FLOWMOS: CONTINUOUS LATENT TRANSPORT FOR SPATIAL MULTI-OMICS INTEGRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial multi-omics integration aims to combine molecular measurements from different modalities with spatial information to learn a unified representation that can be used for downstream biological analysis. However, significant differences among modalities in terms of dimensionality, sparsity, noise, and biological semantics make it difficult to obtain stable and discriminative unified representations, while the transformation from modality specific representations to the shared space is typically modeled implicitly. To address these challenges, we proposes FlowMOS, a flow matching representation learning framework for spatial multi-omics integration. FlowMOS introduces shared space flow matching to explicitly model the continuous transfer process from each modality specific representation to a shared latent space. To further improve the robustness of representation learning against incomplete and noisy measurements, mask consistency regularization is employed to simulate local feature missingness and sequencing noise in the input data. Extensive experiments on several real world spatial multi-omics datasets demonstrate that FlowMOS effectively captures modality to shared transformations, while the unified representation achieve robust performance in downstream task.

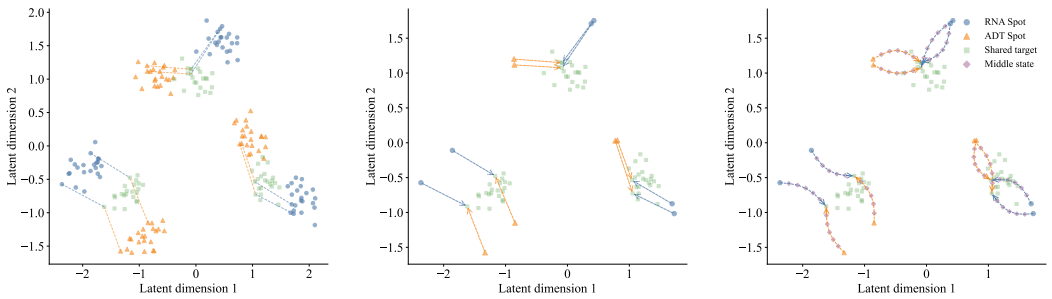
1 INTRODUCTION

Spatial multi-omics technology provides entirely new insights into biological research by simultaneously measuring multiple molecular layers at the same spatial location (Takei et al., 2025; Hsieh et al., 2022). While preserving spatial location information, this technology captures molecular information from the transcriptome (Zheng et al., 2017), surface proteins (Grandi et al., 2022), chromatin accessibility (McDonough et al., 2015), epigenetic signals (Deng et al., 2022) and other modalities, offering a comprehensive perspective on cell identity and spatial domain division. However, these modalities differ significantly in dimensionality, sparsity, noise, and biological semantics. Therefore, learning a unified latent representation that can integrate information from different modalities while preserving biological meaning has become a central challenge in spatial multi-omics analysis.

Spatial multi-omics integration is not a simple concatenation of feature information from different modalities, existing methods often employ probabilistic generative models and graph neural networks for integration. For example, MultiVI (Ashuach et al., 2023) uses modality specific variational autoencoders to map data from different modalities onto a shared latent space, and then integrates them by constraining the distance between the latent representations of different modalities. SpatialGlue (Long et al., 2024) constructs spatial and feature graph representations for different modalities and integrates their latent representations through a cross modal attention mechanism. However, in these methods, the transformation process from modality specific representations to a shared space is modeled implicitly. Due to modality differences, when representing different modalities of the same spot in a unified latent representation space, the transformation direction and magnitude may vary.

To better illustrate this concept, we present a simple example in Figure 1. As shown in Figure 1(a), although different modality representations describe the same biological sample, the patterns by which they align with a unified latent representation may differ. Existing integration methods primarily focus on obtaining a unified latent representation, as shown in Figure 1(b), but do not explic-

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.



(a) Modality specific discrepancy. (b) Endpoint based integration. (c) Flow Matching based transport.

Figure 1: Illustration of spatial multi-omics integration, contrasting endpoint based mapping with Flow Matching based continuous transport from modality specific representations to a shared latent space.

itly describe how each modality changes as it progresses toward this shared space. Flow matching offers a different perspective (Huguet et al., 2022; Tong et al., 2020). As shown in Figure 1(c), it not only specifies the final unified latent representation but also models the entire transformation process from modality specific representations to the unified latent representation through continuous trajectories. In this way, observations from different modalities can converge toward the final goal along distinct paths, thereby providing additional guidance for learning better unified latent representations.

Based on the above perspectives, we view spatial multi-omics integration as a continuous process of transferring data from different modalities to a shared space, and propose a flow matching representation learning framework (FlowMOS) for spatial multi-omics integration. Unlike methods that focus on the final unified latent representation, FlowMOS introduces flow matching to explicitly model the continuous transition from modality specific representations to a unified latent representation, enabling the model to constrain the process by which different modalities converge toward the shared space. During training, FlowMOS randomly masks the input features of each modality and constrains the masked representations to remain consistent with the complete representations, thereby enhancing the model's robustness to noise and missing feature information. Our research contributions are summarized as follows:

- We propose FlowMOS, a flow matching representation learning framework for spatial multi-omics integration. By explicitly modeling the continuous transition process from modality specific representations to a unified latent representation, we provide a new perspective on representation learning for spatial multi-omics integration.
- We formulate a joint optimization objective that combines flow matching with mask consistency regularization. In this framework, flow matching is used to constrain the continuous transition from modality specific representations to a unified latent representation, while the mask consistency regularization constraint enhances the stability of the unified latent representation against noise and the loss of local features.
- We conducted a systematic evaluation of FlowMOS on multiple real world spatial multi-omics datasets. Experimental results demonstrate that this method improves the quality of unified latent representations and the performance of downstream spatial domain classification.

2 METHOD

2.1 PROBLEM FORMULATION

The goal of spatial multi-omics integration is to fully leverage molecular and spatial information from different spatial omics modalities to learn a unified low dimensional representation that supports subsequent biological analysis tasks. For a spatial multi-omics dataset containing

N spots and M omics modalities, the raw feature matrix of the m-th modality is denoted as $X^{(m)} \in \mathbb{R}^{N \times d_m}$, $m = 1, \cdots, M$, where d_m represents the feature dimension of the m-th modality, all modalities share the same spatial coordinates $S \in \mathbb{R}^{N \times 2}$. After preprocessing steps such as quality control, normalization, feature selection, and dimensionality reduction, the input feature representation for the m-th modality is denoted as $F^{(m)} = \{f_i^{(m)}\}_{i=1}^{N}$, where $f_i^{(m)} \in \mathbb{R}^{d_m}$ denotes the preprocessed feature vector of the i-th spot in the m-th modality. By integrating modal feature information and spatial information, we obtain the modality specific latent representation $\mathcal{Z}^{(m)} = \{z_i^{(m)}\}_{i=1}^{N}, \quad z_i^{(m)} \in \mathbb{R}^d$, where d denotes the dimension of the latent representation. Finally, the modality specific latent representations are fused into a unified latent representation $\mathcal{Z} = \{z_i\}_{i=1}^{N}, \quad z_i \in \mathbb{R}^d$, which serves as the final spatial multi-omics integration result and is used for downstream tasks such as spatial domain identification and clustering analysis. Existing spatial multi-omics integration methods typically focus on obtaining the final unified latent representation $\mathcal{Z}$, while the process by which different modalities transition from modality specific representations to a shared space is often not explicitly modeled. Furthermore, sequencing noise and local feature missingness can also affect the stability of representation learning. To address this, we model spatial multi-omics integration as a continuous transmission process from modality specific latent representations $\mathcal{Z}^{(m)}$ to a unified latent representation $\mathcal{Z}$, and learn a time dependent velocity field $v_\theta(\mathcal{Z}^{(m)}, t)$. Building on this, we use a flow matching loss to explicitly constrain the continuous transition from different modalities to the shared space, while introducing masked representation consistency regularization to enhance the model's robustness to local feature loss and noise. Ultimately, we obtain a stable, unified latent representation for spatial multi-omics, as illustrated in Figure 2.

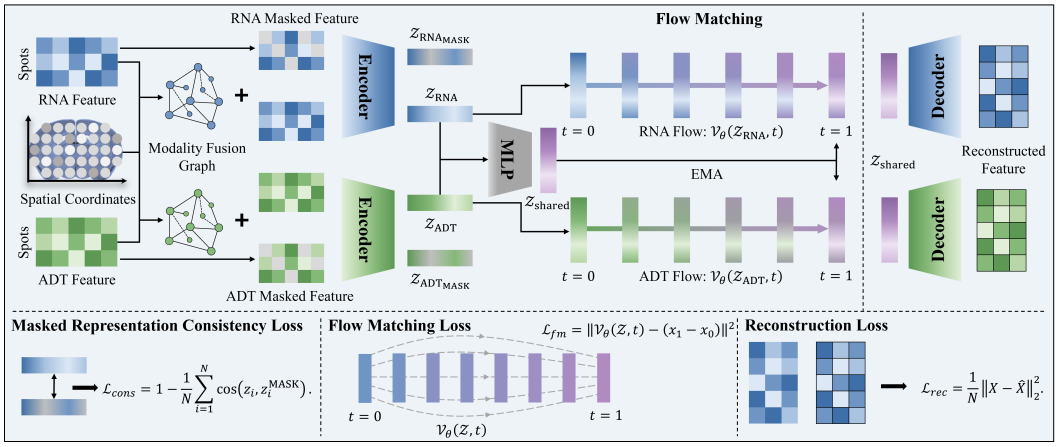


Figure 2: The framework of the proposed FlowMOS. For each modality, FlowMos constructs a modality fusion graph and encodes it to obtain modality specific representation and unified latent representation. Shared space flow matching explicitly models the continuous transport from modality specific representations to the unified latent representation. Subsequently mask consistency regularization improves robustness to local feature perturbations during spatial multi-omics integration.

2.2 Spatial Multi-Omics Graph Encoding

To simultaneously capture spatial relationships and feature relationships within different omics modalities, FlowMOS first constructs a spatial graph and a modality specific feature graph for each omics modality based on the KNN algorithm. For the m-th modality, a spatial adjacency graph $G^s \in \mathbb{R}^{N \times N}$ is constructed based on the spatial coordinates S, and a modality specific feature graph $G_F^{(m)} \in \mathbb{R}^{N \times N}$ is constructed based on the preprocessed modality features $F^{(m)}$. The two graphs are then stacked as $G^{(m)} = [G^s, G_F^{(m)}]$ and fused using a modality specific 1×1 convolution $\widehat{G}^{(m)} = \text{Conv}_{1\times1}^{(m)}(G^{(m)})$ to obtain the modality specific fusion graph. After obtaining the modality specific fused graph $\widehat{G}^{(m)}$, FlowMOS uses a modality specific graph encoder to map the input features of the m-th modality to a latent space. The encoding process is as follows:

$$\mathcal{Z}^{(m)} = \widehat{G}^{(m)} F^{(m)} W_E^{(m)}, \tag{1}$$

where $W_E^{(m)}$ represents the learnable feature mapping matrix for the m-th modality, and $\mathcal{Z}^{(m)} \in \mathbb{R}^{N \times d}$ represents the modality specific latent representation for the m-th modality. After obtaining the modality specific latent representations for all modalities, the final unified latent representation is derived through a concatenation operation and a multilayer perceptron:

$$\mathcal{Z}_{shared} = f_\phi(\mathcal{Z}^{(1)} || \mathcal{Z}^{(2)} || \cdots || \mathcal{Z}^{(M)}), \tag{2}$$

where $|| \cdot ||$ denotes the concatenation operation, and $f_\phi(\cdot)$ denotes a multilayer perceptron. To ensure that the unified latent representation retains the effective features specific to each modality, FlowMOS further introduces modality specific graph decoders $\widehat{F}^{(m)} = G^s \mathcal{Z}_{shared} W_D^{(m)}$, where $W_D^{(m)}$ denotes the learnable parameters of the decoder for the m-th modality, and $\widehat{F}^{(m)}$ denotes the reconstructed features for the m-th modality. The reconstruction loss is defined as:

$$\mathcal{L}_{rec} = \sum_{m=1}^{M} \frac{w_m}{N d_m} ||F^{(m)} - \widehat{F}^{(m)}||_F^2, \tag{3}$$

where w_m represents the weight for the m-th modality's reconstruction loss, used to balance the contributions of different modalities during the reconstruction process. This loss function is used to constrain the graph encoding and decoding process, ensuring that the resulting unified latent representation better preserves modality specific information. Spatial multi-omics graph encoding establishes a mapping relationship between different modalities and the shared space. However, this process still focuses primarily on the construction of the final representation and lacks explicit modeling of the continuous transition from modality specific representations to a unified latent representation. Therefore, FlowMOS further introduces shared space flow matching, which models the mapping from modality specific representation to the shared space as a continuous time dynamical process, thereby providing additional constraints for unified latent representation learning.

2.3 Shared Space Flow Matching

Based on the spatial multi-omics graph representation learning, we can obtain modality specific representations $\mathcal{Z}^{(m)}$ and a unified latent representation $\mathcal{Z}_{shared}$. Although this part can establish mapping relationships between different modalities and the shared space, this process primarily focuses on constructing the final unified latent representation, while the continuous transition from modality specific representations to the shared latent space still lacks explicit constraints. To address this, we further introduce shared space flow matching, which models the mapping of the latent representation $\mathcal{Z}^{(m)}$ of the m-th modality to the unified latent representation $\mathcal{Z}_{shared}$ as a continuous time dynamical process, and characterizes the continuous transport trajectory of the latent representation in the shared space by learning a time dependent velocity field (Zhang et al.; Tong et al., 2023).

For the m-th modality, we define a modality specific velocity field $v_{\theta_m}(z, t)$, where $t \sim \mathcal{U}(0, 1)$ represents a time variable randomly sampled from a uniform distribution, and θ_m represents the parameters of the corresponding flow network. We use the modality specific representation $z_0^{(m)} = \mathcal{Z}^{(m)}$ as the transmission starting point. Since the unified latent representation $\mathcal{Z}_{shared}$ is continuously updated during training, directly using the current unified latent representation as the target for flow matching may cause the transmission endpoint to change too rapidly. To obtain a more stable target representation, FlowMOS employs an exponential moving average (EMA) to smooth the updates to the unified latent representation $\bar{\mathcal{Z}}_{shared} = \beta \bar{\mathcal{Z}}_{shared} + (1 - \beta)\,\mathrm{sg}(\mathcal{Z}_{shared})$, where β is the EMA update coefficient, $\mathrm{sg}(\cdot)$ denotes the stop gradient operation, and $z_1 = \bar{\mathcal{Z}}_{shared}$ is used as the transmission endpoint shared by all modalities. FlowMOS constructs continuous intermediate states between the modality specific representations and the unified latent representation via linear interpolation:

$$z_t^{(m)} = (1-t)z_0^{(m)} + tz_1, \tag{4}$$

where, when $t = 0$, the intermediate state corresponds to the original modality specific representation, and when $t = 1$, it reaches the stable unified latent representation. Taking the derivative of this path with respect to time t yields the target transmission velocity:

$$u_t^{(m)} = \frac{\mathrm{d}z_t^{(m)}}{\mathrm{d}t} = \mathrm{sg}(z_1 - z_0^{(m)}). \tag{5}$$

By inputting the intermediate state $z_t^{(m)}$ and time t into the corresponding flow network, we obtain the predicted velocity $\hat{u}_t^{(m)} = v_{\theta_m}\left(z_t^{(m)}, t\right)$. The model learns this continuous transport process by minimizing the mean squared error between the predicted velocity and the target velocity. The flow matching loss is ultimately defined as

$$\mathcal{L}_{fm} = \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}_{t\sim\mathcal{U}(0,1)}[||v_{\theta_m}\left(z_t^{(m)}, t\right) - u_t^{(m)}||_2^2]. \tag{6}$$

In this way, the model no longer relies solely on the relationship between modality specific representations and the unified latent representation at the two endpoints. Instead, it learns the direction of change in the latent representation over time based on randomly sampled intermediate states, thereby explicitly characterizing the continuous transmission process from different modalities to the shared space and providing additional constraints for unified latent representation learning.

2.4 Mask Consistency Regularization

Spatial omics data are characterized by high dimensionality and sparsity. Additionally, measurement noise and the loss of local features that can easily occur during the sequencing process may cause fluctuations in the latent representations, making the model sensitive to perturbations in the input features. To enhance the stability of modality specific latent representations, FlowMOS further introduces mask consistency regularization. By applying random masking to the input features and constraining the consistency of modality specific latent representations before and after perturbations, the encoder is enabled to learn representations that are more robust to local feature loss. For each spatial location of the preprocessed features $F^{(m)}$ of the m-th modality, a masking ratio $r_i^{(m)} \sim \mathcal{U}(r_{\min}, r_{\max})$ is sampled independently, and a binary masking matrix $M^{(m)}$ is randomly generated based on this ratio. The masked modal feature is defined as $\widetilde{F}^{(m)} = F^{(m)} \odot M^{(m)}$, where $\odot$ denotes element wise multiplication, and the dimensions of the masked features are set to zero. The random masking is applied to the preprocessed feature space rather than directly removing entire modalities, enabling the model to simulate local feature loss and noise while preserving the omics structure. Subsequently, the perturbed features $\widetilde{F}(m)$ are fed into a graph encoder that shares parameters with the original branch, and graph information is propagated using the same fusion graph $\widehat{G}^{(m)}$ to obtain the masked modality specific latent representation:

$$\widetilde{\mathcal{Z}}^{(m)} = \widehat{G}^{(m)}\widetilde{F}^{(m)}W_E^{(m)}. \tag{7}$$

To ensure that the model can still recover the latent structure consistent with the original input even when local features are missing, FlowMOS uses the original modality specific representation as a reference target and minimizes the directional discrepancy between it and the masked modality specific representation. The mask consistency loss is defined as

$$\mathcal{L}_{cons} = \frac{1}{M}\sum_{m=1}^{M}[1 - \frac{1}{N}\sum_{i=1}^{N}\frac{\mathrm{sg}\left(z_i^{(m)}\right)^\top \widetilde{z}_i^{(m)}}{\left\|\mathrm{sg}\left(z_i^{(m)}\right)\right\|_2 \left\|\widetilde{z}_i^{(m)}\right\|_2}], \tag{8}$$

where $z_i(m)$ and $\widetilde{z}_i(m)$ denote the original modality specific latent representation and the masked modality specific latent representation of the i-th spot, respectively. By treating the original representation as a fixed reference target, gradients are backpropagated only through the masking branch, thereby prompting the encoder to obtain a latent representation consistent with the complete input from a partially missing input.

Through the constraint of consistency between the latent representation and the randomly masked input, the model is able to reduce over reliance on local features and enhance robustness against feature loss and input perturbations. This regularization term complements the shared space flow matching mechanism, promoting representation learning from two aspects: internal modality stability and continuous transmission from the modality to the shared space.

2.5 Model Training of FlowMOS

Building upon the reconstruction loss, flow matching loss, and mask consistency loss, FlowMOS's training objective also employs a graph structure consistency loss from previous spatial multi-omics methods to prevent significant fluctuations in the graph structure during training. Specifically, for the modal feature graph $G_F^{(m)}$ of the m-th modality, FlowMOS constructs a corresponding reference graph $\bar{G}_F^{(m)}$ based on an exponential moving average $\bar{G}_F^{(m)} = \alpha \bar{G}_F^{(m)} + (1-\alpha)\,\mathrm{sg}\left(G_F^{(m)}\right)$, where α represents the update coefficient of the reference graph. Based on this, the graph structure loss is defined as,

$$\mathcal{L}_g = \frac{1}{M}\sum_{m=1}^{M}\left\| G_F^{(m)} - \mathrm{sg}\left(\bar{G}_F^{(m)}\right)\right\|_F. \tag{9}$$

By integrating the above constraints, we jointly optimize the model training process and the final unified latent representation in an end-to-end manner. The overall training objective of FlowMOS is defined as

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \lambda_1 \mathcal{L}_{fm} + \lambda_2 \mathcal{L}_{cons} + \lambda_3 \mathcal{L}_g. \tag{10}$$

Here, λ_1, λ_2, and λ_3 are used to balance the contributions of the flow matching loss, the mask consistency loss, and the graph structure consistency loss, respectively.

2.6 Shared Representation Inference and Latent Transport

Upon completion of model training, the unified latent representation $\mathcal{Z}_{shared}$ serves as the final integrated representation of spatial multi-omics data. This representation synthesizes molecular feature information from different omics modalities along with shared spatial structural information, and can be further utilized for downstream analyses such as spatial domain identification, clustering, and low dimensional visualization. The flow matching module is primarily used to characterize the continuous transfer process from modality specific representations to the unified latent representation, without replacing the unified latent representation $\mathcal{Z}_{shared}$ as the final integrated result. In addition to obtaining a unified latent representation, the modality specific velocity field trained can also be used to explicitly infer the transmission process of different modalities in the latent space. For the m-th modality, using its modality specific representation $\mathcal{Z}^{(m)}$ as the initial state, the evolution of its latent state over time can be expressed as:

$$\frac{\mathrm{d}z^{(m)}(t)}{\mathrm{d}t} = v_{\theta_m}\left(z^{(m)}(t), t\right), \qquad z^{(m)}(0) = \mathcal{Z}^{(m)}. \tag{11}$$

By integrating from $t = 0$ to $t = 1$, we obtain the potential endpoint of the m-th mode after continuous transmission:

$$\mathcal{Z}_{\text{Flow}}^{(m)} = z^{(m)}(1) = \mathcal{Z}^{(m)} + \int_0^1 v_{\theta_m}\left(z^{(m)}(t), t\right) \mathrm{d}t. \tag{12}$$

This continuous dynamical process is discretely integrated using a numerical ordinary differential equation solver(Chen et al., 2018; Liu et al., 2022), thereby gradually updating the latent state and obtaining the final transmission result. Unlike the training phase, where the velocity field is learned through random sampling of intermediate states, the inference phase starts from the true modality specific representation $\mathcal{Z}^{(m)}$ and performs a full integration along the learned velocity field up to $t = 1$, ultimately yielding the actual transfer endpoint $\mathcal{Z}_{\text{Flow}}^{(m)}$ from the modality to the shared space as learned by the model.

3 Experiments

3.1 Experimental Setup

Datasets. We evaluated FlowMOS using four real world spatial multi omics datasets from different tissue sources and with varying modal combinations. The Human Lymph Node dataset (Long et al., 2024) contains paired RNA and ADT measurements. The Spatial epigenome transcriptome mouse brain dataset (Zhang et al., 2023) contains paired RNA and ATAC modal data. The Mouse thymus stereo-CITE-seq dataset (Liao et al., 2023) is derived from mouse thymus tissue and contains spatially matched RNA and ADT measurements. The SPOTS mouse spleen dataset (Ben-Chetrit et al., 2023) contains transcriptomic and proteomic information from mouse spleen tissue.

Baselines. To validate the effectiveness of the proposed method, we compared FlowMOS with several recent representative methods, including MOFA+ (Argelaguet et al., 2020), CiteFuse (Kim et al., 2020), MultiVI (Ashuach et al., 2023), STAGATE (Dong & Zhang, 2022), PAST (Li et al., 2023), SpatialGlue (Long et al., 2024), and PRAGA (Huang et al., 2025).

Metrics. We employed nine commonly used clustering evaluation metrics to assess the quality of the final unified latent representation learned, including mutual information (MI), normalized mutual information (NMI), adjusted mutual information (AMI), the Fowlkes-Mallows index (FMI), the adjusted Rand index (ARI), the V-measure, the F1 score, the Jaccard similarity coefficient (Jaccard), and completeness (Compl.).

3.2 Quantitative Performance Evaluation

To quantitatively evaluate the quality of the final unified latent representation, clustering experiments were conducted on datasets with reference annotations. As shown in Table 1, the proposed method maintained overall performance comparable to that of PRAGA on the human lymph node dataset, while achieving a more significant improvement on the mouse brain spatial epigenome transcriptome dataset. Specifically, compared to PRAGA, FlowMOS improved the ARI on the Spatial epigenome transcriptome mouse brain dataset from 27.06% to 32.62%, an increase of 5.56%; the FMI increased from 35.07% to 40.02%, an increase of 4.95%; the F1-Score increased from 35.02% to 39.94%, a 4.92% improvement; and the Jaccard coefficient increased from 21.23% to 24.97%, a 3.74% improvement. In addition, the NMI and AMI also increased by 0.97% and 1.04%, respectively. These results demonstrate that this framework is particularly effective at handling spatial multi-omics data with complex cross modal distributions.

3.3 Qualitative Analysis of Integrated Representations

To further evaluate the quality of the integrated representations learned by FlowMOS, we conducted a multi perspective qualitative analysis on Human lymph node and Spatial epigenome transcriptome mouse brain datasets. As shown in Figure 3, UMAP visualizations indicate that FlowMOS generated clusters with clearer structures and greater distinguishability on both the human lymph node and mouse brain datasets, while preserving meaningful relationships between adjacent regions. In the spatial graph, the spatial clustering results exhibit coherent and spatially continuous regions, indicating that the learned representations effectively preserve the spatial organizational structure.

Table 1: Quantitative experimental results are presented for nine metrics on the Human Lymph Node dataset and the Spatial epigenome transcriptome mouse brain dataset. Second best results are underlined, while the best results are highlighted in bold.

Methods	MI(%)	NMI(%)	AMI(%)	FMI(%)	V-Measure(%)	ARI(%)	F1-Score(%)	Jaccard(%)	Compl.(%)
Human Lymph Node dataset									
MOFA+	65.06	34.91	34.49	37.67	34.91	22.73	36.96	22.67	31.88
CiteFuse	43.35	23.34	22.86	27.57	23.38	12.51	26.15	15.04	26.88
MultiVI	12.03	7.01	6.43	26.16	7.01	3.73	26.15	15.04	7.13
STAGATE	48.45	21.25	21.03	22.36	21.25	12.21	22.36	12.59	21.76
PAST	58.82	33.60	33.14	41.42	33.60	24.64	41.41	26.11	32.42
SpatialGlue	66.52	36.07	35.65	39.16	36.07	23.83	38.79	24.06	33.24
PRAGA	**73.00**	**39.47**	**39.07**	**42.69**	**39.47**	**28.28**	**42.33**	**26.76**	**36.29**
FlowMOS	72.17 ± 1.05	39.04 ± 0.86	38.64 ± 0.86	42.12 ± 1.35	39.04 ± 0.86	27.76 ± 1.42	41.56 ± 1.47	26.24 ± 1.17	35.91 ± 1.08
Spatial epigenome transcriptome mouse brain dataset									
MOFA+	19.58	8.64	8.38	15.59	8.64	4.39	15.59	8.45	8.91
CiteFuse	47.48	19.46	19.05	17.96	19.46	8.24	17.89	9.82	16.63
MultiVI	17.88	8.47	8.22	18.12	8.47	3.81	17.58	9.63	9.46
STAGATE	1.42	0.79	0.12	20.83	0.79	0.22	20.73	11.56	0.74
PAST	69.49	29.13	28.76	24.54	29.13	14.63	24.54	13.99	28.50
SpatialGlue	95.54	37.83	37.53	33.78	37.83	26.33	33.01	19.77	35.14
PRAGA	95.55	39.37	39.06	35.07	39.37	27.06	35.02	21.23	37.88
FlowMOS	$\mathbf{95.85 \pm 0.96}$	$\mathbf{40.34 \pm 0.63}$	$\mathbf{40.10 \pm 0.63}$	$\mathbf{40.02 \pm 1.75}$	$\mathbf{40.34 \pm 0.63}$	$\mathbf{32.62 \pm 1.76}$	$\mathbf{39.94 \pm 1.80}$	$\mathbf{24.97 \pm 1.42}$	$\mathbf{39.61 \pm 0.94}$

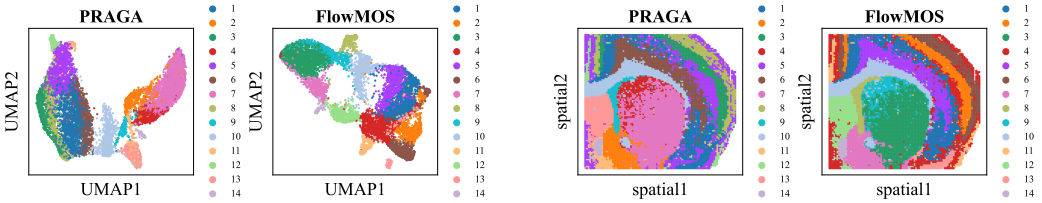


Figure 3: Visualization of qualitative experimental results on the Spatial epigenome transcriptome mouse brain dataset.

Table 2: Ablation study results for the proposed framework on the Human Lymph Node dataset, evaluating the contributions of the reconstruction loss ($\mathcal{L}_{recon}$), graph structure loss ($\mathcal{L}_g$), the mask consistency loss ($\mathcal{L}_{cons}$) and the flow matching loss ($\mathcal{L}_{fm}$). "w/o" denotes "without." The best performance is highlighted in bold.

Methods	MI(%)	NMI(%)	AMI(%)	FMI(%)	V-Measure(%)	ARI(%)	F1-Score(%)	Jaccard(%)	Compl.(%)
FlowMOS	**73.88**	**40.43**	**40.03**	**43.87**	**40.43**	**29.62**	**43.45**	**27.75**	**37.54**
w/o $\mathcal{L}_{\text{recon}}$	70.61	37.71	37.30	41.43	37.71	27.36	40.57	25.45	34.31
w/o $\mathcal{L}_{\text{g}}$	73.38	38.88	38.48	40.29	38.88	26.21	39.27	24.43	35.14
w/o $\mathcal{L}_{\text{cons}}$	73.46	40.27	39.87	43.51	40.27	29.07	43.14	27.50	37.46
w/o $\mathcal{L}_{\text{fm}}$	73.69	39.61	39.21	42.89	39.61	28.78	42.28	26.80	36.23

3.4 Ablation Study and Parameter Sensitivity Analysis

To further investigate the contributions of each objective function and FlowMOS's robustness to hyperparameter variations, we conducted ablation experiments and parameter sensitivity experiments on the Human Lymph Node dataset. As shown in Table 2, removing any single loss term results in a decline in performance across most evaluation metrics, indicating that these objective functions provide complementary supervision in the unified embedding representation learning process. The parameter sensitivity results in Figure 4 further demonstrate that FlowMOS maintains relatively stable performance across a wide range of parameter settings. The model performs stably within a moderate range of masking ratios, whereas excessively high masking ratios lead to a significant decline in clustering quality. The weights of the flow matching and consistency regularization terms also exhibit stable performance within a moderate range, but excessively high values degrade several clustering metrics. Furthermore, the model is generally robust to variations in the EMA decay coefficient β. Overall, these results indicate that FlowMOS is insensitive to minor hyperparameter changes and achieves reliable performance when the regularization strength and masking ratio are set within reasonable ranges.

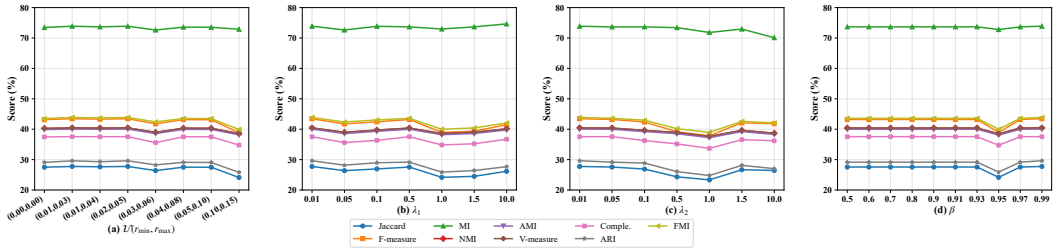


Figure 4: Parameter sensitivity analysis of different hyperparameters.

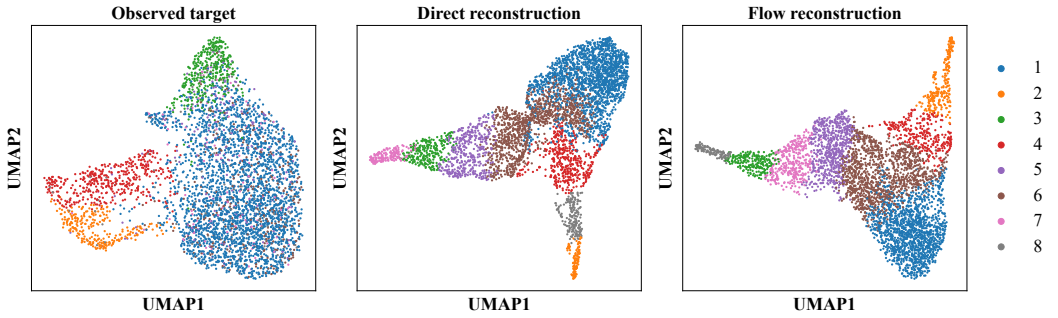


Figure 5: Latent transport analysis results on the Mouse Thymus Stereo-CITE-seq dataset.

3.5 Latent Transport Analysis

To further investigate whether the velocity field learned by FlowMOS captures meaningful cross-modal transformation information, we evaluated the representations obtained via latent transport through a cross modal recovery task. On the Mouse thymus stereo-CITE-seq dataset, RNA modality representations were propagated along the learned velocity field and subsequently decoded into the ADT target modality to obtain Flow-based reconstruction results. We compared the resulting representations with the observed target modality and with direct reconstruction results obtained without latent transport. As shown in Figure 5, the observed target modality exhibits distinct cluster specific structures in latent space, whereas direct reconstruction only partially recovered these structural patterns. This result indicates that the learned latent transport is not only an auxiliary regularization mechanism during training but also captures informative transformation directions that can be transferred to cross modal recovery, further validating the effectiveness of explicitly modeling continuous transitions between heterogeneous omics representations.

4 Conclusion

This paper proposes FlowMOS, a flow matching representation learning framework for spatial multi-omics integration. Based on flow matching, FlowMOS models the transformation from modality specific representations to a shared latent space as a continuous transfer process. Additionally, mask consistency regularization is introduced to enhance robustness against local feature loss and input perturbations. Experimental results demonstrate that FlowMOS's quantitative and qualitative performance across multiple spatial multi-omics datasets validates the potential of flow matching as an effective representation learning principle for integrating heterogeneous spatial multi-omics data.

Reproducibility Statement

We have made significant efforts to ensure the reproducibility of our work. Detailed descriptions of the FlowMOS architecture, training objectives, latent transport mechanism, and evaluation protocols are provided in Sections 2 and 3 of the main text, with additional implementation details, hyperparameter settings, and dataset information included in the Appendix. All experiments are conducted

on publicly available spatial multi-omics datasets. Upon acceptance, we will publicly release the FlowMOS codebase, preprocessing scripts, and experiment configurations to enable end-to-end reproduction of the reported results and facilitate further research on spatial multi-omics integration.

AI USE STATEMENT

Large language models, including ChatGPT, were only used during manuscript revision for language polishing. The authors reviewed and revised the AI-assisted content and take full responsibility for the research ideas, experimental design, data analysis, and scientific conclusions.

REFERENCES

Ricard Argelaguet, Damien Arnol, Danila Bredikhin, Yonatan Deloro, Britta Velten, John C Marioni, and Oliver Stegle. Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome biology*, 21(1):111, 2020.

Tal Ashuach, Mariano I Gabitto, Rohan V Koodli, Giuseppe-Antonio Saldi, Michael I Jordan, and Nir Yosef. Multivi: deep generative model for the integration of multimodal data. *Nature methods*, 20(8):1222–1231, 2023.

Nir Ben-Chetrit, Xiang Niu, Ariel D Swett, Jesus Sotelo, Maria S Jiao, Caitlin M Stewart, Catherine Potenski, Paulius Mielinis, Patrick Roelli, Marlon Stoeckius, et al. Integration of whole transcriptome spatial profiling with protein markers. *Nature biotechnology*, 41(6):788–793, 2023.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Yanxiang Deng, Marek Bartosovic, Petra Kukanja, Di Zhang, Yang Liu, Graham Su, Archibald Enninful, Zhiliang Bai, Gonçalo Castelo-Branco, and Rong Fan. Spatial-cut&tag: spatially resolved chromatin modification profiling at the cellular level. *Science*, 375(6581):681–686, 2022.

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.

Zhihua Du, Qiyi Chen, Weiliang Huang, Jinmiao Chen, and Xubin Zheng. Smart: spatial multi-omic aggregation using graph neural networks and metric learning. *Nature Communications*, 17 (1):2876, 2026.

Fiorella C Grandi, Hailey Modi, Lucas Kampman, and M Ryan Corces. Chromatin accessibility profiling by atac-seq. *Nature protocols*, 17(6):1518–1552, 2022.

Wan-Chen Hsieh, Bugi Ratno Budiarto, Yi-Fu Wang, Chih-Yu Lin, Mao-Chun Gwo, Dorothy Kazuno So, Yi-Shiuan Tzeng, and Shih-Yu Chen. Spatial multi-omics analyses of the tumor immune microenvironment. *Journal of Biomedical Science*, 29(1):96, 2022.

T Huang, T Liu, M Babadi, W Jin, and R Ying. Scalable generation of spatial transcriptomics from histology images via whole-slide flow matching. arxiv 2025. *arXiv preprint arXiv:2506.05361*.

Xinlei Huang, Zhiqi Ma, Dian Meng, Yanran Liu, Shiwei Ruan, Qingqiang Sun, Xubin Zheng, and Ziyue Qiao. Praga: prototype-aware graph adaptive aggregation for spatial multi-modal omics analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 326–333, 2025.

Guillaume Huguet, Daniel Sumner Magruder, Alexander Tong, Oluwadamilola Fasina, Manik Kuchroo, Guy Wolf, and Smita Krishnaswamy. Manifold interpolating optimal-transport flows for trajectory inference. *Advances in neural information processing systems*, 35:29705–29718, 2022.

Hani Jieun Kim, Yingxin Lin, Thomas A Geddes, Jean Yee Hwa Yang, and Pengyi Yang. Citefuse enables multi-modal analysis of cite-seq data. *Bioinformatics*, 36(14):4137–4143, 2020.

Zhen Li, Xiaoyang Chen, Xuegong Zhang, Rui Jiang, and Shengquan Chen. Latent feature extraction with a prior-based self-attention framework for spatial transcriptomics. *Genome Research*, 33(10):1757–1773, 2023.

Sha Liao, Yang Heng, Weiqing Liu, Jinqiong Xiang, Yong Ma, Ligen Chen, Xiuwen Feng, Dongmei Jia, Diyan Liang, Caili Huang, et al. Integrated spatial transcriptomic and proteomic analysis of fresh frozen tissue based on stereo-seq. *BioRxiv*, pp. 2023–04, 2023.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

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 multi-omics with spatialglue. *Nature methods*, 21(9):1658–1667, 2024.

Alicia A McDonough, Luciana C Veiras, Jacqueline N Minas, and Donna Lee Ralph. Considerations when quantitating protein abundance by immunoblot. *American Journal of Physiology-Cell Physiology*, 308(6):C426–C433, 2015.

Alex Morehead, Lazar Atanackovic, Akshata Hegde, Yanli Wang, Frimpong Boadu, Joel Selvaraj, Alexander Tong, Aditi Krishnapriyan, and Jianlin Cheng. Flow matching for generative modelling in bioinformatics and computational biology. *Nature Machine Intelligence*, pp. 1–18, 2026.

Santanu Subhash Rathod, Francesco Ceccarelli, Sean B Holden, Pietro Liò, Xiao Zhang, and Jovan Tanevski. Contextflow: Context-aware flow matching for trajectory inference from spatial omics data. *arXiv preprint arXiv:2510.02952*, 2025.

Yodai Takei, Yujing Yang, Jonathan White, Isabel N Goronzy, Jina Yun, Meera Prasad, Lincoln J Ombelets, Simone Schindler, Prashant Bhat, Mitchell Guttman, et al. Spatial multi-omics reveals cell-type-specific nuclear compartments. *Nature*, 641(8064):1037–1047, 2025.

Alexander Tong, Jessie Huang, Guy Wolf, David Van Dijk, and Smita Krishnaswamy. Trajectorynet: A dynamic optimal transport network for modeling cellular dynamics. In *International conference on machine learning*, pp. 9526–9536. PMLR, 2020.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *arXiv preprint arXiv:2302.00482*, 2023.

Sui Wang, Xinyu Wang, Qiangwei Peng, and Tiejun Li. Stitch: Spatial transcriptomics imputation via flow matching with internal learning. *bioRxiv*, pp. 2026–06, 2026.

Chengyang Zhang, Bo Li, Bob Zhang, Yuansong Zeng, Yuhao Yi, and Jiancheng Lv. Pssd: Progressive spatial-semantic decoupling for flow-based gene expression prediction from histology images. *Bioinformatics*, 42(8):btag598, 2026.

Di Zhang, Yanxiang Deng, Petra Kukanja, Eneritz Agirre, Marek Bartosovic, Mingze Dong, Cong Ma, Sai Ma, Graham Su, Shuozhen Bao, et al. Spatial epigenome–transcriptome co-profiling of mammalian tissues. *Nature*, 616(7955):113–122, 2023.

Jiawei Zhang, Kangjun Jin, Shuai Xiao, and Jiachen Yang. scchord: A probabilistic manifold rectification framework for rna-to-protein translation. In *Forty-third International Conference on Machine Learning*.

Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8(1):14049, 2017.

Yuansheng Zhou, Xue Xiao, Lei Dong, Chen Tang, Guanghua Xiao, and Lin Xu. Cooperative integration of spatially resolved multi-omics data with cosmos. *Nature communications*, 16(1): 27, 2025.

A Related works

A.1 Spatial multi-omics integration

Spatial multi-omics approaches aim to integrate molecular information from different modalities with spatial information. COSMOS (Zhou et al., 2025) employs graph neural networks to encode each omics modality separately and combine them using modality specific weights, subsequently obtaining a spatially coherent unified latent representation through contrastive learning and spatial regularization. PRAGA (Huang et al., 2025) introduces a learnable dynamic graph to mitigate the influence of fixed neighborhood graph structures and employs dynamic prototype contrastive learning to optimize the unified latent representation. SMART (Du et al., 2026) combines graph neural networks with metric learning to aggregate spatial and modality specific information into a unified latent representation. Although these methods have achieved good results, they primarily focus on how spatial and multi-omics information should be fused in the final latent representation. The transformation from modality specific representations to a unified latent representation has not been modeled as an explicit continuous process.

A.2 Flow Matching in Spatial Omics

Flow Matching (Lipman et al., 2022) has recently been introduced into spatial omics research to describe complex spatial variations and changes in molecular states (Morehead et al., 2026). STFlow (Huang et al.) utilizes Flow Matching to generate whole slide spatial transcriptomic expression from histological images. By learning conditional vector fields at the slide level, it can predict the joint gene expression across different spatial locations holistically. PSSD (Zhang et al., 2026) employs conditional Flow Matching to predict spatial gene expression from histological images and models expression differences between different spatial regions through the gradual decoupling of spatial and semantic information. ContextFlow (Rathod et al., 2025) incorporates spatial and ligand-receptor information into Flow Matching to infer trajectories of cellular state changes in longitudinal spatial omics data. STITCH (Wang et al., 2026) combines optimal transport with conditional Flow Matching to reconstruct missing spatial regions and further generate corresponding transcriptomic expression. These studies demonstrate that Flow Matching can effectively describe continuous change processes in spatial omics. However, current methods of this type primarily focus on spatial transcriptomics tasks. Unlike the above works, FlowMOS applies Flow Matching to spatial multi-omics representation learning, explicitly modeling the continuous transition from modality specific representations to a unified latent representation, and using this as a constraint in representation learning to further enhance the stability of the unified latent representation.

B Data Preprocessing

Because data from different omics modalities differ significantly in terms of dimensionality, sparsity, and statistical characteristics, each modality must be preprocessed separately before constructing the graph structure. Subsequently, the resulting low dimensional feature representations will serve as modality specific inputs for spatial multi-omics integration.

B.1 RNA Preprocessing

For the RNA modality, genes detected in fewer than 10 spatial spots were first excluded to reduce the impact of extremely sparse features. Subsequently, the Seurat v3 was used to identify highly variable genes, and the top 3000 highly variable genes were retained. The RNA count matrix was normalized by scaling the total expression count of each spot to 10,000, followed by a logarithmic transformation. Feature scaling was then applied where appropriate. Next, PCA analysis was performed on the selected highly variable genes to obtain a compact transcriptomic representation. In most RNA and protein datasets, the number of PCA dimensions was selected based on the corresponding protein modality, however, in the RNA and ATAC datasets, a fixed 50 dimensional PCA representation was used for the RNA modality.

B.2 Protein Preprocessing

For protein modalities (Antibody-Derived Tag (ADT) measurements), single cell based Centralized Log Ratio (CLR) normalization is applied to mitigate differences in protein abundance across different spatial observations. During quality control, low abundance protein features are removed as needed, and the normalized protein measurements are then standardized. PCA is subsequently applied to obtain a compact representation of protein features. The number of retained PCA dimensions is typically set to one less than the number of available protein features, thereby preserving most of the protein feature space while providing a low dimensional representation suitable for graph construction and subsequent representation learning.

B.3 ATAC Preprocessing

For the ATAC modality, chromatin accessibility spectra are represented using Latent Semantic Indexing (LSI), which is better suited to the high dimensional and sparse nature of accessibility data than traditional PCA. Prior to dimension reduction, ATAC observations are aligned with corresponding RNA observations based on common spatial spot identifiers to ensure precise correspondence between modalities after quality control. An LSI representation comprising 51 components is computed based on the ATAC feature matrix. The resulting LSI embedding is directly used as a low dimensional ATAC feature representation for subsequent graph construction.

C Implementation Details of FlowMOS

C.1 Hyperparameter Settings

The common hyperparameter settings used in FlowMOS are summarized in Table 3.

Table 3: Main implementation parameters of FlowMOS.

Parameter	Setting
Latent dimension	64
Initial graph K	10
KNN neighborhood size	20
Graph EMA coefficient α	0.9
λ_{cons}	0.01
Mask range $(r_{\min}, r_{\max})$	$(0.02, 0.05)$
Mask start epoch	10
λ_{fm}	0.01
Flow start epoch	10
Flow warm-up epochs	3
Flow hidden dimension	256
Time embedding dimension	32
Number of Flow blocks	2
Flow learning rate	1×10^{-4}
mclust PCA dimensions	20

Table 4: Dataset specific configurations used in the main experiments.

Dataset	Epochs	K_{cluster}	Modality weights	β
HLN	100	10	RNA:1, ADT:1	0.99
MB	300	14	RNA:1, ATAC:10	0.99

C.2 Training Process

The overall training process of FlowMOS is summarized in Algorithm 1.

Algorithm 1 Training Process of FlowMOS

Input: Preprocessed modality features $F^{(m)}$, spatial coordinates S

Construct the shared spatial adjacency graph G^S
Construct modality specific feature graphs $G_F^{(m)}$
Construct modality specific fused graphs $\widehat{G}^{(m)}$
for each training epoch **do**
  for each modality $m = 1, \ldots, M$ **do**
    Fuse spatial and feature graphs to obtain modality specific fused graphs $\widehat{G}^{(m)}$
    Encode $F^{(m)}$ and $\widehat{G}^{(m)}$ into $\mathcal{Z}^{(m)}$
  end for
  Fuse modality specific embeddings to obtain $\mathcal{Z}_{\text{shared}}$
  Decode $\mathcal{Z}_{shared}$ and compute reconstruction loss $\mathcal{L}_{\text{rec}}$
  Update the reference feature graph and compute graph structure loss $\mathcal{L}_g$
  if mask consistency is activated **then**
    Randomly mask modality features
    Encode masked features into $\widetilde{\mathcal{Z}}^{(m)}$
    Compute mask consistency loss $\mathcal{L}_{\text{cons}}$
  end if
  if Flow Matching is activated **then**
    Update $\mathcal{Z}_{shared}$ using EMA
    for each modality $m = 1, \ldots, M$ **do**
      Sample $t \sim \mathcal{U}(0, 1)$
      Construct $z_t^{(m)}$
      Predict velocity $v_{\theta_m}(z_t^{(m)}, t)$
    end for
    Compute flow matching loss $\mathcal{L}_{fm}$
  end if
  Compute

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \lambda_1 \mathcal{L}_{fm} + \lambda_2 \mathcal{L}_{cons} + \lambda_3 \mathcal{L}_g$$

  Update all trainable parameters
end for
return $\mathcal{Z}_{shared}$

D Evaluation Metrics

We employ nine commonly used clustering evaluation metrics to highlight the performance of our model, including Mutual Information (MI), Normalized Mutual Information (NMI), Adjusted Mutual Information (AMI), Fowlkes-Mallows Index (FMI), Adjusted Rand Index (ARI), Variation of Information Measure (V-Measure), F1-Score, Jaccard Similarity Coefficient (Jaccard), and Completeness (Compl.).

Let $Y = \{y_1, \ldots, y_n\}$ denote the ground truth labels and $\hat{Y} = \{\hat{y}_1, \ldots, \hat{y}_n\}$ denote the predicted labels for n samples. Let $C = \{C_1, \ldots, C_K\}$ and $\hat{C} = \{\hat{C}_1, \ldots, \hat{C}_{\hat{K}}\}$ represent the sets of ground truth and predicted clusters, respectively, then calculation formulas for the above metrics are as follows.

MI

MI measures the amount of shared information between Y and $\hat{Y}$

$$\text{MI}(Y, \hat{Y}) = \sum_{i=1}^{K} \sum_{j=1}^{\hat{K}} p(C_i, \hat{C}_j) \log \frac{p(C_i, \hat{C}_j)}{p(C_i)p(\hat{C}_j)}. \tag{13}$$

NMI

NMI normalizes MI to account for the scale of entropies

$$\mathrm{NMI}(Y, \hat{Y}) = \frac{2\,\mathrm{MI}(Y, \hat{Y})}{H(Y) + H(\hat{Y})}, \tag{14}$$

where the entropy is defined as

$$H(Y) = -\sum_{i=1}^{K} p(C_i) \log p(C_i). \tag{15}$$

AMI

AMI adjusts MI by correcting for chance agreement

$$\mathrm{AMI}(Y, \hat{Y}) = \frac{\mathrm{MI}(Y, \hat{Y}) - \mathbb{E}[\mathrm{MI}]}{\max\{H(Y), H(\hat{Y})\} - \mathbb{E}[\mathrm{MI}]}. \tag{16}$$

FMI

FMI evaluates clustering quality based on pairwise precision and recall

$$\mathrm{FMI} = \sqrt{\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \cdot \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}}. \tag{17}$$

ARI

ARI measures the similarity between two clusterings while correcting for random chance

$$\mathrm{ARI} = \frac{\sum_{ij} \binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}, \tag{18}$$

where $n_{ij} = |C_i \cap \hat{C}_j|$, $a_i = |C_i|$, and $b_j = |\hat{C}_j|$.

V-MEASURE

V-Measure is defined as the harmonic mean of Homogeneity and Completeness

$$\text{V-Measure} = \frac{2 \cdot \mathrm{Homogeneity} \cdot \mathrm{Completeness}}{\mathrm{Homogeneity} + \mathrm{Completeness}}. \tag{19}$$

Homogeneity is given by

$$\mathrm{Homogeneity} = 1 - \frac{H(Y \mid \hat{Y})}{H(Y)}. \tag{20}$$

F1-SCORE

The F1-score is computed based on pairwise precision and recall

$$\mathrm{F1} = \frac{2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}, \tag{21}$$

where

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \quad \mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{22}$$

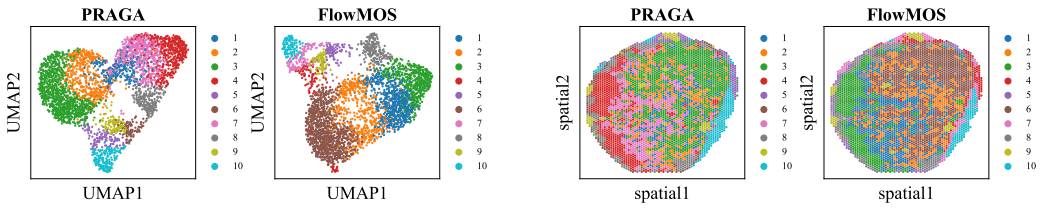


Figure 6: Visualization of qualitativeexp erimental results on the Human Lymph Node dataset.

JACCARD

The Jaccard Index measures the overlap between predicted and ground truth pairs

$$\text{Jaccard} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}. \tag{23}$$

COMPL.

Completeness evaluates whether samples from the same ground truth class are assigned to a single predicted cluster:

$$\text{Completeness} = 1 - \frac{H(\hat{Y} \mid Y)}{H(\hat{Y})}. \tag{24}$$

E ADDITIONAL EXPERIMENTAL RESULTS

To conduct a more comprehensive evaluation of FlowMOS, we further analyzed the learned unified latent representation from multiple perspectives, including UMAP visualizations of latent space organization, spatial graph visualizations, connectivity among clusters, the correspondence between clusters and ground truth labels, and cross modal reconstruction. These supplementary results complement the quantitative experiments presented in the main text and provide further evidence of the structural characteristics of the unified latent representation and the information captured by the learned Flow scenes.

E.1 ADDITIONAL VISUALIZATION ON THE HUMAN LYMPH NODE DATASET

Figure 6 shows the UMAP and spatial visualization results of the unified latent representations learned by the FlowMOS and PRAGA methods on the Human Lymph Node dataset. In the UMAP space, the clusters inferred by FlowMOS exhibit a clearer organizational structure, with greater compactness within classes while preserving transitions between closely related regions. When mapped back to the original tissue coordinates, the corresponding clusters form spatially coherent domains and preserve the overall organizational structure of the lymph node tissue. These results further demonstrate that FlowMOS has learned shared representations that better preserve both molecular distinguishability and spatial structural information simultaneously.

E.2 PAGA BASED CONNECTIVITY ANALYSIS

To further investigate the global relationships among clusters identified in a unified latent space, we performed a PAGA based connectivity analysis on the Human Lymph Node dataset. As shown in Figure 7, the generated graph summarizes the coarse grained connectivity among different inferred clusters and reveals structured relationships between adjacent regions. Multiple clusters are interconnected by relatively strong graph edges, suggesting that the learned representations preserve progressive relationships between spatially adjacent or molecularly related regions. These observations indicate that FlowMOS preserves not only the separability of local clusters but also higher order relationships between spatial domains.

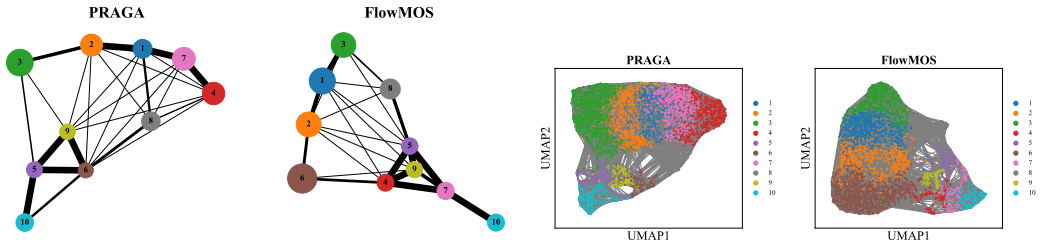


Figure 7: Cell trajectory inference results on the Human Lymph Node dataset.

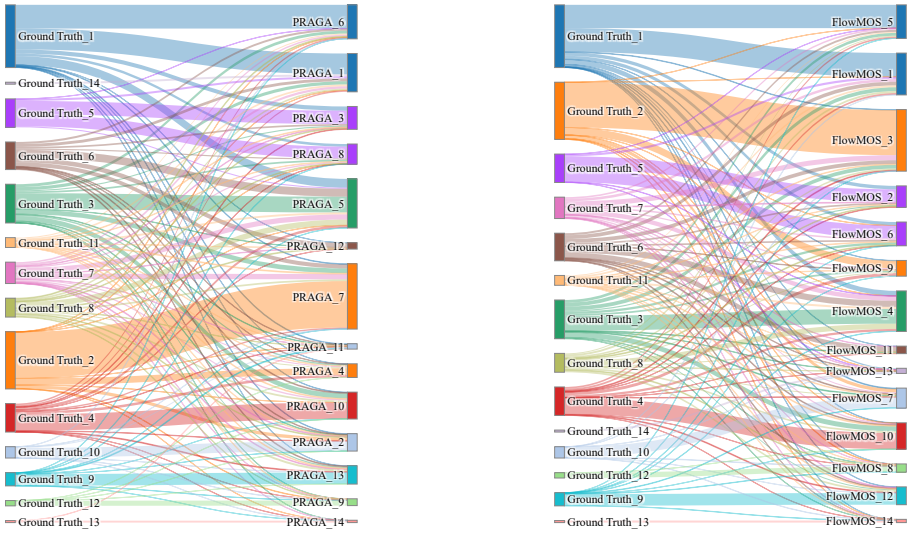


Figure 8: Sankey plot visualization on the Spatial epigenome transcriptome mouse brain dataset.

E.3 Cluster Correspondence Analysis

In addition, we used a Sankey diagram to compare the correspondence between the clustering results of FlowMOS and PRAGA on the Spatial epigenome transcriptome mouse brain dataset and the sample labels from the reference domains. As shown in Figure 8, the width of the flow lines intuitively illustrates the distribution of samples from each reference domain across the clusters identified by the FlowMOS representation. Compared to PRAGA, FlowMOS exhibits a highly concentrated correspondence between its primary reference groups and a small number of inferred clusters, whereas several closely related groups in PRAGA show partial overlap. This analysis complements numerical clustering metrics by clearly demonstrating the correspondence between reference labels and inferred partitions.

E.4 Additional Cross Modal Recovery on the SPOTS Dataset

To further investigate whether the learned flow field can capture transferable cross modal information across other datasets, we conducted additional recovery experiments on the SPOTS mouse spleen dataset. Starting from RNA specific latent representations, we used the learned velocity field to transport the source representations into a shared latent space, and then decoded them into ADT latent features.

Figure 9 compares the observed ADT features, the directly reconstructed ADT features, and the reconstructed ADT features obtained via Flow based latent transfer. The reconstructed features restore the primary structural organization of the observed protein modalities, and several cluster specific

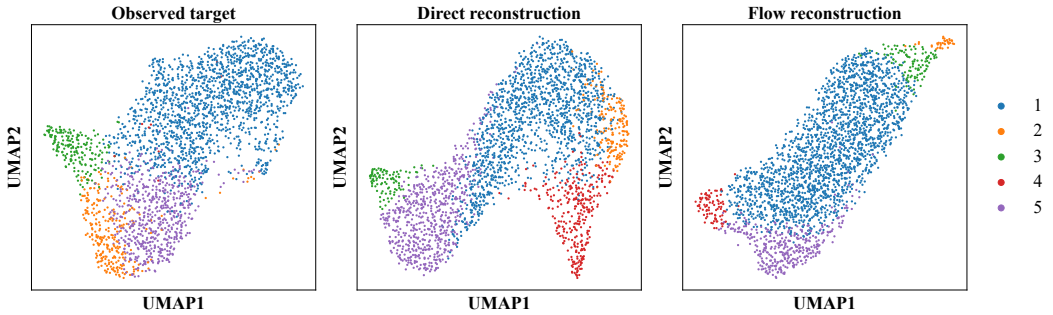


Figure 9: Latent transport analysis results on the SPOTS mouse spleen dataset.

patterns remain recognizable after cross modal reconstruction. This result further demonstrates that the learned velocity field captures meaningful information regarding the transformation from modal to shared space, and that such information can be transferred to cross-modal reconstruction on multi-omics datasets in independent spaces.

F LIMITATIONS

FlowMOS currently focuses on spatially matched multi-omics data and assumes that different modalities are measured at corresponding spatial locations. In addition, the continuous variable used in Flow Matching represents latent transport rather than biological time, and the current implementation adopts a relatively simple linear transport path. The additional Flow networks and ODE-based inference also introduce extra computational cost. Future work will extend FlowMOS to unpaired or partially paired spatial multi-omics data, explore more flexible transport paths, and evaluate the framework on larger datasets with more diverse modality combinations.