dataset (HLN)** (Long et al. 2024) is a spatial transcriptome analysis dataset derived from human lymph node sections, which contain RNA sequencing data, Antibody Derived Tag (ADT) data, spatial coordinates, and manual annotations for 3484 spots. **Mouse Brain Dataset (MB)** (Zhang et al. 2023) is a spatial epigenome transcriptome mouse brain dataset, which is collected from juvenile mouse brain sections (P22) with paired ATAC-RNA dataset and three spatial CUT & Tag-RNA dataset; MB has 9196 spots used after SpatialGlue preprocessing. Labels are unavailable, so RNA/ATAC cluster consistency is used for validation. **Simulation Dataset** (Long et al. 2024) is a spatial multi-modal omics simulation dataset, which has ADT-seq, ATAC-seq, and RNA-seq modalities and spatial coordinates, with three modalities and explicit labels for 1296 spots, used to stress-test model behavior under known ground truth. **Mouse Thymus (Stereo-CITE-seq)** (Liao et al. 2023) is a spatial transcriptome analysis dataset. A multi-modal slice of mouse thymus providing RNA and ADT plus spatial coordinates for 4697 spots. **SPOTS Mouse Spleen (SPOTS)** (Ben-Chetrit et al. 2023) is derived from simultaneously sequenced mouse spleen tissue samples, with transcriptome, proteome sequencing, and retaining 2568 spots.

Dataset	Epoch	Learning rate	Weight decay
HLN	50	0.01	5×10^{-3}
MB	200	0.001	2×10^{-2}
Simulation	300	0.01	5×10^{-2}
SPOTS	200	0.01	5×10^{-3}
Stereo-CITE-seq	300	0.01	5×10^{-2}

Table 1: Experimental settings.

Baselines and Metrics We compare our approach with eight state-of-the-art baselines: MOFA+ (Argelaguet et al. 2020), MultiVI (Ashuach et al. 2023), TotalVI (Gayoso et al. 2021), CiteFuse (Kim et al. 2020), STAGATE (Dong and Zhang 2022), PAST (Li et al. 2023), SpatialGlue (Long et al. 2024), PRAGA (Huang et al. 2025). For fairness, we follow the data preprocessing steps described in the original papers of each baseline. Performance is assessed by nine clustering metrics: MI, NMI, AMI, FMI, ARI, V-measure, F1-score, Jaccard, and Completeness.

Preprocessing and hyperparameters All datasets are preprocessed and configured according to the settings in prior work. For fair comparison, all baselines follow the hyperparameter configurations reported in their original papers. Our model uses the following settings: number of training epochs (Epoch), learning rate, and weight decay, as detailed in Table 1. The augmentation magnitude β is set to 0 for the first 100 epochs and 1 thereafter. For the modality weights λ, we use $\{0.05, 0.5, 1\}$ on HLN (epochs 0-5, 5-15, >15) and $\{0.01, 0.2, 0.5\}$ on all other datasets (epochs 0-100, 100-200, >200).

Comparison Experiments

Quantitative results Tables 2, 3, and 4 present the scores of CRCT against eight state-of-the-art baselines on the three benchmarks described. CRCT is never worse than second and attains the best result in almost every metric. On the **HLN** dataset, CRCT outperforms the second-best method by +1.46 NMI, +0.98 FMI and +1.72 ARI. On the more challenging **MB** dataset, the margins are even larger: +1.72 NMI, +7.37 FMI and +7.76 ARI. Even on the synthetic benchmark, where PRAGA is close to optimal, CRCT still adds about +0.31 NMI, +0.20 FMI and +0.26 ARI.

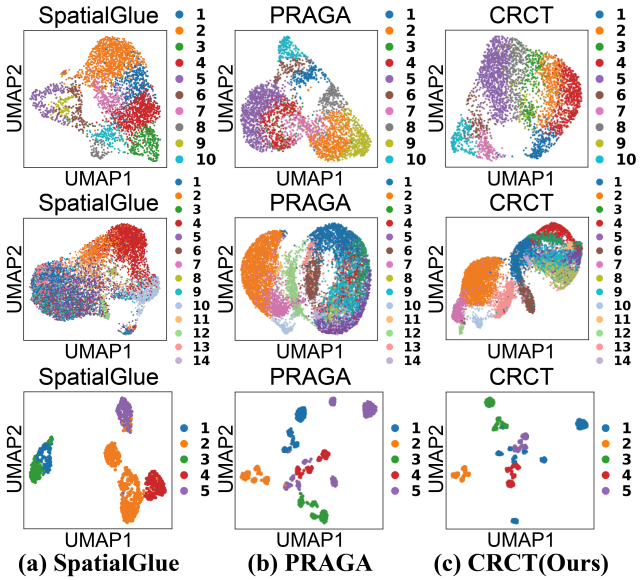


Figure 3: UMAP embeddings on HLN (top), MB (middle) and Simulation (bottom). Columns correspond to SpatialGlue, PRAGA and CRCT (ours).

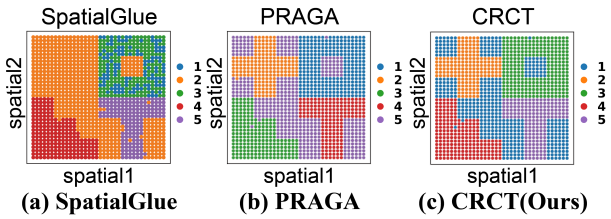
Methods	MI(%)	NMI(%)	AMI(%)	FMI(%)	ARI(%)	V-Measure(%)	F1-Score(%)	Jaccard(%)
MOFA+	65.06	34.91	34.49	37.67	22.73	34.91	36.96	22.67
CiteFuse	45.35	23.34	22.86	27.57	12.51	23.38	26.15	15.04
TotalVI	25.51	14.72	14.14	26.59	6.45	14.72	26.55	15.31
MultiVI	12.03	7.01	6.43	26.16	3.73	7.01	26.15	15.04
STAGATE	1.42	0.79	0.12	20.83	0.22	0.79	20.73	11.56
PAST	58.82	33.60	33.14	41.42	24.64	33.60	41.41	26.11
SpatialGlue	66.52	36.07	35.65	39.16	23.83	36.07	38.79	24.06
PRAGA	71.66	37.99	37.58	39.84	25.61	37.99	38.90	24.15
CRCT (Ours)	**71.40**	**39.45**	**39.04**	**42.40**	**27.33**	**39.45**	**42.18**	**26.73**
Δ	-0.26	+1.46	+1.46	+2.56	+1.72	+1.46	+3.28	+2.58

Table 2: Comparison of CRCT with other baseline methods across the Human Lymph Node Dataset. The Δ row shows the improvement over the second-best result.

Methods	MI(%)	NMI(%)	AMI(%)	FMI(%)	ARI(%)	V-Measure(%)	F1-Score(%)	Jaccard(%)
MOFA+	19.58	8.64	8.38	15.59	4.39	8.64	15.59	8.45
CiteFuse	47.48	19.46	19.05	17.96	8.24	19.46	17.89	9.82
MultiVI	17.88	8.47	8.22	18.12	3.81	8.47	17.58	9.63
STAGATE	48.45	21.25	21.03	22.36	12.21	21.25	22.36	12.59
PAST	69.49	29.13	28.76	24.54	14.63	29.13	24.54	13.99
SpatialGlue	95.54	37.83	37.53	33.78	26.33	37.83	33.01	19.77
PRAGA	95.55	39.37	39.06	35.07	27.06	39.37	35.02	21.23
CRCT (Ours)	**96.09**	**41.09**	**40.85**	**42.44**	**34.82**	**41.09**	**42.43**	**26.93**
Δ	+0.54	+1.72	+1.79	+7.37	+7.76	+1.72	+7.41	+5.70

Table 3: Comparison of CRCT with other baseline methods across the Mouse Brain Dataset. The Δ row shows the improvement over the second-best result.

Qualitative visualisation Figure 3 illustrates the two-dimensional UMAP projections of HLN, MB, and Simulation embeddings produced by SpatialGlue, PRAGA and CRCT. Three consistent observations emerge: (i) Clusters produced by CRCT are more compact and mutually separated than those of the baselines, especially for rare or boundary cell types (e.g., cluster 7 in MB). (ii) On the Simulation dataset, CRCT recovers all five classes cleanly; SpatialGlue merges clusters 1 and 2, while PRAGA shows mild mixing. (iii) Across datasets, CRCT preserves global geometry better, avoiding the elongated or folded manifolds visible in the baseline plots.



(a) SpatialGlue **(b) PRAGA** **(c) CRCT(Ours)**

Figure 4: Spatial domain reconstruction on the Simulation dataset. CRCT best matches the ground truth, while SpatialGlue and PRAGA exhibit domain mixing.

Figure 4 further compares spatial domain assignments on Simulation. SpatialGlue incorrectly merges the upper-left domains, PRAGA mislabels a small fraction of spots, whereas CRCT faithfully reconstructs the ground-truth cross-shaped pattern with only a few isolated errors. These visual trends corroborate the quantitative gains, confirming that covariance-guided augmentation and dynamic prototype refinement jointly improve both cluster compactness and spatial coherence.

Ablation Study

To assess the individual role of implicit semantic augmentation, we created three simplified variants of our model. The first variant, denoted $\mathcal{L}_{\text{aug}}$, computes reconstruction exclusively on augmented embeddings, discarding the real ones. The second variant, $\mathcal{L}_{\text{cluster}}$, applies augmentation only to the clustering term; reconstruction is performed on real embeddings. Ablation results in Table 5 demonstrate that each variant improves over the baseline, indicating that covariance-guided noise helps alleviate overfitting and highlights minority clusters. The full CRCT configuration, which applies augmentation to both reconstruction and clustering, achieves the highest performance on every dataset, showing that the two loss components act in a complementary fashion: recon-

Methods	MI(%)	NMI(%)	AMI(%)	FMI(%)	ARI(%)	V-Measure(%)	F1-Score(%)	Jaccard(%)
MOFA+	1.02	0.58	-0.23	21.32	0.39	0.58	21.27	11.90
CiteFuse	1.23	0.66	-0.10	17.17	0.03	0.66	16.56	9.03
TotalVI	1.36	0.72	-0.02	15.93	-0.09	0.72	15.03	8.12
MultiVI	1.22	0.77	-0.05	25.20	-0.01	0.77	25.05	14.32
STAGATE	7.40	3.91	3.91	17.25	1.56	3.91	16.23	8.83
PAST	2.09	1.18	1.18	19.17	0.07	1.18	18.91	10.44
SpatialGlue	150.13	96.97	96.97	98.21	97.69	96.98	98.21	96.48
PRAGA	152.07	98.26	98.26	98.97	98.67	98.26	98.97	97.97
CRCT (Ours)	**152.62**	**98.57**	**98.56**	**99.17**	**98.93**	**98.56**	**99.17**	**98.35**
Δ	+0.55	+0.31	+0.30	+0.20	+0.26	+0.30	+0.20	+0.38

Table 4: Comparison of CRCT with other baseline methods across the Spatial Multi-modal Omics Simulation Dataset. The Δ row shows the improvement over the second-best result.

Methods	MI(%)	NMI(%)	AMI(%)	FMI(%)	ARI(%)	V-Measure(%)	F1-Score(%)	Jaccard(%)
PRAGA	71.66	37.99	37.58	39.84	25.61	37.99	38.90	24.15
$\mathcal{L}_{\text{cluster}}$	71.34	38.87	38.46	42.13	27.35	38.87	41.78	26.41
$\mathcal{L}_{\text{aug}}$	72.12	39.20	38.79	41.39	26.36	39.20	41.07	25.84
CRCT (Ours)	**71.40**	**39.45**	**39.04**	**42.40**	**27.33**	**39.45**	**42.18**	**26.73**
Δ vs PRAGA	-0.26	+1.46	+1.46	+2.56	+1.72	+1.46	+3.28	+2.58

Table 5: Ablation Study on HLN Dataset. The Δ row shows the improvement over the PRAGA baseline.

struction maintains cross-modal fidelity, whereas prototype-aware clustering sharpens decision boundaries.

Parameter Sensitivity Experiments

We conduct the augmentation loss weight λ sensitivity analysis on MB dataset. The λ controls the intensity of prototype-based implicit augmentation. Figure 5 shows that performance increases with λ up to 0.5 and declines beyond, validating the effectiveness of moderate augmentation strength in balancing diversity and semantic consistency. The optimal λ value of 0.5 represents a balance between augmentation diversity and semantic consistency. Values below 0.3 provide insufficient augmentation, while values above 0.7 introduce excessive noise that degrades clustering performance.

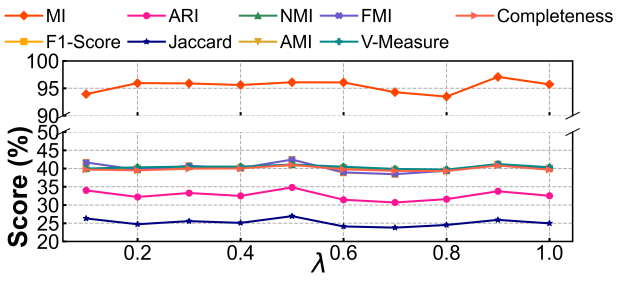


Figure 5: Parameter sensitivity experiments for λ on MB dataset.

Conclusion

In this paper, we presented CRCT, a novel framework designed to address the challenge of clustering rare cell types in spatial multi-modal omics data. By leveraging implicit semantic data augmentation (ISDA) and adaptive graph learning, CRCT overcomes key limitations such as data sparsity, high dimensionality, and class imbalance. Unlike traditional methods that rely on explicit synthetic samples, CRCT dynamically estimates intra-class covariance matrices and perturbs features along semantically meaningful directions in the latent space. This approach effectively enhances the representation of rare cell populations while preserving biological fidelity. Our extensive experiments on four real-world datasets (HLN, MB, Stereo-CITE-seq, and SPOTS) and one synthetic benchmark demonstrate CRCT's superior performance, achieving improvements of up to +1.7 NMI and +7.8 ARI over state-of-the-art baselines. We believe this work not only advances the field of spatial multi-omics analysis but also inspires further research into methods for uncovering rare yet biologically significant cell states.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grants No. 62406100 and No. 92570118. Tianjin Natural Science Foundation under Grants No. 24JCQNJC00320, Beijing Postdoctoral Research Foundation.

References

Argelaguet, R.; Arnol, D.; Bredikhin, D.; Deloro, Y.; Velten, B.; Marioni, J. C.; and Stegle, O. 2020. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome biology*, 21: 1–17.

Ashuach, T.; Gabitto, M. I.; Koodli, R. V.; Saldi, G.-A.; Jordan, M. I.; and Yosef, N. 2023. MultiVI: deep generative model for the integration of multimodal data. *Nature Methods*, 20(8): 1222–1231.

Ballard, J. L.; Wang, Z.; Li, W.; Shen, L.; and Long, Q. 2024. Deep learning-based approaches for multi-omics data integration and analysis. *BioData Mining*, 17(1): 38.

Ben-Chetrit, N.; Niu, X.; Swett, A. D.; Sotelo, J.; Jiao, M. S.; Stewart, C. M.; Potenski, C.; Mielinis, P.; Roelli, P.; Stoeckius, M.; et al. 2023. Integration of whole transcriptome spatial profiling with protein markers. *Nature biotechnology*, 41(6): 788–793.

Chen, X.; Shen, Y.; Draper, W.; Buenrostro, J. D.; Litzenburger, U.; Cho, S. W.; Satpathy, A. T.; Carter, A. C.; Ghosh, R. P.; East-Seletsky, A.; et al. 2016. ATAC-see reveals the accessible genome by transposase-mediated imaging and sequencing. *Nature methods*, 13(12): 1013–1020.

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*, 1–9.

Conesa, A.; Madrigal, P.; Tarazona, S.; Gomez-Cabrero, D.; Cervera, A.; McPherson, A.; Szcześniak, M. W.; Gaffney, D. J.; Elo, L. L.; Zhang, X.; et al. 2016. A survey of best practices for RNA-seq data analysis. *Genome biology*, 17: 1–19.

Cui, S.; Wang, S.; Zhuo, J.; Li, L.; Huang, Q.; and Tian, Q. 2020. Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3941–3950.

Dezem, F. S.; Morosini, N. S.; Arjumand, W.; DuBose, H.; and Plummer, J. 2024. Spatially resolved single-cell omics: methods, challenges, and future perspectives. *Annual Review of Biomedical Data Science*, 7.

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

Gayoso, A.; Steier, Z.; Lopez, R.; Regier, J.; Nazor, K. L.; Streets, A.; and Yosef, N. 2021. Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nature methods*, 18(3): 272–282.

Gravis, G.; Boher, J.-M.; Joly, F.; Soulié, M.; Albiges, L.; Priou, F.; Latorzeff, I.; Delva, R.; Krakowski, I.; Laguerre, B.; et al. 2016. Androgen deprivation therapy (ADT) plus docetaxel versus ADT alone in metastatic non castrate prostate cancer: impact of metastatic burden and long-term survival analysis of the randomized phase 3 GETUG-AFU15 trial. *European urology*, 70(2): 256–262.

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.

Kim, H. J.; Lin, Y.; Geddes, T. A.; Yang, J. Y. H.; and Yang, P. 2020. CiteFuse enables multi-modal analysis of CITE-seq data. *Bioinformatics*, 36(14): 4137–4143.

Kriebel, A. R.; and Welch, J. D. 2022. UINMF performs mosaic integration of single-cell multi-omic datasets using nonnegative matrix factorization. *Nature communications*, 13(1): 780.

Lee, J.; Yoo, M.; and Choi, J. 2022. Recent advances in spatially resolved transcriptomics: challenges and opportunities. *BMB reports*, 55(3): 113.

Li, Z.; Chen, X.; Zhang, X.; Jiang, R.; and Chen, S. 2023. Latent feature extraction with a prior-based self-attention framework for spatial transcriptomics. *Genome Research*, 33(10): 1757–1773.

Liao, S.; Heng, Y.; Liu, W.; Xiang, J.; Ma, Y.; Chen, L.; Feng, X.; Jia, D.; Liang, D.; Huang, C.; et al. 2023. Integrated spatial transcriptomic and proteomic analysis of fresh frozen tissue based on stereo-seq. *bioRxiv*, 2023–04.

Lock, E. F.; Hoadley, K. A.; Marron, J. S.; and Nobel, A. B. 2013. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *The annals of applied statistics*, 7(1): 523.

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.

Meng, D.; Feng, Y.; Yuan, K.; Yu, Z.; Cao, Q.; Cheng, L.; and Zheng, X. 2025. scMMAE: masked cross-attention network for single-cell multimodal omics fusion to enhance unimodal omics. *Briefings in Bioinformatics*, 26(1): bbaf010.

Moses, L.; and Pachter, L. 2022. Museum of spatial transcriptomics. *Nature methods*, 19(5): 534–546.

Rao, A.; Barkley, D.; França, G. S.; and Yanai, I. 2021. Exploring tissue architecture using spatial transcriptomics. *Nature*, 596(7871): 211–220.

Schuster, V.; Dann, E.; Krogh, A.; and Teichmann, S. A. 2024. multiDGD: A versatile deep generative model for multi-omics data. *Nature Communications*, 15(1): 10031.

Singh, A.; Shannon, C. P.; Gautier, B.; Rohart, F.; Vacher, M.; Tebbutt, S. J.; and Lê Cao, K.-A. 2019. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics*, 35(17): 3055–3062.

Su, H.; Wang, B.; Liu, D.; Li, J.; Feng, C.-B.; and Vong, C.-M. 2025. Towards Fully Test-Time Adaptation via Variance Balancing and Semantic Augmentation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Subramanian, I.; Verma, S.; Kumar, S.; Jere, A.; and Anamika, K. 2020. Multi-omics data integration, interpretation, and its application. *Bioinformatics and biology insights*, 14: 1177932219899051.

Tian, L.; Chen, F.; and Macosko, E. Z. 2023. The expanding vistas of spatial transcriptomics. *Nature Biotechnology*, 41(6): 773–782.

Wang, M.; houcheng su; Li, J.; Li, C.; Yin, N.; Shen, L.; and Guo, J. 2025a. GraphCL: Graph-based Clustering for Semi-Supervised Medical Image Segmentation. In *Forty-second International Conference on Machine Learning*.

Wang, M.; Ren, W.; Zhang, Y.; Fan, Y.; Shi, D.; Jing, L.; and Yin, N. 2025b. Gaussian Mixture Model for Graph Domain Adaptation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*.

Wang, M.-z. 2025. SimProF: A Simple Probabilistic Framework for Unsupervised Domain Adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 21153–21161.

Wang, Y.; Huang, G.; Song, S.; Pan, X.; Xia, Y.; and Wu, C. 2021. Regularizing deep networks with semantic data augmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7): 3733–3748.

Williams, C. G.; Lee, H. J.; Asatsuma, T.; Vento-Tormo, R.; and Haque, A. 2022. An introduction to spatial transcriptomics for biomedical research. *Genome medicine*, 14(1): 68.

Xie, B.; Li, S.; Li, M.; Liu, C. H.; Huang, G.; and Wang, G. 2023. Sepico: Semantic-guided pixel contrast for domain adaptive semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7): 9004–9021.

Zhang, D.; Deng, Y.; Kukanja, P.; Agirre, E.; Bartosovic, M.; Dong, M.; Ma, C.; Ma, S.; Su, G.; Bao, S.; et al. 2023. Spatial epigenome–transcriptome co-profiling of mammalian tissues. *Nature*, 616(7955): 113–122.

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.