SPATIALLY AWARE NONLINEAR DIMENSION REDUCTION IMPROVES THE EFFICIENCY OF SPATIAL DOMAIN DETECTION IN SPATIAL MULTI-OMICS STUDIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial multi-omics technologies enable the joint characterization of multiple molecular modalities while preserving spatial context, but integrating heterogeneous nonlinear measurements with complex spatial dependencies remains challenging. Here, we introduce SpaLVM, a Bayesian latent variable model that jointly captures cross-modality variation and spatial dependencies through Gaussian process priors, enabling nonlinear dimensionality reduction of spatial multi-omics data. We further develop a sparse, structured variational inference scheme with inducing points for efficient and scalable computation, achieving an approximately 10-fold speedup over the original SpaLVM implementation. Across simulated and real datasets spanning diverse technologies and tissue structures, SpaLVM consistently outperforms existing spatial single-modal, spatial multimodal, non-spatial single-modal, and non-spatial multimodal methods in dimensionality reduction and spatial domain detection. SpaLVM provides a flexible and scalable framework for nonlinear integration and spatial domain detection in spatial multi-omics data.

1 INTRODUCTION

Spatial domain detection, the process of identifying distinct anatomical or functional regions based on spatial omics profiles, is essential for elucidating tissue architecture and cellular organization. Recent technological advances have enabled simultaneous profiling of multiple molecular layers, such as the genome, epigenome, transcriptome, and proteome, at single-cell resolution, offering unprecedented insight into cellular heterogeneity and function (Vandereyken et al., 2023). Among these, spatial multi-omics technologies add an essential dimension by preserving spatial context, enabling the study of how molecular programs are organized across tissue architectures. These integrative assays are transforming our understanding of development, disease progression, and tissue microenvironments by capturing both cell-intrinsic regulatory states and cell-extrinsic spatial cues Deng et al. (2023); Liu et al. (2023).

Despite their promise, spatial multi-omics datasets pose substantial analytical challenges. Each modality brings different noise profiles, resolution, and dynamic ranges, while spatial constraints introduce complex correlations among neighboring cells. Effective analysis thus requires integrative approaches that (i) reconcile heterogeneous nonlinear omics data and (ii) model complex spatial relationships in a principled manner. While early computational efforts have made strides in analyzing spatial transcriptomics or integrating two omics layers, comprehensive frameworks for scalable and flexible spatial multi-omics analysis remain lacking.

Several current approaches attempt to address these issues. Methods developed for non-spatial single-cell multi-omics integration, such as Seurat v4 (Hao et al., 2021), TotalVI (Gayoso et al., 2021) and scMDC (Lin et al., 2022), model modality-specific noise and variation, but typically ignore spatial context, treating cells as independent and exchangeable. Conversely, spatially informed tools like SpatialPCA (Shang & Zhou, 2022), STAGATE (Dong & Zhang, 2022), GraphST (Long et al., 2023), and SpaVAE (Tian et al., 2024) explicitly model spatial dependencies using Gaussian processes priors or graph neural networks, but are often designed for single-modality data, requiring additional modifications to accommodate multi-omics data.

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.

More recently, deep learning approaches (SpatialGlue (Long et al., 2024), COSMOS (Zhou et al., 2025), and MISO (Coleman et al., 2025)) and statistical models (MEFISTO (Velten et al., 2022) and SMOPCA (Chen et al., 2025)) have extended spatial modeling from single-modality to dual-modality or multi-modality settings. However, these methods are either inherently linear models or heavily reliant on linear dimensionality reduction as a preprocessing step. This reliance often leads to information loss and restricts their capacity to capture complex nonlinear patterns, ultimately compromising downstream performance (See **Appendix B:Related Work** for detailed descriptions and comparisons).

The Gaussian process latent variable model (GPLVM) (Lawrence, 2003; Lawrence & Hyvärinen, 2005) and its Bayesian extension (Titsias & Lawrence, 2010) offer a potential nonlinear, probabilistic alternative. GPLVM-based methods have been applied to scRNA-seq (nonspatial single modality) analysis (Campbell & Yau, 2015; Ahmed et al., 2019; Verma & Engelhardt, 2020; Lalchand et al., 2022; Zhao et al., 2024). However, these approaches cannot explicitly capture complex spatial dependencies or jointly model multiple modalities. More recently, Okamoto et al. extended GPLVM to analyze Spatial Transcriptomics (spatial single modality) by incorporating cell locations as fixed covariates in the latent space, yielding a semi-supervised GPLVM (ssGPLVM). In contrast to ssGPLVM, we extend the framework to multimodal settings and introduce GP priors over cell coordinates to capture spatial dependencies among neighboring cells within a fully Bayesian framework (See **Appendix B:Related Work** for detailed descriptions and comparisons).

In this article, we introduce a novel spatial multi-omics latent variable model, SpaLVM, which jointly models multiple modalities and spatial dependencies to learn a unified cell representation. By capturing cross-modality variation, SpaLVM enables the extraction of biologically meaningful signals while preserving the spatial structure and local relationships among tissue locations. The resulting latent representation can be directly paired with K-means clustering, facilitating more accurate spatial domain detection and downstream analysis. We make the following key contributions:

1. We propose a novel spatially aware nonlinear model SpaLVM for integrating and analyzing spatial multi-omics data.
2. We develop a sparse, structured variational inference scheme that employs inducing points for both the spatial Gaussian process priors and the modality-specific mapping GPs. This approach reduces the computational time by approximately 10-fold compared with the raw SpaLVM implementation, enabling faster and more scalable inference.
3. We show that SpaLVM consistently outperforms state-of-the-art methods for dimensionality reduction and spatial domain detection on both simulated and real spatial multi-omics datasets spanning diverse technologies and tissue structures.

2 METHOD

2.1 SPALVM MODEL

Consider a spatial multi-omics dataset with M modalities measured simultaneously at N cells (or spots). The expression matrix of the m-th modality is represented by $\boldsymbol{Y}_m = [\boldsymbol{y}_{m1}, \ldots, \boldsymbol{y}_{mD_m}] \in \mathbb{R}^{N \times D_m}$, where D_m denotes the number of features in this modality, and $\boldsymbol{y}_{md} \in \mathbb{R}^N$ represents the vector of observations for feature d. The corresponding spatial coordinates are $\boldsymbol{S} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N]^\top \in \mathbb{R}^{N \times P}$. Our objective is to conduct dimension reduction on the measurement matrices and learn a unified latent representation, denoted as $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^\top \in \mathbb{R}^{N \times L}$. The factor matrix $\boldsymbol{X}$ contains L factors and the ℓ-th column $\boldsymbol{x}_{:\ell}$ is a N-vector that represents the ℓ-th factor values across all N cells. Mathematically, for each modality m and feature d, we consider

$$\boldsymbol{y}_{md} = \boldsymbol{f}_{md}(\boldsymbol{X}) + \boldsymbol{\epsilon}_{md}, \qquad \boldsymbol{\epsilon}_{md} \sim \mathcal{N}(\boldsymbol{0}, \sigma_m^2 \boldsymbol{I}_N), \tag{1}$$

where $m \in \{1, \ldots, M\}$ and $d \in \{1, \ldots, D_m\}$. The noise variance σ_m^2 is shared across features within modality m.

Following Lawrence & Hyvärinen (2005), each feature is generated nonlinearly from the latent representation:

$$\boldsymbol{y}_{md} \mid \boldsymbol{f}_{md} \sim \mathcal{N}(\boldsymbol{y}_{md} \mid \boldsymbol{f}_{md}, \sigma_m^2 \boldsymbol{I}_N), \qquad \boldsymbol{f}_{md}(\boldsymbol{X}) \sim \mathcal{GP}(0, \boldsymbol{K}_{ff}^{(m)}), \tag{2}$$

where $\boldsymbol{K}_{ff}^{(m)} \in \mathbb{R}^{N\times N}$ controls the smoothness of the mapping, with entries $(K_{ff}^{(m)})_{ij} = k_f^{(m)}(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta}_f^{(m)})$. $k_f^{(m)}$ is the kernel function for modality m, with kernel parameters $\boldsymbol{\theta}_f^{(m)}$. For notational simplicity, we suppress kernel parameters in subsequent equations whenever they are clear from context.

We place a Gaussian process prior on each latent dimension over the observed spatial coordinates, inducing correlations among spatially proximate observations. Let $\boldsymbol{x}_{:\ell} \in \mathbb{R}^N$ denote the ℓ-th latent dimension across the N spatial locations. We assume

$$p(\boldsymbol{X} \mid \boldsymbol{S}) = \prod_{\ell=1}^{L} \mathcal{N}(\mathbf{x}_{:\ell} \mid \mathbf{0}, \boldsymbol{\Sigma}), \tag{3}$$

where $(\boldsymbol{\Sigma})_{ij} = k_x(\mathbf{s}_i, \mathbf{s}_j; \boldsymbol{\theta}_x)$, with $\boldsymbol{\theta}_x$ denoting the learnable kernel parameters.

Assuming features and modalities are conditionally independent given $\boldsymbol{X}$, the joint conditional distribution factorizes as

$$p(\boldsymbol{Y}_{1:M}, \boldsymbol{F}_{1:M}, \boldsymbol{X} \mid \boldsymbol{S}) = p(\boldsymbol{X} \mid \boldsymbol{S}) \prod_{m=1}^{M} \prod_{d=1}^{D_m} p(\boldsymbol{y}_{md} \mid \boldsymbol{f}_{md}) p(\boldsymbol{f}_{md} \mid \boldsymbol{X}), \tag{4}$$

where $\boldsymbol{Y}_{1:M} = \{\boldsymbol{Y}_m\}_{m=1}^M$, $\boldsymbol{F}_{1:M} = \{\boldsymbol{F}_m(\boldsymbol{X})\}_{m=1}^M$, and $\boldsymbol{F}_m(\boldsymbol{X}) = \{\boldsymbol{f}_{md}(\boldsymbol{X})\}_{d=1}^{D_m}$.

Equation (4) defines the generative model for the multi-omics data. Since optimizing the exact marginal likelihood is intractable, we introduce variational approximations to construct a tractable evidence lower bound (ELBO) based on modality-specific marginal likelihoods. The resulting learning objective is detailed in Section 2.4, with complete derivations provided in **Appendix A**.

2.2 Spatial variational inference

To approximate the posterior $p(\mathbf{X} \mid \mathbf{Y}_{1:M}, \mathbf{S})$, we formulate a structured Gaussian variational distribution $q(\mathbf{X})$ using a full-covariance fixed-point parameterization (Section 2.2.1), which retains the full spatial covariance across observations. To make it computationally tractable for large N, we further introduce a sparse inducing-variable parameterization (Section 2.2.2), which approximates the full spatial structure through a set of inducing locations and thereby reduces the time and memory costs.

2.2.1 Full-Covariance Fixed-Point Parameterization

It factorizes over latent dimensions while retaining the full spatial covariance across observations:

$$q(\mathbf{X}) = \prod_{\ell=1}^{L} \mathcal{N}(\mathbf{x}_{:\ell} \mid \mathbf{r}_\ell, \mathbf{V}_\ell). \tag{5}$$

Optimizing the $N \times N$ covariance matrices $\boldsymbol{V}_\ell$ directly would involve $O(N^2)$ variational parameters per latent dimension. To reduce the number of free variational parameters and accelerate convergence, we reparameterize $q(\mathbf{X})$ using the fixed-point form

$$\boldsymbol{V}_\ell = (\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Lambda}_\ell)^{-1}, \qquad \boldsymbol{r}_\ell = \boldsymbol{\Sigma}\bar{\boldsymbol{r}}_\ell, \tag{6}$$

where $\boldsymbol{\Lambda}_\ell = \mathrm{diag}(\boldsymbol{\lambda}_\ell)$ with $\boldsymbol{\lambda}_\ell \geq \mathbf{0}$. This diagonal precision-correction form is motivated by the stationary condition of the variational objective, while the non-negativity constraint ensures $0 \prec \boldsymbol{V}_\ell \preceq \boldsymbol{\Sigma}$ (See **Appendix A.7**). The variational parameters are thus reduced from the N^2 entries of $\boldsymbol{V}_\ell$ to the $2N$ parameters $(\bar{\boldsymbol{r}}_\ell, \boldsymbol{\lambda}_\ell)$. Under this parameterization, the spatial KL term

$$\mathcal{K}_X = D_{\mathrm{KL}}(q(\boldsymbol{X}) \| p(\boldsymbol{X} \mid \boldsymbol{S})) \tag{7}$$

has a closed-form expression since both distributions are Gaussian (**Appendix A.5**). This term regularizes the latent representation towards the spatial GP priors.

2.2.2 SPARSE INDUCING-VARIABLE PARAMETERIZATION

Although the fixed-point parameterization reduces the number of free variational parameters, it still requires computations involving the full spatial covariance matrix over the N observations, which becomes prohibitive when N is large.

To further improve scalability, we introduce a sparse inducing-variable parameterization that preserves the Gaussian form of $q(\mathbf{X})$ in equation (5). Specifically, we introduce J spatial inducing locations $\overline{\boldsymbol{S}} \in \mathbb{R}^{J\times P}$, with $J \ll N$, and define the inducing variables for each latent dimension as $\mathbf{h}_\ell = x_\ell(\overline{\boldsymbol{S}}) \in \mathbb{R}^J$.

Let $\boldsymbol{K}^x_{\bar{s}\bar{s}} = k_x(\overline{\boldsymbol{S}}, \overline{\boldsymbol{S}})$, $\boldsymbol{K}^x_{s\bar{s}} = k_x(\boldsymbol{S}, \overline{\boldsymbol{S}})$, and let $\boldsymbol{L}^x$ denote the Cholesky factor such that $\boldsymbol{K}^x_{\bar{s}\bar{s}} = \boldsymbol{L}^x(\boldsymbol{L}^x)^\top$. We whiten the inducing variables via the transformation $\mathbf{h}_\ell = \boldsymbol{L}^x\mathbf{v}_\ell$, where $p(\mathbf{v}_\ell) = \mathcal{N}(\mathbf{0}, \boldsymbol{I}_J)$. We then choose a Gaussian variational distribution $q(\mathbf{v}_\ell) = \mathcal{N}(\boldsymbol{\eta}_\ell, \mathrm{diag}(\boldsymbol{\omega}_\ell))$, with $\boldsymbol{\omega}_\ell > \mathbf{0}$.

Using the structured variational distribution $q(\mathbf{x}_{:\ell}, \mathbf{v}_\ell) = p(\mathbf{x}_{:\ell} \mid \mathbf{v}_\ell, \boldsymbol{S})q(\mathbf{v}_\ell)$, we define $\boldsymbol{B}^x = \boldsymbol{K}^x_{s\bar{s}}(\boldsymbol{L}^x)^{-\top}$, marginalizing $\mathbf{v}_\ell$ gives $q(\mathbf{x}_{:\ell}) = \mathcal{N}(\mathbf{x}_{:\ell} \mid \mathbf{r}_\ell, \mathbf{V}_\ell)$, where

$$\boldsymbol{r}_\ell = \boldsymbol{B}^x\boldsymbol{\eta}_\ell, \qquad \boldsymbol{V}_\ell = \boldsymbol{\Sigma} + \boldsymbol{B}^x\left[\mathrm{diag}(\boldsymbol{\omega}_\ell) - \boldsymbol{I}_J\right](\boldsymbol{B}^x)^\top. \tag{8}$$

Thus, the variational posterior retains the same Gaussian form as in equation (5), while $\mathbf{r}_\ell$ and $\mathbf{V}_\ell$ are now determined by the spatial variational parameters $\{\boldsymbol{\eta}_\ell, \boldsymbol{\omega}_\ell\}_{\ell=1}^L$.

Because $q(\mathbf{X})$ is fully determined by $\{q(\mathbf{v}_\ell)\}_{\ell=1}^L$, under this structured variational parameterization, the KL term reduces to

$$\mathcal{K}_X = \sum_{\ell=1}^{L} D_{\mathrm{KL}}\left(q(\mathbf{v}_\ell) \,\|\, p(\mathbf{v}_\ell)\right). \tag{9}$$

This term is available in closed form because both distributions are Gaussian. This sparse parameterization replaces the $O(LN^3)$ spatial cost of the full-covariance construction by operations involving $J \ll N$ as detailed in **Appendix A.8**, while preserving the spatial GP prior through the exact conditional $p(\mathbf{x}_{:\ell} \mid \mathbf{v}_\ell, \boldsymbol{S})$.

2.3 SPARSE MAPPING GP INFERENCE

To handle the complex dependence of $\boldsymbol{K}_{ff}$ on the latent factors in the nonlinear mapping, we employ a sparse Gaussian process variational inference framework. We therefore introduce T inducing inputs $\boldsymbol{Z} \in \mathbb{R}^{T\times L}$ and the corresponding inducing variable $\boldsymbol{u}_{md} = \boldsymbol{f}_{md}(\boldsymbol{Z}) \in \mathbb{R}^T$. The prior over $\boldsymbol{u}_{md}$ is $p(\boldsymbol{u}_{md}) = \mathcal{N}(\mathbf{0}, \boldsymbol{K}^{(m)}_{uu})$, where $(\boldsymbol{K}^{(m)}_{uu})_{ij} = k^{(m)}_f(\boldsymbol{z}_i, \boldsymbol{z}_j)$.

Denote $\boldsymbol{U}_m = \{\boldsymbol{u}_{md}\}_{d=1}^{D_m}$. For modality m, we use the augmented variational family

$$q(\boldsymbol{F}_m, \boldsymbol{U}_m, \boldsymbol{X}) = q(\boldsymbol{X})\prod_{d=1}^{D_m} p(\boldsymbol{f}_{md} \mid \boldsymbol{u}_{md}, \boldsymbol{X})q(\boldsymbol{u}_{md}), \tag{10}$$

where $q(\boldsymbol{u}_{md})$ is a Gaussian distribution.

Because $\boldsymbol{X}$ is random under $q(\boldsymbol{X})$, the resulting sparse-GP bound depends on expectations of kernel matrices with respect to the latent posterior. These expectations are summarized by the standard Ψ-statistics $\psi_0^{(m)} = \mathrm{tr}(\mathbb{E}_{q(\boldsymbol{X})}[\boldsymbol{K}^{(m)}_{ff}])$; $\boldsymbol{\Psi}_1^{(m)} = \mathbb{E}_{q(\boldsymbol{X})}[\boldsymbol{K}^{(m)}_{fu}]$; $\boldsymbol{\Psi}_2^{(m)} = \mathbb{E}_{q(\boldsymbol{X})}[\boldsymbol{K}^{(m)}_{uf}\boldsymbol{K}^{(m)}_{fu}]$, where $\boldsymbol{\Psi}_1^{(m)} \in \mathbb{R}^{N\times T}$ and $\boldsymbol{\Psi}_2^{(m)} \in \mathbb{R}^{T\times T}$.

For the Gaussian variational posterior and the ARD-RBF kernel $k_f^{(m)}$, these expectations admit analytic expressions; their closed forms are provided in **Appendix A.4**. They depend on $q(\boldsymbol{X})$ only through the posterior means and the marginal variances $(\boldsymbol{V}_\ell)_{nn}$.

The optimal variational distribution $q(\boldsymbol{u}_{md})$ can be eliminated analytically, yielding a collapsed feature-specific sparse GP term $\mathcal{H}_{md}$, which is the standard sparse GP lower bound for dimension

d. We define the modality-specific data-fit term as

$$\mathcal{H}_m = \sum_{d=1}^{D_m} \mathcal{H}_{md}, \tag{11}$$

which quantifies how well the shared latent representation explains modality m. Its closed form depends on $\psi_0^{(m)}$, $\mathbf{\Psi}_1^{(m)}$, $\mathbf{\Psi}_2^{(m)}$, $\boldsymbol{K}_{uu}^{(m)}$, σ_m^2, and $\boldsymbol{Y}_m$, and is given in **Appendix A.3**.

2.4 Multimodal objective and optimization

We now combine the spatial latent and mapping components to obtain the learning objective. Starting from the generative model in equation (4), the marginal likelihood for modality m is obtained by integrating out the latent representation and GP function values

$$p(\boldsymbol{Y}_m \mid \boldsymbol{S}) = \int p(\boldsymbol{Y}_m, \boldsymbol{F}_m, \boldsymbol{X} \mid \boldsymbol{S})\, d\boldsymbol{F}_m\, d\boldsymbol{X}. \tag{12}$$

Because the mapping covariance depends nonlinearly on the uncertain latent representation $\boldsymbol{X}$, this marginal likelihood is not available in closed form. Using the structured posterior introduced in Section 2.2 and the sparse-GP construction above, Jensen's inequality gives

$$\log p(\boldsymbol{Y}_m \mid \boldsymbol{S}) \geq \mathcal{H}_m - \mathcal{K}_X = \mathcal{L}_m. \tag{13}$$

More generally, we associate each modality with a fixed nonnegative weight w_m and define the multimodal evidence criterion

$$\mathcal{J} = \sum_{m=1}^{M} w_m \log p(\boldsymbol{Y}_m \mid \boldsymbol{S}). \tag{14}$$

Applying equation (13) to each modality yields the multimodal variational objective

$$\mathcal{L} = \sum_{m=1}^{M} w_m \mathcal{L}_m = \sum_{m=1}^{M} w_m \mathcal{H}_m - \left(\sum_{m=1}^{M} w_m \right) \mathcal{K}_X. \tag{15}$$

The modality weights act on the complete modality-specific variational bounds. Normalizing by the fixed constant $\sum_m w_m$ yields a weighted average of modality-specific data-fit terms with one shared spatial KL regularizer, preserving its relative strength as the number of modalities increases. This composite marginal-evidence objective differs from the standard joint-likelihood ELBO (**Appendix A.6**). Following Chen et al. (2025), We use $w_m = 1$ throughout.

We jointly maximize $\mathcal{L}$ using gradient-based optimization. Analytic gradients and computational complexity are provided in **Appendices A.9 – A.11**. After convergence, $\mathbb{E}_{q(\boldsymbol{X})}[\boldsymbol{X}] \in \mathbb{R}^{N \times L}$ is used as the integrated low-dimensional cell representation for downstream analysis. Implementation details are provided in **Appendix C.1**.

3 Experiments

We evaluated SpaLVM using two simulation studies and two real spatial multi-omics datasets. Simulation Study I (Scenarios 1–3), proposed by Townes & Engelhardt (2023); Long et al. (2024); Liu et al. (2025), synthesized transcriptomic and proteomic matrices mapped onto ten ggblocks patterns. Simulation Study II (Scenarios 4–5) constructed semi-synthetic benchmarks by placing real single-cell multi-omics samples into the same ten spatial patterns. Across both simulation studies, 10 independent replicates were generated per scenario using distinct random seeds. Lastly, we applied SpaLVM to two real-world datasets: the human tonsil dataset and the mouse embryonic brain dataset (E15.5), generated by Spatial-CITE-seq (RNA and protein) and MISAR-seq (RNA and ATAC), respectively. **Datasets and download links were summarized in Table S1.**

We benchmarked SpaLVM against nine methods, including five recently published spatial multi-omics approaches (COSMOS, MISO, SMOPCA, SpaVAE, and SpatialGlue), two state-of-the-art single-cell multi-omics methods (scMDC and Multigrate), and two widely used single-modality

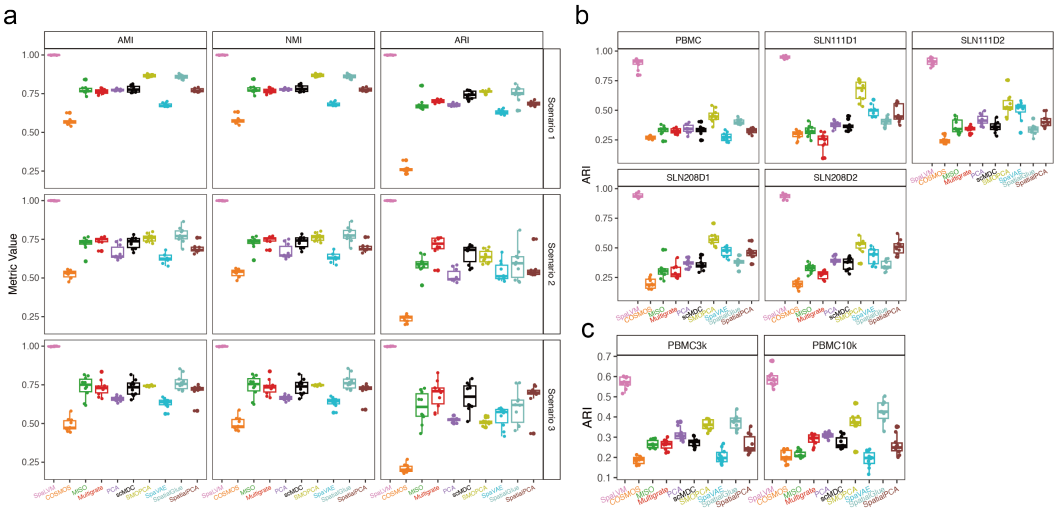

Figure 1: Simulation results. (a) Clustering results for Simulation Study I (Scenarios 1–3). (b) Clustering results on five CITE-seq datasets in Scenario 4. (c) Clustering results on two SMAGE-seq datasets in Scenario 5. AMI and NMI for Scenarios 4–5 are reported in **Figures S8–S9**.

baselines (PCA and SpatialPCA). Following Chen et al. (2025), we pre-processed and normalized each modality separately and concatenated them into a single input, on which SpatialPCA and PCA were then executed. **Table S2 provides the code and tutorial links for these competing methods.**

When ground-truth domain labels are available (as in Simulation Studies I and II, and the mouse embryonic brain dataset), we evaluate the spatial domain detection performance using Adjusted Mutual Information (AMI), Normalized Mutual Information (NMI), and the Adjusted Rand Index (ARI). When ground truth is absent (the human tonsil dataset), we instead report Moran's I and the Local Inverse Simpson's Index (LISI) to evaluate spatial coherence and smoothness. To further support the validity of the spatial domains identified by SpaLVM, we performed differential expression analysis, which highlighted distinct, biologically meaningful expression profiles across clusters. **Appendix C.3 provides detailed descriptions of all evaluation metrics.**

3.1 Simulation Study I

Detailed simulation procedures for Simulation Study I (Scenarios 1-3) were provided in Appendix C.4. Figures S1–S3 show example clustering results from SpaLVM and other methods for Scenario 1 – 3, respectively. **Figure 1(a)** summarizes the AMI, NMI, and ARI values for the ten methods across all three scenarios over the ten independent replicates. SpaLVM achieved the highest scores across all metrics in all three scenarios. Notably, its performance advantage persisted even as the number of shared patterns decreased from Scenario 1 to Scenario 3 (**Figure 1(a)**, **Figures S1–S3**), with SpaLVM consistently ranking first on average across the three metrics.

We further present the results in a unified manner in **Figures S4–S6**. It is evident that when two metrics are considered simultaneously, SpaLVM consistently maintains its superior position. Finally, we ranked all methods based on the median metric values across multiple runs for each scenario. The resulting average ranks and standard errors across the different scenarios are presented in **Figure S7**. Based on these average rankings, SpaLVM consistently outperforms competing methods.

3.2 Simulation Study II

Detailed simulation procedures for Simulation Study II (Scenarios 4-5) were provided in Appendix C.5. Figure 1(b)–(c) presents the ARI results of different methods on the five CITE-seq (Scenario 4) and two SMAGE-seq datasets (Scenario 5). To save space, only ARI is shown in the main text, while the complete AMI, NMI, and ARI results are provided in **Figures S8–S9**. Among

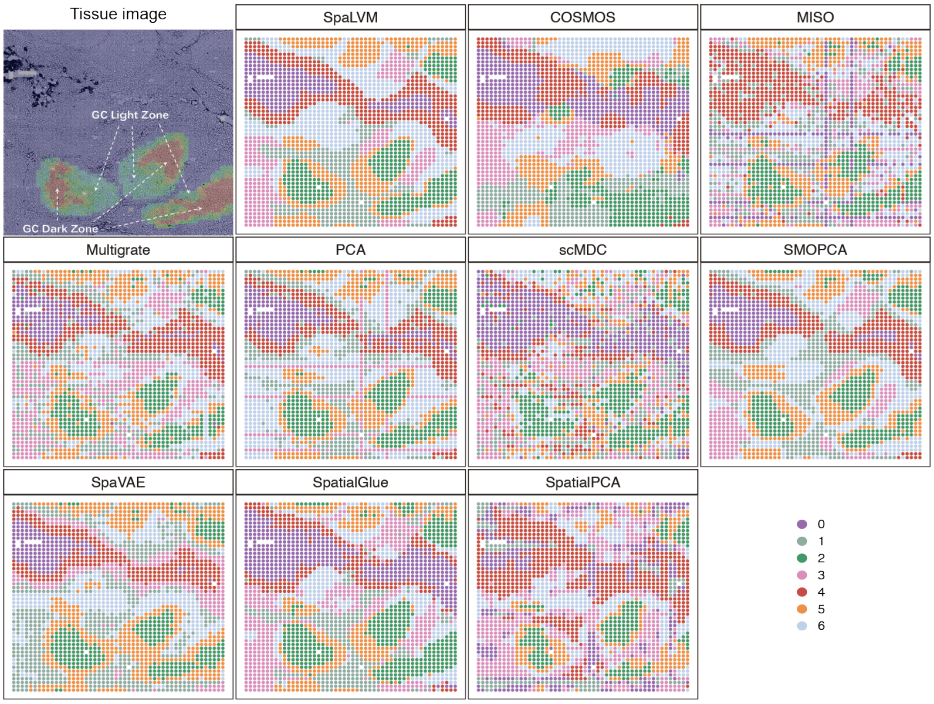

Figure 2: Spatial domains detected by SpaLVM and other methods on the human tonsil data.

these results, SpaLVM achieved the highest AMI, NMI, and ARI scores across all datasets, demonstrating its superiority over both deep learning-based and statistical methods.

Like in Scenarios 1–3, we also present the results in a unified manner in **Figures S10–S12** and **Figures S14–S16**. Furthermore, we rank the different methods based on the median metric values across multiple runs in **Figures S13 and S17**. **Figures S18–S24** show the latent representations learned by SpaLVM and different competing methods via UMAP, with colors denoting the ground-truth labels. It can be clearly seen that across all these evaluations, SpaLVM consistently outperforms all other approaches, indicating its advantages in modeling data with complex spatial structures.

3.3 Application to Spatial Transcriptome–Proteome Data

To demonstrate the capability in modeling real spatial multi-omics data, we applied SpaLVM to analyze a real spatial-CITE-seq data of human tonsil tissue (Liu et al., 2023). **The data information and preprocessing steps are summarized in Appendix C.6. Figure 2** shows the clustering results obtained by SpaLVM and the competing methods. Spatial clusters derived from SpaLVM demonstrate a strong alignment with tissue morphology and marker protein expressions (**Figure 2** and **Figure S25**). Following Liu et al. (2023), we approximately identify the clusters as follows: Cluster 4 represents the crypt epithelia; Clusters 5 and 2 correspond to the germinal center (GC) light and dark zones, respectively; Cluster 3 marks specific T-cell zones; Clusters 6 and 1 are localized within extrafollicular regions; and Cluster 0 represents peripheral blood cells. Quantitatively, SpaLVM achieved the highest Moran's I score, and its LISI was among the lowest across all ten evaluated methods (**Figure 3**). Although COSMOS produced a marginally lower LISI score, its clustering results are erroneous and biologically invalid. This inaccuracy is clearly demonstrated by the lack of alignment with the marker proteins established by Liu et al. (2023) (**Figure S25**).

We further conducted differential expression (DE) analyses using Seurat's FindAllMarkers function to identify cluster-specific markers (adjusted p-value < 0.05) based on the SpaLVM clusters. The top 10 most significant genes and proteins for each cluster were visualized in **Figure 4**. Notably, SpaLVM correctly identified all marker proteins mentioned by Liu et al. (2023). These markers span several biologically distinct categories: (1) surface proteins associated with B cells and follicular dendritic cells (CD19, CD21 and CD23); (2) functional protein markers such as immunoglobulins

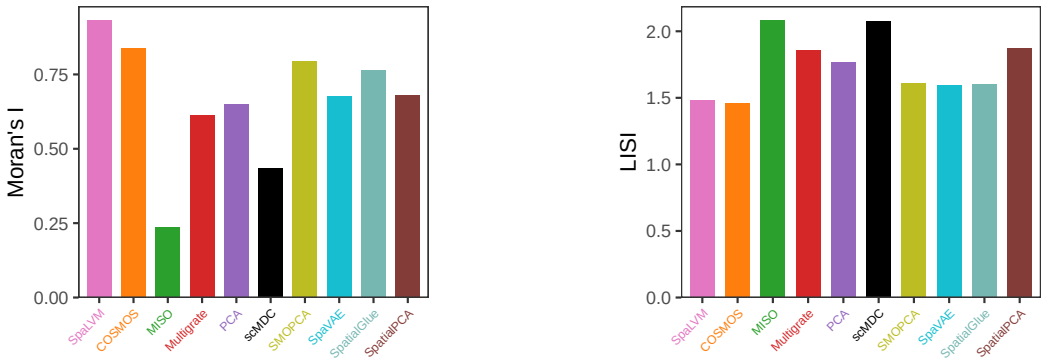

Figure 3: Spatial smoothness and domain coherence across different methods on the human tonsil data, measured by Moran's I (higher is smoother) and LISI (lower is smoother) scores.

showing spatially distinct distribution of GC B cells (IgM), matured B cells (IgG) and naive B cells (IgD), in agreement with B cell maturation, class switch and migration; (3) markers enriched in the extracellular region (CD90, Notch3) and crypt (Mac2); (4) T-cell markers (CD3, CD4, and CD45RA) that define specific T-cell zones and subtypes; and (5) additional key markers, including CD9, CD171, and CD32. These marker protein expressions are shown in **Figure S25**.

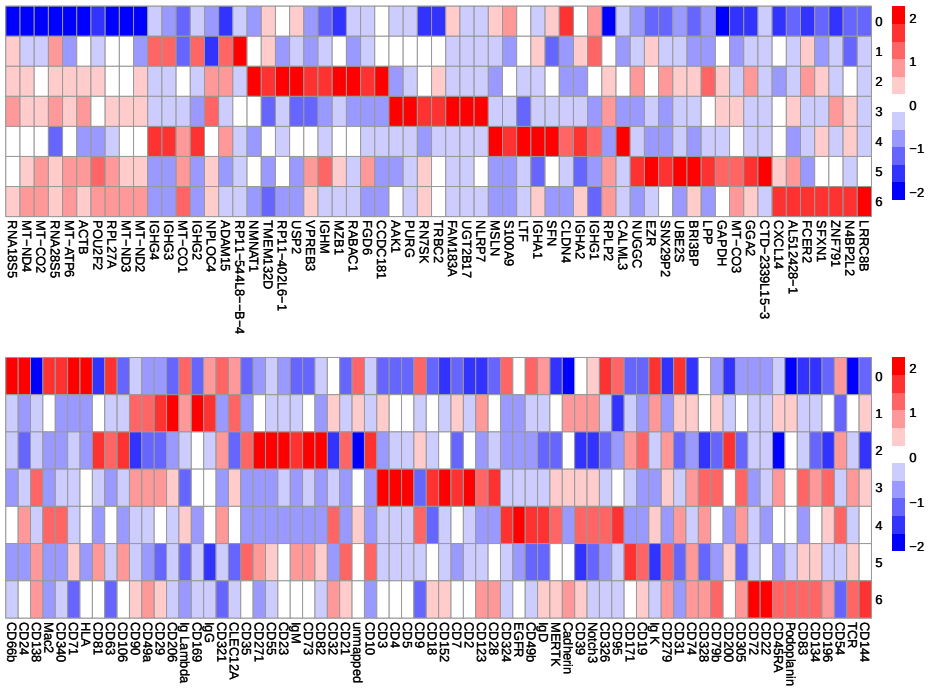

Figure 4: Differentially expressed genes (top) and surface proteins (bottom) across the SpaLVM detected domains for the human tonsil data. Rows are domains and columns are markers, and color shows the mean expression.

Six of the seven domains showed distinct marker genes (**Figure 4** (top) and **Figure 2**). The GC light-zone domain (Cluster 5) over-expressed NUGGC. NUGGC encodes a nuclear GTPase associated with GC B cells (Richter et al., 2009). The GC dark-zone domain (Cluster 2) over-expressed the immunoglobulin heavy-chain gene IGHM, together with MZB1. MZB1 supports immunoglobulin heavy-chain assembly and secretion in antibody-secreting cells (Rosenbaum et al.,

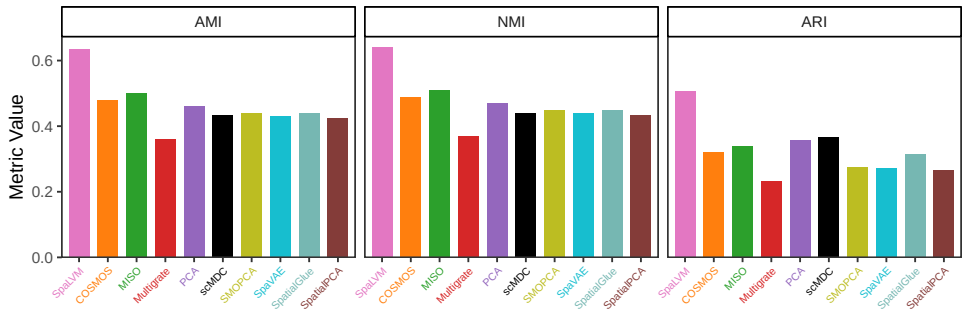

Figure 5: Clustering performance of different methods on the mouse embryonic (E15.5) brain data.

2014). The crypt-epithelial domain (Cluster 4) over-expressed MSLN and CLDN4, which encode the epithelial membrane proteins mesothelin and claudin-4. In the extrafollicular regions, Cluster 6 over-expressed FCER2 and CXCL14, whereas Cluster 1 over-expressed the immunoglobulin genes IGHG3, IGHG4, and IGHG2 . The T-cell zone (Cluster 3) showed elevated expression of TRBC2, which encodes the constant region of the T-cell receptor β chain, consistent with its T-cell-associated identity (Maciocia et al., 2017). Despite the lack of benchmark ground truth, the consistency across these evaluations reinforces the reliable performance and advantages of SpaLVM.

3.4 Application to Spatial Epigenomics–Transcriptomics Data

We further analyzed a real MISAR-seq (Jiang et al., 2023) dataset from the mouse embryonic brain at embryonic day 15.5 (E15.5), which jointly profiles chromatin accessibility (ATAC) and gene expression. **The data information and preprocessing steps are summarized in Appendix C.7. Figure 5** displays the results of SpaLVM along with the other competing methods. SpaLVM consistently outperformed competing methods across all evaluation metrics, achieving the highest AMI, NMI, and ARI scores. **Figures S26–S27** show the ground-truth labels and predicted clusters generated by the different methods. SpaLVM, SMOPCA and SpatialPCA, designed to capture spatial correlations among cells, yield notably smoother results than PCA (**Figures S26–S27**). We visualized the latent embeddings of each method using UMAP, with ground-truth labels indicated by different colors (**Figure S28**). The latent representation from SpaLVM effectively separates the different labels, whereas the Diencephalon, Hindbrain, and some other regions tend to be entangled in the embeddings produced by the other methods (e.g., COSMOS, MISO, Multigrate, SpaVAE and SpatialGlue). These results collectively highlight the clear advantages of SpaLVM in modeling spatial epigenomics and transcriptomics data.

3.5 Ablation Study and Runtime

To isolate the contributions of spatial modeling and multimodal integration, we compare the full SpaLVM with three ablated variants that use either only a single modality or remove the spatial prior. We put the ablation study results in **Tables S3–S5** and **Figures S29–S32**.

The proposed sparse spatial inducing-variable variational parameterization could greatly reduce the computational costs (See **Appendices A.11,C.2** and **Tables S6–S7**).

4 Conclusion

SpaLVM provides a powerful and flexible framework for dependency-aware nonlinear dimension reduction in spatial multi-omics analysis. By explicitly incorporating cell location information, SpaLVM preserves neighborhood structure in latent space while effectively capturing modality-specific nonlinear variation across multiple omics layers. Extensive experiments on both simulated and real spatial multi-omics datasets demonstrate that SpaLVM delivers superior performance compared to existing approaches, with enhanced efficiency and robustness, laying a solid foundation for advancing spatial multi-omics data analysis.

AI USE STATEMENT

In this work, we used generative AI tools to assist with the writing and presentation of the paper. Generative AI was not used to formulate mathematical claims, develop theoretical models or conceptual frameworks, provide critical ingredients for proving mathematical claims, design research methodology or experiments, implement the proposed method, or interpret experimental results.

Additionally, we used generative AI tools to edit software code and to improve the readability, grammar, and clarity of the manuscript.

All AI-assisted mathematical content was reviewed and verified by the authors. AI-assisted code edits were inspected and tested by the authors, and all AI-assisted manuscript edits were reviewed before inclusion in the final paper. We take responsibility for the final content of this work, including all text, mathematical statements, proofs, code, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide detailed information to facilitate reproducibility of the proposed method and the experimental results. The model formulation, spatial prior, variational inference procedure, and multimodal training objective are described in **Section 2**, with the spatial variational formulation and training objective detailed in **Sections 2.2.1 and 2.4**, respectively. Additional theoretical derivations are provided in **Appendix A**.

The data-generation procedures for Simulation Studies I and II are described in **Appendices C.4 and C.5**, respectively. The preprocessing procedures for the real spatial transcriptome–proteome and spatial epigenomics–transcriptomics datasets are provided in **Appendices C.6 and C.7**, respectively. Implementation details and the computational environment are provided in **Appendix C.1**, and runtime information is reported in **Appendix C.2**.

The sources of the experimental datasets and the implementations of the competing methods are summarized in Supplementary **Tables S1 and S2**, respectively. The corresponding supplementary experimental visualizations and result figures are provided in **Section E**.

REFERENCES

Sumon Ahmed, Magnus Rattray, and Alexis Boukouvalas. Grandprix: scaling up the bayesian gplvm for single-cell data. *Bioinformatics*, 35(1):47–54, 2019.

Ricard Argelaguet, Britta Velten, Damien Arnol, Sascha Dietrich, Thorsten Zenz, John C Marioni, Florian Buettner, Wolfgang Huber, and Oliver Stegle. Multi-omics factor analysis-a framework for unsupervised integration of multi-omics data sets. *Molecular systems biology*, 14(6):e8124, 2018.

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*, 21:1–17, 2020.

Kieran Campbell and Christopher Yau. Bayesian gaussian process latent variable models for pseudotime inference in single-cell rna-seq data. *bioRxiv*, pp. 026872, 2015.

Francesco Paolo Casale, Adrian Dalca, Luca Saglietti, Jennifer Listgarten, and Nicolo Fusi. Gaussian process prior variational autoencoders. *Advances in neural information processing systems*, 31, 2018.

Mo Chen, Ruihua Cheng, Jianuo He, Jun Chen, and Jie Zhang. Smopca: spatially aware dimension reduction integrating multi-omics improves the efficiency of spatial domain detection. *Genome Biology*, 26(1), 2025. ISSN 1474-760X. doi: 10.1186/s13059-025-03576-9.

Kyle Coleman, Amelia Schroeder, Melanie Loth, Daiwei Zhang, Jeong Hwan Park, Ji-Youn Sung, Niklas Blank, Alexis J. Cowan, Xuyu Qian, Jianfeng Chen, Jiahui Jiang, Hanying Yan, Laith Z. Samarah, Jean R. Clemenceau, Inyeop Jang, Minji Kim, Isabel Barnfather, Joshua D. Rabinowitz, Yanxiang Deng, Edward B. Lee, Alexander Lazar, Jianjun Gao, Emma E. Furth, Tae Hyun

Hwang, Linghua Wang, Christoph A. Thaiss, Jian Hu, and Mingyao Li. Resolving tissue complexity by multimodal spatial omics modeling with miso. *Nature Methods*, 22(3):530–538, 2025. ISSN 1548-7105. doi: 10.1038/s41592-024-02574-2.

Yanxiang Deng, Zhiliang Bai, and Rong Fan. Microtechnologies for single-cell and spatial multiomics. *Nature Reviews Bioengineering*, 1(10):769–784, 2023.

Van Hoan Do, Francisca Rojas Ringeling, and Stefan Canzar. Linear-time cluster ensembles of large-scale single-cell rna-seq and multimodal data. *Genome research*, 31(4):677–688, 2021.

Kangning Dong and Shihua Zhang. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nature Communications*, 13(1), 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-29439-6.

Adam Gayoso, Zoe Steier, Romain Lopez, Jeffrey Regier, Kristopher L Nazor, Aaron Streets, and Nir Yosef. Joint probabilistic modeling of single-cell multi-omic data with totalvi. *Nature methods*, 18(3):272–282, 2021.

Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck, Shiwei Zheng, Andrew Butler, Maddie J Lee, Aaron J Wilk, Charlotte Darby, Michael Zager, et al. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587, 2021.

Fuqing Jiang, Xin Zhou, Yingying Qian, Miao Zhu, Li Wang, Zhuxia Li, Qingmei Shen, Minhan Wang, Fangfang Qu, Guizhong Cui, et al. Simultaneous profiling of spatial gene expression and chromatin accessibility during mouse brain development. *Nature Methods*, 20(7):1048–1057, 2023.

Peng Jiang, Sishuo Chen, Xingye Wu, Juan Liu, and Tian Tian. Siva: diagonal integration of spatial multi-omics data via spatially informed variational autoencoders and anchor guidance. *Bioinformatics*, 42(Supplement_1):btag247, 2026.

Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods*, 16(12):1289–1296, 2019.

Vidhi Lalchand, Aditya Ravuri, Emma Dann, Natsuhiko Kumasaka, Dinithi Sumanaweera, Rik GH Lindeboom, Shaista Madad, Sarah A Teichmann, and Neil D Lawrence. Modelling technical and biological effects in scrna-seq data with scalable gplvms. *arXiv preprint arXiv:2209.06716*, 2022.

Neil Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. *Advances in neural information processing systems*, 16, 2003.

Neil Lawrence and Aapo Hyvärinen. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal of machine learning research*, 6(11), 2005.

Xiang Lin, Tian Tian, Zhi Wei, and Hakon Hakonarson. Clustering of single-cell multi-omics data with a multimodal deep learning method. *Nature communications*, 13(1):7705, 2022.

Yang Liu, Marcello DiStasio, Graham Su, Hiromitsu Asashima, Archibald Enninful, Xiaoyu Qin, Yanxiang Deng, Jungmin Nam, Fu Gao, Pino Bordignon, et al. High-plex protein and whole transcriptome co-mapping at cellular resolution with spatial cite-seq. *Nature Biotechnology*, 41 (10):1405–1409, 2023.

Yang Liu, Kexin Ma, Haoran Xu, Ke Xu, Yunfei Hu, Zhenhan Lin, Jiangli Lin, Bo Han, Shuaicheng Li, Zhixiang Lin, Xin Maizie Zhou, and Lu Zhang. Interpretable spatial multi-omics data integration and dimension reduction with spamv. *bioRxiv*, 2025. doi: 10.1101/2025.08.02.668264.

Yahui Long, Kok Siong Ang, Mengwei Li, Kian Long Kelvin Chong, Raman Sethi, Chengwei Zhong, Hang Xu, Zhiwei Ong, Karishma Sachaphibulkij, Ao Chen, Li Zeng, Huazhu Fu, Min Wu, Lina Hsiu Kim Lim, Longqi Liu, and Jinmiao Chen. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with graphst. *Nature Communications*, 14(1), 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-36796-3.

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.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

Paul M. Maciocia, Patrycja A. Wawrzyniecka, Brian Philip, Ida Ricciardelli, Ayse U. Akarca, Shimobi C. Onuoha, Mateusz Legut, David K. Cole, Andrew K. Sewell, Giuseppe Gritti, Joan Somja, Miguel A. Piris, Karl S. Peggs, David C. Linch, Teresa Marafioti, and Martin A. Pule. Targeting the t cell receptor β-chain constant region for immunotherapy of t cell malignancies. *Nature Medicine*, 23(12):1416–1423, 2017. doi: 10.1038/nm.4444.

Lauren Okamoto, Andrew Jones, Archit Verma, and Barbara E Engelhardt. Spatially-aware dimension reduction of transcriptomics data.

Giovanni Palla, Hannah Spitzer, Michal Klein, David Fischer, Anna Christina Schaar, Louis Benedikt Kuemmerle, Sergei Rybakov, Ignacio L Ibarra, Olle Holmberg, Isaac Virshup, et al. Squidpy: a scalable framework for spatial omics analysis. *Nature methods*, 19(2):171–178, 2022.

Kathleen Richter, Sukhdev Brar, Madhumita Ray, Prapaporn Pisitkun, Silvia Bolland, Laurent Verkoczy, and Marilyn Diaz. Speckled-like pattern in the germinal center (slip-gc), a nuclear gtpase expressed in activation-induced deaminase-expressing lymphomas and germinal center b cells. *Journal of Biological Chemistry*, 284(44):30652–30661, 2009. doi: 10.1074/jbc.M109.014506.

Marc Rosenbaum, Virginia Andreani, Tanya Kapoor, Simone Herp, Henrik Flach, Marlena Duchniewicz, and Rudolf Grosschedl. Mzb1 is a grp94 cochaperone that enables proper immunoglobulin heavy chain biosynthesis upon er stress. *Genes & Development*, 28(11):1165–1178, 2014. ISSN 1549-5477. doi: 10.1101/gad.240762.114.

Lulu Shang and Xiang Zhou. Spatially aware dimension reduction for spatial transcriptomics. *Nature communications*, 13(1):7203, 2022.

Shiquan Sun, Jiaqiang Zhu, and Xiang Zhou. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nature methods*, 17(2):193–200, 2020.

Tian Tian, Jie Zhang, Xiang Lin, Zhi Wei, and Hakon Hakonarson. Dependency-aware deep generative models for multitasking analysis of spatial omics data. *Nature Methods*, 21(8):1501–1513, 2024.

Michalis Titsias and Neil D Lawrence. Bayesian gaussian process latent variable model. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 844–851. JMLR Workshop and Conference Proceedings, 2010.

F William Townes and Barbara E Engelhardt. Nonnegative spatial factorization applied to spatial genomics. *Nature methods*, 20(2):229–238, 2023.

Katy Vandereyken, Alejandro Sifrim, Bernard Thienpont, and Thierry Voet. Methods and applications for single-cell and spatial multi-omics. *Nature Reviews Genetics*, 24(8):494–515, 2023.

Petar Velickovic, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax. *stat*, 1050:21, 2018.

Britta Velten, Jana M Braunger, Ricard Argelaguet, Damien Arnol, Jakob Wirbel, Danila Bredikhin, Georg Zeller, and Oliver Stegle. Identifying temporal and spatial patterns of variation from multimodal data using mefisto. *Nature methods*, 19(2):179–186, 2022.

Archit Verma and Barbara E Engelhardt. A robust nonlinear low-dimensional manifold for single cell rna-seq data. *BMC bioinformatics*, 21(1):324, 2020.

Sarah Zhao, Aditya Ravuri, Vidhi Lalchand, and Neil D Lawrence. Scalable amortized gplvms for single cell transcriptomics data. *arXiv preprint arXiv:2405.03879*, 2024.

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), 2025. ISSN 2041-1723. doi: 10.1038/s41467-024-55204-y.

A Detailed Variational Derivation and Complexity

This appendix expands the derivations summarised in Section 2. Appendix A.1 fixes the kernels and auxiliary notation; Appendix A.2 derives the per-modality bound of equation (13); Appendix A.3 gives the closed form of the collapsed mapping term $\mathcal{H}_{md}$ and $\mathcal{H}_m$; Appendix A.4 gives the analytic Ψ-statistics; Appendix A.5 gives the spatial KL term for the fixed-point variational parameterization; Appendix A.6 discusses the multimodal objective; Appendices A.7 and A.8 derive the fixed-point and sparse spatial inducing-variable variational parameterizations, respectively; Appendices A.9 and A.10 give their corresponding analytic gradients; and Appendix A.11 compares their time and memory complexity. All symbols follow the main text.

A.1 Kernels and Auxiliary Notation

The spatial prior of equation (3) uses the covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{N\times N}$ with $(\boldsymbol{\Sigma})_{ij} = k_x(\boldsymbol{s}_i, \boldsymbol{s}_j)$, shared by all L latent dimensions and parameterized by $\boldsymbol{\theta}_x$. Each modality carries its own mapping kernel $k_f^{(m)}$, taken to be the ARD-RBF kernel

$$k_f^{(m)}(\boldsymbol{x}, \boldsymbol{x}') = \sigma_{f,m}^2 \exp\left(-\frac{1}{2}\sum_{\ell=1}^{L} b_\ell^{(m)}(x_\ell - x'_\ell)^2\right), \qquad \boldsymbol{\theta}_f^{(m)} = \{\sigma_{f,m}^2, b_1^{(m)}, \dots, b_L^{(m)}\}, \quad (16)$$

with non-negative inverse squared length-scales $b_\ell^{(m)}$. The associated kernel matrices are $\boldsymbol{K}_{ff}^{(m)} = k_f^{(m)}(\boldsymbol{X}, \boldsymbol{X}) \in \mathbb{R}^{N\times N}$, $\boldsymbol{K}_{fu}^{(m)} = k_f^{(m)}(\boldsymbol{X}, \boldsymbol{Z}) \in \mathbb{R}^{N\times T}$, $\boldsymbol{K}_{uu}^{(m)} = k_f^{(m)}(\boldsymbol{Z}, \boldsymbol{Z}) \in \mathbb{R}^{T\times T}$, and $\boldsymbol{K}_{uf}^{(m)} = (\boldsymbol{K}_{fu}^{(m)})^\top$. The inducing inputs $\boldsymbol{Z}$ are shared across modalities, while the inducing variables $\boldsymbol{u}_{md} = \boldsymbol{f}_{md}(\boldsymbol{Z})$ are feature-specific.

It is convenient to introduce, for each modality,

$$\boldsymbol{A}_m = \boldsymbol{\Psi}_2^{(m)} + \sigma_m^2 \boldsymbol{K}_{uu}^{(m)} \in \mathbb{R}^{T\times T}, \qquad \boldsymbol{B}_m = \frac{1}{\sigma_m^2}\left(\boldsymbol{I}_N - \boldsymbol{\Psi}_1^{(m)} \boldsymbol{A}_m^{-1} (\boldsymbol{\Psi}_1^{(m)})^\top\right) \in \mathbb{R}^{N\times N}. \quad (17)$$

We also write $v_{\ell n} = (\boldsymbol{V}_\ell)_{nn}$ for the marginal posterior variance of latent dimension ℓ at location n, and $r_{\ell n} = (\boldsymbol{r}_\ell)_n$ for the corresponding posterior mean. Throughout, we fix a modality m and drop the superscript (m) whenever it is unambiguous.

A.2 Per-Modality Evidence Lower Bound

Fix a modality m, we first augment the model of equation (4) with the inducing variables $\boldsymbol{U}_m$. Because $\boldsymbol{u}_{md} = \boldsymbol{f}_{md}(\boldsymbol{Z})$ are function values of the same GP, $\int p(\boldsymbol{f}_{md} \mid \boldsymbol{u}_{md}, \boldsymbol{X}) p(\boldsymbol{u}_{md})\, d\boldsymbol{u}_{md} = p(\boldsymbol{f}_{md} \mid \boldsymbol{X})$, so the augmentation leaves the marginal likelihood of equation (12) unchanged. Applying Jensen's inequality to the augmented model gives

$$\begin{aligned}\log p(\boldsymbol{Y}_m \mid \boldsymbol{S}) &= \log \int q(\boldsymbol{F}_m, \boldsymbol{U}_m, \boldsymbol{X}) \frac{p(\boldsymbol{Y}_m, \boldsymbol{F}_m, \boldsymbol{U}_m, \boldsymbol{X} \mid \boldsymbol{S})}{q(\boldsymbol{F}_m, \boldsymbol{U}_m, \boldsymbol{X})}\, d\boldsymbol{F}_m\, d\boldsymbol{U}_m\, d\boldsymbol{X} \\ &\geq \int q(\boldsymbol{F}_m, \boldsymbol{U}_m, \boldsymbol{X}) \log \frac{p(\boldsymbol{Y}_m, \boldsymbol{F}_m, \boldsymbol{U}_m, \boldsymbol{X} \mid \boldsymbol{S})}{q(\boldsymbol{F}_m, \boldsymbol{U}_m, \boldsymbol{X})}\, d\boldsymbol{F}_m\, d\boldsymbol{U}_m\, d\boldsymbol{X}. \end{aligned} \quad (18)$$

The augmented variational family of equation (10) is chosen so that the conditional factors $p(\boldsymbol{f}_{md} \mid \boldsymbol{u}_{md}, \boldsymbol{X})$ appear identically in the numerator and the denominator of equation (18) and cancel exactly, which is what makes the bound tractable. Using the factorization across features and separating the term that involves $q(\boldsymbol{X})$ alone yields

$$\log p(\boldsymbol{Y}_m \mid \boldsymbol{S}) \geq \sum_{d=1}^{D_m} \mathcal{H}_{md} - D_{\mathrm{KL}}(q(\boldsymbol{X}) \| p(\boldsymbol{X} \mid \boldsymbol{S})) = \mathcal{H}_m - \mathcal{K}_X = \mathcal{L}_m, \quad (19)$$

which is equation (13), with the feature-specific term

$$\mathcal{H}_{md} = \int q(\boldsymbol{X})\, \mathbb{E}_{q(\boldsymbol{f}_{md}, \boldsymbol{u}_{md})}\left[\log \frac{p(\boldsymbol{y}_{md} \mid \boldsymbol{f}_{md})\, p(\boldsymbol{u}_{md})}{q(\boldsymbol{u}_{md})}\right] d\boldsymbol{X}. \tag{20}$$

A.3 Closed Form of the Collapsed Mapping Term

We evaluate equation (20) in four steps.

Step 1: expectation over the mapping GP given the inducing variables. Conditioned on $\boldsymbol{X}$ and $\boldsymbol{u}_{md}$, the sparse GP construction gives $p(\boldsymbol{f}_{md} \mid \boldsymbol{u}_{md}, \boldsymbol{X}) = \mathcal{N}(\boldsymbol{\alpha}_{md}, \boldsymbol{C}_{f|u})$ with

$$\boldsymbol{\alpha}_{md} = \boldsymbol{K}_{fu}\boldsymbol{K}_{uu}^{-1}\boldsymbol{u}_{md}, \qquad \boldsymbol{C}_{f|u} = \boldsymbol{K}_{ff} - \boldsymbol{K}_{fu}\boldsymbol{K}_{uu}^{-1}\boldsymbol{K}_{uf}, \tag{21}$$

where $\boldsymbol{C}_{f|u}$ does not depend on d. Because $p(\boldsymbol{y}_{md} \mid \boldsymbol{f}_{md})$ is Gaussian with isotropic noise, standard Gaussian identities give

$$\mathbb{E}_{p(\boldsymbol{f}_{md} \mid \boldsymbol{u}_{md}, \boldsymbol{X})}\big[\log p(\boldsymbol{y}_{md} \mid \boldsymbol{f}_{md})\big] = \log \mathcal{N}(\boldsymbol{y}_{md} \mid \boldsymbol{\alpha}_{md}, \sigma_m^2 \boldsymbol{I}_N) - \frac{1}{2\sigma_m^2}\operatorname{tr}(\boldsymbol{C}_{f|u}), \tag{22}$$

so that equation (20) becomes

$$\begin{aligned}\mathcal{H}_{md} = \int q(\boldsymbol{X})\Big(\mathbb{E}_{q(\boldsymbol{u}_{md})}\big[\log \mathcal{N}(\boldsymbol{y}_{md} \mid \boldsymbol{\alpha}_{md}, \sigma_m^2 \boldsymbol{I}_N)\big] - \frac{1}{2\sigma_m^2}\operatorname{tr}(\boldsymbol{C}_{f|u})\Big) d\boldsymbol{X} \\ - D_{\mathrm{KL}}(q(\boldsymbol{u}_{md}) \| p(\boldsymbol{u}_{md})).\end{aligned} \tag{23}$$

Step 2: expectation over the latent representation. Only the trace term and the quadratic form in $\boldsymbol{\alpha}_{md}$ depend on $\boldsymbol{X}$, and both depend on it only through the kernel expectations that define the Ψ-statistics. In particular,

$$\mathbb{E}_{q(\boldsymbol{X})}\big[\operatorname{tr}(\boldsymbol{C}_{f|u})\big] = \psi_0^{(m)} - \operatorname{tr}\big((\boldsymbol{K}_{uu}^{(m)})^{-1}\boldsymbol{\Psi}_2^{(m)}\big). \tag{24}$$

Step 3: collapsing the inducing distribution. The remaining dependence on $q(\boldsymbol{u}_{md})$ in equation (23) has the form $\mathbb{E}_{q(u)}[\log g(\boldsymbol{u}_{md})] - D_{\mathrm{KL}}(q(\boldsymbol{u}_{md}) \| p(\boldsymbol{u}_{md}))$ with $\log g(\boldsymbol{u}_{md}) = \mathbb{E}_{q(\boldsymbol{X})}[\log \mathcal{N}(\boldsymbol{y}_{md} \mid \boldsymbol{\alpha}_{md}, \sigma_m^2 \boldsymbol{I}_N)]$. This expression is maximised at $q^\star(\boldsymbol{u}_{md}) \propto g(\boldsymbol{u}_{md})p(\boldsymbol{u}_{md})$, at which point it equals $\log \int g(\boldsymbol{u}_{md})p(\boldsymbol{u}_{md})\, d\boldsymbol{u}_{md}$. Substituting this optimum eliminates $q(\boldsymbol{u}_{md})$ analytically; as in the main text we write $\mathcal{H}_{md}$ for the resulting collapsed value, so that

$$\begin{aligned}\mathcal{H}_{md} = \log \int \exp\big\{\mathbb{E}_{q(\boldsymbol{X})}[\log \mathcal{N}(\boldsymbol{y}_{md} \mid \boldsymbol{\alpha}_{md}, \sigma_m^2 \boldsymbol{I}_N)]\big\} p(\boldsymbol{u}_{md})\, d\boldsymbol{u}_{md} \\ - \frac{\psi_0^{(m)} - \operatorname{tr}((\boldsymbol{K}_{uu}^{(m)})^{-1}\boldsymbol{\Psi}_2^{(m)})}{2\sigma_m^2}.\end{aligned} \tag{25}$$

Step 4: the remaining Gaussian integral. Expanding the quadratic form and taking the expectation under $q(\boldsymbol{X})$,

$$\begin{aligned}\mathbb{E}_{q(\boldsymbol{X})}\big[\log \mathcal{N}(\boldsymbol{y}_{md} \mid \boldsymbol{\alpha}_{md}, \sigma_m^2 \boldsymbol{I}_N)\big] = -\frac{N}{2}\log(2\pi\sigma_m^2) - \frac{1}{2\sigma_m^2}\Big(\boldsymbol{y}_{md}^\top \boldsymbol{y}_{md} - 2\boldsymbol{y}_{md}^\top \boldsymbol{\Psi}_1 \boldsymbol{K}_{uu}^{-1}\boldsymbol{u}_{md} \\ + \boldsymbol{u}_{md}^\top \boldsymbol{K}_{uu}^{-1}\boldsymbol{\Psi}_2 \boldsymbol{K}_{uu}^{-1}\boldsymbol{u}_{md}\Big),\end{aligned} \tag{26}$$

which is quadratic in $\boldsymbol{u}_{md}$. Combining it with the Gaussian prior $p(\boldsymbol{u}_{md}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{K}_{uu})$ leaves a Gaussian integral with precision and linear coefficient

$$\boldsymbol{P} = \boldsymbol{K}_{uu}^{-1}\Big(\tfrac{1}{\sigma_m^2}\boldsymbol{\Psi}_2 + \boldsymbol{K}_{uu}\Big)\boldsymbol{K}_{uu}^{-1}, \qquad \boldsymbol{b} = \frac{1}{\sigma_m^2}\boldsymbol{K}_{uu}^{-1}\boldsymbol{\Psi}_1^\top \boldsymbol{y}_{md}, \tag{27}$$

and evaluating it with $\int \exp\{-\tfrac{1}{2}\boldsymbol{u}^\top \boldsymbol{P}\boldsymbol{u} + \boldsymbol{b}^\top \boldsymbol{u}\} d\boldsymbol{u} = (2\pi)^{T/2}|\boldsymbol{P}|^{-1/2}\exp\{\tfrac{1}{2}\boldsymbol{b}^\top \boldsymbol{P}^{-1}\boldsymbol{b}\}$ gives

$$\begin{aligned}\int \exp\big\{\mathbb{E}_{q(\boldsymbol{X})}[\log \mathcal{N}(\boldsymbol{y}_{md} \mid \boldsymbol{\alpha}_{md}, \sigma_m^2 \boldsymbol{I}_N)]\big\} p(\boldsymbol{u}_{md})\, d\boldsymbol{u}_{md} = \frac{|\boldsymbol{K}_{uu}|^{1/2}}{(2\pi\sigma_m^2)^{N/2}\big|\frac{1}{\sigma_m^2}\boldsymbol{\Psi}_2 + \boldsymbol{K}_{uu}\big|^{1/2}} \\ \times \exp\Big\{-\tfrac{1}{2}\boldsymbol{y}_{md}^\top \boldsymbol{B}_m \boldsymbol{y}_{md}\Big\},\end{aligned} \tag{28}$$

with $\boldsymbol{B}_m$ as in equation (17). Substituting equation (28) into equation (25) yields the closed form

$$\mathcal{H}_{md} = -\frac{N}{2}\log(2\pi\sigma_m^2) + \frac{1}{2}\log|\boldsymbol{K}_{uu}^{(m)}| - \frac{1}{2}\log\left|\tfrac{1}{\sigma_m^2}\boldsymbol{\Psi}_2^{(m)} + \boldsymbol{K}_{uu}^{(m)}\right| - \frac{1}{2}\boldsymbol{y}_{md}^\top \boldsymbol{B}_m \boldsymbol{y}_{md} - \frac{\psi_0^{(m)} - \operatorname{tr}((\boldsymbol{K}_{uu}^{(m)})^{-1}\boldsymbol{\Psi}_2^{(m)})}{2\sigma_m^2}. \tag{29}$$

Aggregation across features. The Ψ-statistics, $\boldsymbol{K}_{uu}^{(m)}$, and $\boldsymbol{B}_m$ do not depend on the feature index d. Summing equation (29) over $d = 1, \ldots, D_m$ replaces $\sum_d \boldsymbol{y}_{md}^\top \boldsymbol{B}_m \boldsymbol{y}_{md}$ with $\operatorname{tr}(\boldsymbol{Y}_m^\top \boldsymbol{B}_m \boldsymbol{Y}_m)$, while the remaining terms are multiplied by D_m. By $|\sigma_m^{-2}\boldsymbol{\Psi}_2 + \boldsymbol{K}_{uu}| = \sigma_m^{-2T}|\boldsymbol{A}_m|$, the modality-specific data-fit term in equation (11) becomes

$$\mathcal{H}_m = -\frac{ND_m}{2}\log(2\pi\sigma_m^2) + \frac{D_m}{2}\log\frac{\sigma_m^{2T}|\boldsymbol{K}_{uu}^{(m)}|}{|\boldsymbol{A}_m|} - \frac{1}{2}\operatorname{tr}(\boldsymbol{Y}_m^\top \boldsymbol{B}_m \boldsymbol{Y}_m) - \frac{D_m}{2\sigma_m^2}\Big(\psi_0^{(m)} - \operatorname{tr}((\boldsymbol{K}_{uu}^{(m)})^{-1}\boldsymbol{\Psi}_2^{(m)})\Big). \tag{30}$$

The last term is the trace penalty that measures how much of $\boldsymbol{K}_{ff}^{(m)}$ is not explained by the inducing variables. All determinants and inverses in equation (30) act on $T \times T$ matrices.

A.4 Analytic Ψ-Statistics

For the ARD-RBF kernel of equation (16) and the Gaussian posterior of equation (5), the kernel expectations admit closed forms. Because $q(\boldsymbol{X})$ factorizes over latent dimensions and the kernel factorizes over dimensions as well, each expectation reduces to a product of one-dimensional Gaussian integrals of the form $\int \mathcal{N}(x \mid r, v)\exp\{-a(x-c)^2\}dx = (1+2av)^{-1/2}\exp\{-a(r-c)^2/(1+2av)\}$.

The term $\psi_0^{(m)}$. Since $k_f^{(m)}(\boldsymbol{x}, \boldsymbol{x}) = \sigma_{f,m}^2$ for any $\boldsymbol{x}$, the diagonal of $\boldsymbol{K}_{ff}^{(m)}$ is deterministic and

$$\psi_0^{(m)} = \operatorname{tr}\big(\mathbb{E}_{q(\boldsymbol{X})}[\boldsymbol{K}_{ff}^{(m)}]\big) = \sum_{n=1}^{N}\mathbb{E}_{q(\boldsymbol{x}_n)}\big[k_f^{(m)}(\boldsymbol{x}_n, \boldsymbol{x}_n)\big] = N\sigma_{f,m}^2. \tag{31}$$

In particular $\psi_0^{(m)}$ does not depend on the variational parameters.

The term $\boldsymbol{\Psi}_1^{(m)}$. Applying the one-dimensional identity with $a = b_\ell^{(m)}/2$ and $c = z_{t\ell}$,

$$(\boldsymbol{\Psi}_1^{(m)})_{nt} = \mathbb{E}_{q(\boldsymbol{x}_n)}\big[k_f^{(m)}(\boldsymbol{x}_n, \boldsymbol{z}_t)\big] = \sigma_{f,m}^2\prod_{\ell=1}^{L}\frac{1}{\sqrt{b_\ell^{(m)}v_{\ell n} + 1}}\exp\left(-\frac{b_\ell^{(m)}(r_{\ell n} - z_{t\ell})^2}{2\,(b_\ell^{(m)}v_{\ell n} + 1)}\right). \tag{32}$$

The term $\boldsymbol{\Psi}_2^{(m)}$. The product of two ARD-RBF kernels evaluated at the same input satisfies

$$k_f^{(m)}(\boldsymbol{x}, \boldsymbol{z}_t)\,k_f^{(m)}(\boldsymbol{x}, \boldsymbol{z}_{t'}) = \sigma_{f,m}^4\exp\left(-\sum_{\ell=1}^{L} b_\ell^{(m)}\Big[(x_\ell - \bar{z}_{tt'\ell})^2 + \tfrac{1}{4}(z_{t\ell} - z_{t'\ell})^2\Big]\right), \qquad \bar{z}_{tt'\ell} = \frac{z_{t\ell} + z_{t'\ell}}{2}, \tag{33}$$

so that, applying the same identity with $a = b_\ell^{(m)}$ and $c = \bar{z}_{tt'\ell}$,

$$(\boldsymbol{\Psi}_2^{(m)})_{tt'} = \sigma_{f,m}^4\sum_{n=1}^{N}\prod_{\ell=1}^{L}\frac{1}{\sqrt{2b_\ell^{(m)}v_{\ell n} + 1}}\exp\left(-\frac{b_\ell^{(m)}(z_{t\ell} - z_{t'\ell})^2}{4} - \frac{b_\ell^{(m)}(r_{\ell n} - \bar{z}_{tt'\ell})^2}{2b_\ell^{(m)}v_{\ell n} + 1}\right). \tag{34}$$

Equations (31–34) show that the Ψ-statistics depend on $q(\boldsymbol{X})$ only through the posterior means $r_{\ell n}$ and the marginal variances $v_{\ell n} = (\boldsymbol{V}_\ell)_{nn}$; the off-diagonal entries of $\boldsymbol{V}_\ell$ do not enter the mapping term, although they are retained by the posterior and do enter the KL term of Appendix A.5. During optimization the Ψ-statistics are recomputed from the current $q(\boldsymbol{X})$, and their analytic forms enable efficient gradient evaluation.

A.5 Spatial KL for the Fixed-Point Variational Parameterization

Both $q(\boldsymbol{X})$ in equation (5) and $p(\boldsymbol{X} \mid \boldsymbol{S})$ in equation (3) factorize over latent dimensions into Gaussians, so the KL term of equation (7) is the sum of L Gaussian KL divergences

$$\mathcal{K}_X = \frac{1}{2} \sum_{\ell=1}^{L} \left(\operatorname{tr}(\boldsymbol{\Sigma}^{-1} \boldsymbol{V}_\ell) + \boldsymbol{r}_\ell^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{r}_\ell - N + \log \frac{|\boldsymbol{\Sigma}|}{|\boldsymbol{V}_\ell|} \right). \tag{35}$$

Substituting equations (30) and (35) into equation (15) gives the closed-form objective $\mathcal{L}$ that is maximised during training.

A.6 Interpretation of the Multimodal Objective

The multimodal objective in equation (15) is obtained by summing the modality-specific marginal-evidence bounds. Its apparent repetition of the shared spatial KL term can be understood more clearly through an equivalent normalized form. Let

$$W = \sum_{m=1}^{M} w_m, \qquad \alpha_m = \frac{w_m}{W}, \tag{36}$$

where $W > 0$ and $\sum_{m=1}^{M} \alpha_m = 1$. Since W is fixed during optimization, multiplying the objective by the positive constant $1/W$ does not change its maximizers. Therefore, equation (15) is equivalent, for optimization, to

$$\bar{\mathcal{L}} = \frac{\mathcal{L}}{W} = \sum_{m=1}^{M} \alpha_m \mathcal{H}_m - \mathcal{K}_X. \tag{37}$$

Equation (37) shows that the objective balances a weighted average of the modality-specific data-fit terms against a single shared spatial KL regularizer. Equivalently, it provides a lower bound on the composite marginal-evidence criterion

$$\bar{\mathcal{J}} = \sum_{m=1}^{M} \alpha_m \log p(\boldsymbol{Y}_m \mid \boldsymbol{S}), \tag{38}$$

which can also be written as the logarithm of the weighted geometric mean of the modality-specific marginal evidences,

$$\bar{\mathcal{J}} = \log \prod_{m=1}^{M} p(\boldsymbol{Y}_m \mid \boldsymbol{S})^{\alpha_m}. \tag{39}$$

This composite objective is deliberately different from the standard ELBO of the single joint likelihood. Under conditional independence of the modalities given the shared latent representation, the corresponding joint ELBO would take the form

$$\mathcal{L}_{\text{joint}} = \sum_{m=1}^{M} \mathcal{H}_m - \mathcal{K}_X. \tag{40}$$

For equal modality weights, expressing this joint ELBO on the same average-per-modality scale gives

$$\frac{1}{M} \mathcal{L}_{\text{joint}} = \frac{1}{M} \sum_{m=1}^{M} \mathcal{H}_m - \frac{1}{M} \mathcal{K}_X. \tag{41}$$

Hence, relative to the average per-modality data-fit contribution, the effective strength of the shared KL regularization decreases as the number of modalities increases.

In contrast, with equal weights $w_m = 1$, our normalized composite objective becomes

$$\bar{\mathcal{L}} = \frac{1}{M} \sum_{m=1}^{M} \mathcal{H}_m - \mathcal{K}_X, \tag{42}$$

so the coefficient of the shared spatial KL term remains fixed at one as additional modalities are introduced. This preserves the relative strength of the spatial regularization with respect to the average modality-specific data fit, rather than allowing it to diminish automatically with the number of modalities.

The formulation also permits unequal fixed modality weights through $\alpha_m = w_m/W$. In all experiments reported in this work, however, we use equal unit weights, $w_m = 1$ for every modality, and do not tune modality weights on individual datasets.

A.7 Derivation of the Fixed-Point Variational Parameterization

The parameterization in equation (6) is motivated by the stationary form of the variational objective. Differentiating equation (15) with respect to $\boldsymbol{V}_\ell$ and $\boldsymbol{r}_\ell$, and using $\partial \mathcal{K}_X/\partial \boldsymbol{V}_\ell = \frac{1}{2}(\boldsymbol{\Sigma}^{-1} - \boldsymbol{V}_\ell^{-1})$ and $\partial \mathcal{K}_X/\partial \boldsymbol{r}_\ell = \boldsymbol{\Sigma}^{-1}\boldsymbol{r}_\ell$ from equation (35), the stationarity conditions $\partial \mathcal{L}/\partial \boldsymbol{V}_\ell = \mathbf{0}$ and $\partial \mathcal{L}/\partial \boldsymbol{r}_\ell = \mathbf{0}$ yield

$$\boldsymbol{V}_\ell^{-1} = \boldsymbol{\Sigma}^{-1} + \boldsymbol{\Lambda}_\ell, \qquad \boldsymbol{r}_\ell = \boldsymbol{\Sigma}\bar{\boldsymbol{r}}_\ell, \tag{43}$$

where $\boldsymbol{\Lambda}_\ell = -2(\sum_m w_m)^{-1} \sum_m w_m \, \partial \mathcal{H}_m/\partial \boldsymbol{V}_\ell$ and $\bar{\boldsymbol{r}}_\ell = (\sum_m w_m)^{-1} \sum_m w_m \, \partial \mathcal{H}_m/\partial \boldsymbol{r}_\ell$. Since $\mathcal{H}_m$ depends on $\boldsymbol{V}_\ell$ only through its diagonal (Appendix A.4), the precision correction $\boldsymbol{\Lambda}_\ell$ is diagonal at any interior stationary point. This stationary condition motivates the diagonal structure in equation (6); but it does not, by itself, impose a sign constraint on its diagonal entries. We therefore restrict $\boldsymbol{\lambda}_\ell \geq \mathbf{0}$ in the variational parameterization.

This restriction provides a direct positive-definiteness guarantee. Because $\boldsymbol{\Sigma} \succ \mathbf{0}$, we have $\boldsymbol{\Sigma}^{-1} \succ \mathbf{0}$, and $\boldsymbol{\lambda}_\ell \geq \mathbf{0}$ implies $\boldsymbol{\Lambda}_\ell \succeq \mathbf{0}$. Hence

$$\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Lambda}_\ell \succ \mathbf{0}, \tag{44}$$

and therefore

$$\boldsymbol{V}_\ell = (\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Lambda}_\ell)^{-1} \succ \mathbf{0}. \tag{45}$$

Moreover, $\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Lambda}_\ell \succeq \boldsymbol{\Sigma}^{-1}$, and the order-reversing property of matrix inversion for positive-definite matrices gives

$$\mathbf{0} \prec \boldsymbol{V}_\ell \preceq \boldsymbol{\Sigma}. \tag{46}$$

Thus the non-negative diagonal correction defines a valid structured Gaussian variational covariance while reducing the N^2 entries of $\boldsymbol{V}_\ell$ to the $2N$ parameters $(\bar{\boldsymbol{r}}_\ell, \boldsymbol{\lambda}_\ell)$ per latent dimension.

A.8 Derivation of the Sparse Inducing-Variable Variational Parameterization

The fixed-point variational parameterization in Appendix A.7 provides a structured representation of the Gaussian posterior in equation (5), but still requires operations involving $N \times N$ spatial covariance matrices. We now derive the sparse spatial variational parameterization introduced in Section 2.2.2. This construction uses a different variational parameterization: rather than treating $\boldsymbol{r}_\ell$ and $\boldsymbol{V}_\ell$ as free variational quantities, they are induced by a lower-dimensional variational distribution over spatial inducing variables.

Spatial inducing variables and the GP conditional. Let $\overline{\boldsymbol{S}} \in \mathbb{R}^{J \times P}$ denote the spatial inducing locations, with $J \ll N$, and define

$$\boldsymbol{h}_\ell = x_\ell(\overline{\boldsymbol{S}}) \in \mathbb{R}^J. \tag{47}$$

Because $\boldsymbol{x}_{:\ell}$ and $\boldsymbol{h}_\ell$ are evaluations of the same spatial Gaussian process, their joint prior is

$$\begin{bmatrix} \boldsymbol{x}_{:\ell} \\ \boldsymbol{h}_\ell \end{bmatrix} \,\Big|\, \boldsymbol{S}, \overline{\boldsymbol{S}} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{K}^x_{s\bar{s}} \\ \boldsymbol{K}^x_{\bar{s}s} & \boldsymbol{K}^x_{\bar{s}\bar{s}} \end{bmatrix} \right), \tag{48}$$

where

$$\boldsymbol{K}^x_{\bar{s}\bar{s}} = k_x(\overline{\boldsymbol{S}}, \overline{\boldsymbol{S}}), \qquad \boldsymbol{K}^x_{s\bar{s}} = k_x(\boldsymbol{S}, \overline{\boldsymbol{S}}), \qquad \boldsymbol{K}^x_{\bar{s}s} = (\boldsymbol{K}^x_{s\bar{s}})^\top. \tag{49}$$

It follows immediately that

$$p(\boldsymbol{h}_\ell) = \mathcal{N}\left(\mathbf{0}, \boldsymbol{K}^x_{\bar{s}\bar{s}}\right). \tag{50}$$

Applying the standard conditional Gaussian identity to equation (48) gives

$$\begin{aligned} p(\boldsymbol{x}_{:\ell} \mid \boldsymbol{h}_\ell, \boldsymbol{S}) = \mathcal{N}\Big(& \boldsymbol{K}^x_{s\bar{s}}(\boldsymbol{K}^x_{\bar{s}\bar{s}})^{-1}\boldsymbol{h}_\ell, \\ & \boldsymbol{\Sigma} - \boldsymbol{K}^x_{s\bar{s}}(\boldsymbol{K}^x_{\bar{s}\bar{s}})^{-1}\boldsymbol{K}^x_{\bar{s}s} \Big). \end{aligned} \tag{51}$$

This is the exact conditional distribution implied by the original spatial GP prior; no additional approximation is introduced in this step.

Whitening the spatial inducing variables. Let

$$\boldsymbol{K}^x_{\bar{s}\bar{s}} = \boldsymbol{L}^x(\boldsymbol{L}^x)^\top \tag{52}$$

be the Cholesky decomposition, and introduce the whitened inducing variables through

$$\boldsymbol{h}_\ell = \boldsymbol{L}^x \boldsymbol{v}_\ell. \tag{53}$$

Since $p(\boldsymbol{h}_\ell) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{L}^x(\boldsymbol{L}^x)^\top)$, the corresponding prior is

$$p(\boldsymbol{v}_\ell) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_J). \tag{54}$$

Define

$$\boldsymbol{B}^x = \boldsymbol{K}^x_{s\bar{s}}(\boldsymbol{L}^x)^{-\top}. \tag{55}$$

Using

$$(\boldsymbol{K}^x_{\bar{s}\bar{s}})^{-1} = (\boldsymbol{L}^x)^{-\top}(\boldsymbol{L}^x)^{-1},$$

the conditional mean in equation (51) becomes

$$\begin{aligned}\boldsymbol{K}^x_{s\bar{s}}(\boldsymbol{K}^x_{\bar{s}\bar{s}})^{-1}\boldsymbol{h}_\ell &= \boldsymbol{K}^x_{s\bar{s}}(\boldsymbol{L}^x)^{-\top}(\boldsymbol{L}^x)^{-1}\boldsymbol{L}^x\boldsymbol{v}_\ell \\ &= \boldsymbol{B}^x\boldsymbol{v}_\ell,\end{aligned} \tag{56}$$

and similarly

$$\boldsymbol{K}^x_{s\bar{s}}(\boldsymbol{K}^x_{\bar{s}\bar{s}})^{-1}\boldsymbol{K}^x_{\bar{s}s} = \boldsymbol{B}^x(\boldsymbol{B}^x)^\top. \tag{57}$$

Therefore,

$$p(\boldsymbol{x}_{:\ell} \mid \boldsymbol{v}_\ell, \boldsymbol{S}) = \mathcal{N}\left(\boldsymbol{B}^x\boldsymbol{v}_\ell, \boldsymbol{\Sigma} - \boldsymbol{B}^x(\boldsymbol{B}^x)^\top\right). \tag{58}$$

We choose the whitened variational distribution

$$q(\boldsymbol{v}_\ell) = \mathcal{N}\left(\boldsymbol{\eta}_\ell, \boldsymbol{\Omega}_\ell\right), \qquad \boldsymbol{\Omega}_\ell = \mathrm{diag}(\boldsymbol{\omega}_\ell), \qquad \boldsymbol{\omega}_\ell > \boldsymbol{0}. \tag{59}$$

The structured variational family is then

$$q(\boldsymbol{x}_{:\ell}, \boldsymbol{v}_\ell) = p(\boldsymbol{x}_{:\ell} \mid \boldsymbol{v}_\ell, \boldsymbol{S})q(\boldsymbol{v}_\ell). \tag{60}$$

Thus, the exact GP conditional is retained and the only free variational parameters at the spatial inducing layer are $\boldsymbol{\eta}_\ell$ and $\boldsymbol{\omega}_\ell$.

Marginal posterior over the latent representation. The mapping GP operates on $\boldsymbol{X}$, rather than directly on the auxiliary variables $\boldsymbol{v}_\ell$. We therefore marginalize the whitened inducing variables,

$$q(\boldsymbol{x}_{:\ell}) = \int p(\boldsymbol{x}_{:\ell} \mid \boldsymbol{v}_\ell, \boldsymbol{S})q(\boldsymbol{v}_\ell)\, d\boldsymbol{v}_\ell. \tag{61}$$

Because both factors in the integrand are Gaussian and the conditional mean is linear in $\boldsymbol{v}_\ell$, this integral is available analytically.

By the law of total expectation,

$$\begin{aligned}\mathbb{E}_{q(\boldsymbol{x}_{:\ell})}[\boldsymbol{x}_{:\ell}] &= \mathbb{E}_{q(\boldsymbol{v}_\ell)}\left[\mathbb{E}[\boldsymbol{x}_{:\ell} \mid \boldsymbol{v}_\ell]\right] \\ &= \boldsymbol{B}^x\mathbb{E}_{q(\boldsymbol{v}_\ell)}[\boldsymbol{v}_\ell] \\ &= \boldsymbol{B}^x\boldsymbol{\eta}_\ell.\end{aligned} \tag{62}$$

Hence,

$$\boldsymbol{r}_\ell = \boldsymbol{B}^x\boldsymbol{\eta}_\ell. \tag{63}$$

For the covariance, the law of total covariance gives

$$\begin{aligned}\mathrm{Cov}_{q(\boldsymbol{x}_{:\ell})}(\boldsymbol{x}_{:\ell}) = \mathbb{E}_{q(\boldsymbol{v}_\ell)}\left[\mathrm{Cov}(\boldsymbol{x}_{:\ell} \mid \boldsymbol{v}_\ell)\right] \\ + \mathrm{Cov}_{q(\boldsymbol{v}_\ell)}\left[\mathbb{E}(\boldsymbol{x}_{:\ell} \mid \boldsymbol{v}_\ell)\right].\end{aligned} \tag{64}$$

Using equation (58), the first term is

$$\boldsymbol{\Sigma} - \boldsymbol{B}^x(\boldsymbol{B}^x)^\top, \tag{65}$$

whereas the second term is

$$\boldsymbol{B}^x \boldsymbol{\Omega}_\ell (\boldsymbol{B}^x)^\top. \tag{66}$$

Therefore,

$$\begin{aligned} \boldsymbol{V}_\ell &= \boldsymbol{\Sigma} - \boldsymbol{B}^x (\boldsymbol{B}^x)^\top + \boldsymbol{B}^x_\ell (\boldsymbol{B}^x)^\top \\ &= \boldsymbol{\Sigma} + \boldsymbol{B}^x \left(\boldsymbol{\Omega}_\ell - \boldsymbol{I}_J\right) (\boldsymbol{B}^x)^\top \\ &= \boldsymbol{\Sigma} + \boldsymbol{B}^x \left[\operatorname{diag}(\boldsymbol{\omega}_\ell) - \boldsymbol{I}_J\right] (\boldsymbol{B}^x)^\top. \end{aligned} \tag{67}$$

Combining equations (63) and 67 yields

$$q(\boldsymbol{x}_{:\ell}) = \mathcal{N}\left(\boldsymbol{x}_{:\ell} \mid \boldsymbol{r}_\ell, \boldsymbol{V}_\ell\right), \tag{68}$$

which has the same Gaussian marginal form as equation (5). Under the sparse spatial parameterization, however, $\boldsymbol{r}_\ell$ and $\boldsymbol{V}_\ell$ are induced by $\boldsymbol{\eta}_\ell$ and $\boldsymbol{\omega}_\ell$ rather than optimized as independent variational parameters.

Augmented spatial KL. Introducing the spatial inducing variables augments the spatial prior from $p(\boldsymbol{X} \mid \boldsymbol{S})$ to a joint distribution over the latent variables and the whitened inducing variables. Correspondingly, applying Jensen's inequality to the augmented model yields the spatial regularization

$$D_{\mathrm{KL}}\left(q(\boldsymbol{X}, \boldsymbol{V}) \,\|\, p(\boldsymbol{X}, \boldsymbol{V} \mid \boldsymbol{S})\right), \qquad \boldsymbol{V} = \{\boldsymbol{v}_\ell\}_{\ell=1}^L. \tag{69}$$

Because both the prior and variational distribution factorize over latent dimensions, this KL decomposes as

$$D_{\mathrm{KL}}\left(q(\boldsymbol{X}, \boldsymbol{V}) \,\|\, p(\boldsymbol{X}, \boldsymbol{V} \mid \boldsymbol{S})\right) = \sum_{\ell=1}^{L} D_{\mathrm{KL}}\left(q(\boldsymbol{x}_{:\ell}, \boldsymbol{v}_\ell) \,\|\, p(\boldsymbol{x}_{:\ell}, \boldsymbol{v}_\ell \mid \boldsymbol{S})\right). \tag{70}$$

Spatial KL under the structured variational family. For a single latent dimension, the augmented spatial prior factorizes as

$$p(\boldsymbol{x}_{:\ell}, \boldsymbol{v}_\ell \mid \boldsymbol{S}) = p(\boldsymbol{x}_{:\ell} \mid \boldsymbol{v}_\ell, \boldsymbol{S}) p(\boldsymbol{v}_\ell), \tag{71}$$

whereas equation (60) gives

$$q(\boldsymbol{x}_{:\ell}, \boldsymbol{v}_\ell) = p(\boldsymbol{x}_{:\ell} \mid \boldsymbol{v}_\ell, \boldsymbol{S}) q(\boldsymbol{v}_\ell).$$

Consequently,

$$\begin{aligned} &D_{\mathrm{KL}}\left(q(\boldsymbol{x}_{:\ell}, \boldsymbol{v}_\ell) \,\|\, p(\boldsymbol{x}_{:\ell}, \boldsymbol{v}_\ell \mid \boldsymbol{S})\right) \\ &= \iint p(\boldsymbol{x}_{:\ell} \mid \boldsymbol{v}_\ell, \boldsymbol{S}) q(\boldsymbol{v}_\ell) \log \frac{p(\boldsymbol{x}_{:\ell} \mid \boldsymbol{v}_\ell, \boldsymbol{S}) q(\boldsymbol{v}_\ell)}{p(\boldsymbol{x}_{:\ell} \mid \boldsymbol{v}_\ell, \boldsymbol{S}) p(\boldsymbol{v}_\ell)} \, d\boldsymbol{x}_{:\ell} \, d\boldsymbol{v}_\ell \\ &= \int q(\boldsymbol{v}_\ell) \log \frac{q(\boldsymbol{v}_\ell)}{p(\boldsymbol{v}_\ell)} \, d\boldsymbol{v}_\ell \\ &= D_{\mathrm{KL}}\left(q(\boldsymbol{v}_\ell) \,\|\, p(\boldsymbol{v}_\ell)\right), \end{aligned} \tag{72}$$

where the exact conditional distribution cancels between the numerator and denominator. Summing over the L latent dimensions therefore gives the spatial regularization term used in equation (9),

$$\mathcal{K}_X = \sum_{\ell=1}^{L} D_{\mathrm{KL}}\left(q(\boldsymbol{v}_\ell) \,\|\, p(\boldsymbol{v}_\ell)\right). \tag{73}$$

Since $p(\boldsymbol{v}_\ell) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_J)$ and $q(\boldsymbol{v}_\ell) = \mathcal{N}(\boldsymbol{\eta}_\ell, \boldsymbol{\Omega}_\ell)$, the standard Gaussian KL identity gives

$$D_{\mathrm{KL}}\left(q(\boldsymbol{v}_\ell) \,\|\, p(\boldsymbol{v}_\ell)\right) = \frac{1}{2}\left[\operatorname{tr}(\boldsymbol{\Omega}_\ell) + \boldsymbol{\eta}_\ell^\top \boldsymbol{\eta}_\ell - J - \log|\boldsymbol{\Omega}_\ell|\right]. \tag{74}$$

Because $\boldsymbol{\Omega}_\ell = \operatorname{diag}(\boldsymbol{\omega}_\ell)$, this can equivalently be written as

$$\mathcal{K}_X = \frac{1}{2} \sum_{\ell=1}^{L} \sum_{j=1}^{J} \left(\omega_{\ell j} + \eta_{\ell j}^2 - 1 - \log \omega_{\ell j}\right). \tag{75}$$

The whitening transformation is invertible, so the same regularizer may also be expressed before whitening as

$$D_{\mathrm{KL}}\left(q(\boldsymbol{h}_\ell)\,\|\,p(\boldsymbol{h}_\ell)\right) = D_{\mathrm{KL}}\left(q(\boldsymbol{v}_\ell)\,\|\,p(\boldsymbol{v}_\ell)\right). \tag{76}$$

Whitening therefore changes only the parameterization of the inducing variables and not the corresponding KL divergence.

Relation to the fixed-point variational parameterization. The stationary representation derived in Appendix A.7,

$$\boldsymbol{V}_\ell = (\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Lambda}_\ell)^{-1}, \qquad \boldsymbol{r}_\ell = \boldsymbol{\Sigma}\bar{\boldsymbol{r}}_\ell, \tag{77}$$

and the sparse representation derived above,

$$\boldsymbol{V}_\ell = \boldsymbol{\Sigma} + \boldsymbol{B}^x\left[\mathrm{diag}(\boldsymbol{\omega}_\ell) - \boldsymbol{I}_J\right](\boldsymbol{B}^x)^\top, \qquad \boldsymbol{r}_\ell = \boldsymbol{B}^x\boldsymbol{\eta}_\ell, \tag{78}$$

are two different parameterizations of a Gaussian latent posterior. The former describes the full-covariance variational construction, whereas the latter restricts the posterior to the family induced by the J spatial inducing variables. In the sparse formulation, $\boldsymbol{\eta}_\ell$ and $\boldsymbol{\omega}_\ell$ are the free spatial variational parameters; the parameters $\bar{\boldsymbol{r}}_\ell$ and $\boldsymbol{\lambda}_\ell$ of the full formulation are not simultaneously optimized.

Finally, the mapping-GP terms in Appendix A.4 depend on the latent posterior only through $\boldsymbol{r}_\ell$ and the marginal variances $v_{\ell n} = (\boldsymbol{V}_\ell)_{nn}$. From equation (67),

$$v_{\ell n} = k_x(\boldsymbol{s}_n, \boldsymbol{s}_n) + \sum_{j=1}^{J}(B^x_{nj})^2\left(\omega_{\ell j} - 1\right). \tag{79}$$

Thus the marginal posterior moments required by the mapping-GP Ψ-statistics follow directly from the sparse spatial inducing parameterization.

A.9 Gradients for the Fixed-Point Variational Parameterization

The gradients for the fixed-point variational parameterization below follow from equations (30) and (35). The corresponding gradients for the sparse spatial inducing-variable variational parameterization are given in Appendix A.10. We first record how $\mathcal{H}_m$ responds to the quantities it depends on, and then apply the chain rule.

A common form. Every parameter that enters $\mathcal{H}_m$ does so only through $\psi_0^{(m)}$, $\boldsymbol{\Psi}_1^{(m)}$, $\boldsymbol{\Psi}_2^{(m)}$, and $\boldsymbol{K}_{uu}^{(m)}$. Differentiating equation (30) with respect to a generic scalar parameter ξ gives

$$\begin{aligned}\frac{\partial \mathcal{H}_m}{\partial \xi} = &-\frac{D_m}{2\sigma_m^2}\frac{\partial \psi_0^{(m)}}{\partial \xi} + \frac{1}{\sigma_m^2}\,\mathrm{tr}\left(\frac{\partial(\boldsymbol{\Psi}_1^{(m)})^\top}{\partial \xi}\boldsymbol{Y}_m\boldsymbol{Y}_m^\top\boldsymbol{\Psi}_1^{(m)}\boldsymbol{A}_m^{-1}\right) + \frac{1}{2\sigma_m^2}\,\mathrm{tr}\left(\frac{\partial \boldsymbol{\Psi}_2^{(m)}}{\partial \xi}\boldsymbol{W}_m\right)\\ &+\frac{1}{2}\,\mathrm{tr}\left(\frac{\partial \boldsymbol{K}_{uu}^{(m)}}{\partial \xi}\left(\boldsymbol{W}_m - \frac{D_m}{\sigma_m^2}(\boldsymbol{K}_{uu}^{(m)})^{-1}\boldsymbol{\Psi}_2^{(m)}(\boldsymbol{K}_{uu}^{(m)})^{-1}\right)\right),\end{aligned} \tag{80}$$

where

$$\boldsymbol{W}_m = D_m(\boldsymbol{K}_{uu}^{(m)})^{-1} - \sigma_m^2 D_m\boldsymbol{A}_m^{-1} - \boldsymbol{A}_m^{-1}(\boldsymbol{\Psi}_1^{(m)})^\top\boldsymbol{Y}_m\boldsymbol{Y}_m^\top\boldsymbol{\Psi}_1^{(m)}\boldsymbol{A}_m^{-1} \in \mathbb{R}^{T\times T}. \tag{81}$$

equation (80) applies verbatim to $\xi \in \{r_{\ell n}, v_{\ell n}\}$, for which $\partial \boldsymbol{K}_{uu}^{(m)}/\partial\xi = \boldsymbol{0}$ and $\partial\psi_0^{(m)}/\partial\xi = 0$ by equation (31), to the mapping-kernel parameters $\boldsymbol{\theta}_f^{(m)}$, and to the inducing inputs $\boldsymbol{Z}$; the required derivatives of the Ψ-statistics follow from equations (32) and (34).

Variational parameters. Since $\boldsymbol{r}_\ell = \boldsymbol{\Sigma}\bar{\boldsymbol{r}}_\ell$ by equation (6), we have $\partial\mathcal{K}_X/\partial\boldsymbol{r}_\ell = \boldsymbol{\Sigma}^{-1}\boldsymbol{r}_\ell = \bar{\boldsymbol{r}}_\ell$, and the chain rule gives

$$\frac{\partial \mathcal{L}}{\partial \bar{\boldsymbol{r}}_\ell} = \boldsymbol{\Sigma}\left(\sum_{m=1}^{M} w_m\frac{\partial \mathcal{H}_m}{\partial \boldsymbol{r}_\ell} - \Big(\sum_{m=1}^{M} w_m\Big)\bar{\boldsymbol{r}}_\ell\right). \tag{82}$$

For $\boldsymbol{\lambda}_\ell$, differentiating $\boldsymbol{V}_\ell = (\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Lambda}_\ell)^{-1}$ gives $\partial\boldsymbol{V}_\ell/\partial\lambda_{\ell n} = -\boldsymbol{V}_\ell\boldsymbol{e}_n\boldsymbol{e}_n^\top\boldsymbol{V}_\ell$ with $\boldsymbol{e}_n$ the n-th standard basis vector, so for any function of $\boldsymbol{V}_\ell$ whose derivative is diagonal the chain rule collapses to a Hadamard product,

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\lambda}_\ell} = -(\boldsymbol{V}_\ell \circ \boldsymbol{V}_\ell)\left(\sum_{m=1}^{M} w_m\,\mathrm{diag}\left(\frac{\partial \mathcal{H}_m}{\partial \boldsymbol{V}_\ell}\right) + \frac{1}{2}\Big(\sum_{m=1}^{M} w_m\Big)\boldsymbol{\lambda}_\ell\right), \tag{83}$$

where $\circ$ denotes the Hadamard product and the second term uses $\partial \mathcal{K}_X / \partial \boldsymbol{V}_\ell = \frac{1}{2}(\boldsymbol{\Sigma}^{-1} - \boldsymbol{V}_\ell^{-1}) = -\frac{1}{2}\boldsymbol{\Lambda}_\ell$.

Mapping-kernel parameters and noise variances. The KL term does not depend on $\boldsymbol{\theta}_f^{(m)}$ or on $\boldsymbol{Z}$, so

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}_f^{(m)}} = w_m \frac{\partial \mathcal{H}_m}{\partial \boldsymbol{\theta}_f^{(m)}}, \qquad \frac{\partial \mathcal{L}}{\partial \boldsymbol{Z}} = \sum_{m=1}^{M} w_m \frac{\partial \mathcal{H}_m}{\partial \boldsymbol{Z}}, \tag{84}$$

both obtained from equation (80). The noise variances are optimized in the natural parameterization $\beta_m = 1/\sigma_m^2$, for which

$$\frac{\partial \mathcal{L}}{\partial \beta_m} = w_m \frac{\partial \mathcal{H}_m}{\partial \beta_m} = \frac{w_m}{2}\, g_m, \tag{85}$$

where

$$\begin{aligned} g_m = D_m \big(\operatorname{tr}((\boldsymbol{K}_{uu}^{(m)})^{-1}\boldsymbol{\Psi}_2^{(m)}) + (N-T)\beta_m^{-1} - \psi_0^{(m)}\big) - \operatorname{tr}(\boldsymbol{Y}_m \boldsymbol{Y}_m^\top) \\ + \operatorname{tr}\big(\boldsymbol{A}_m^{-1}(\boldsymbol{\Psi}_1^{(m)})^\top \boldsymbol{Y}_m \boldsymbol{Y}_m^\top \boldsymbol{\Psi}_1^{(m)}\big) + \beta_m^{-2} D_m \operatorname{tr}(\boldsymbol{K}_{uu}^{(m)} \boldsymbol{A}_m^{-1}) \\ + \beta_m^{-1} \operatorname{tr}\big(\boldsymbol{K}_{uu}^{(m)} \boldsymbol{A}_m^{-1} (\boldsymbol{\Psi}_1^{(m)})^\top \boldsymbol{Y}_m \boldsymbol{Y}_m^\top \boldsymbol{\Psi}_1^{(m)} \boldsymbol{A}_m^{-1}\big). \end{aligned} \tag{86}$$

Spatial-kernel parameters. The parameters $\boldsymbol{\theta}_x$ require special attention because $\boldsymbol{r}_\ell$ and $\boldsymbol{V}_\ell$ depend on $\boldsymbol{\Sigma}$ implicitly through equation (6); both the explicit dependence in $\mathcal{K}_X$ and this implicit dependence must be accounted for. Define

$$\tilde{\boldsymbol{B}}_\ell = \boldsymbol{I}_N + \boldsymbol{\Lambda}_\ell^{1/2} \boldsymbol{\Sigma} \boldsymbol{\Lambda}_\ell^{1/2}, \qquad \hat{\boldsymbol{B}}_\ell = \boldsymbol{\Lambda}_\ell^{1/2} \tilde{\boldsymbol{B}}_\ell^{-1} \boldsymbol{\Lambda}_\ell^{1/2}, \tag{87}$$

so that $\boldsymbol{V}_\ell = \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\hat{\boldsymbol{B}}_\ell\boldsymbol{\Sigma}$. Together with $\boldsymbol{r}_\ell = \boldsymbol{\Sigma}\bar{\boldsymbol{r}}_\ell$ this gives the differentials

$$d\boldsymbol{r}_\ell = (d\boldsymbol{\Sigma})\bar{\boldsymbol{r}}_\ell, \qquad d\boldsymbol{V}_\ell = (\boldsymbol{I}_N - \boldsymbol{\Sigma}\hat{\boldsymbol{B}}_\ell)\,(d\boldsymbol{\Sigma})\,(\boldsymbol{I}_N - \hat{\boldsymbol{B}}_\ell\boldsymbol{\Sigma}), \tag{88}$$

while the explicit dependence of $\mathcal{K}_X$ on $\boldsymbol{\Sigma}$ contributes $\frac{1}{2}\sum_\ell(\hat{\boldsymbol{B}}_\ell\boldsymbol{\Sigma}\hat{\boldsymbol{B}}_\ell + \bar{\boldsymbol{r}}_\ell\bar{\boldsymbol{r}}_\ell^\top)$. Collecting the two contributions into the matrix

$$\begin{aligned} \boldsymbol{G}_\ell = (\boldsymbol{I}_N - \boldsymbol{\Sigma}\hat{\boldsymbol{B}}_\ell)\operatorname{diag}\left(\sum_{m=1}^{M} w_m \frac{\partial \mathcal{H}_m}{\partial \boldsymbol{V}_\ell}\right)(\boldsymbol{I}_N - \hat{\boldsymbol{B}}_\ell\boldsymbol{\Sigma}) \\ - \frac{1}{2}\Big(\sum_{m=1}^{M} w_m\Big)\big(\hat{\boldsymbol{B}}_\ell\boldsymbol{\Sigma}\hat{\boldsymbol{B}}_\ell + \bar{\boldsymbol{r}}_\ell\bar{\boldsymbol{r}}_\ell^\top\big) \in \mathbb{R}^{N\times N}, \end{aligned} \tag{89}$$

the gradient with respect to the spatial-kernel parameters is

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}_x} = \sum_{\ell=1}^{L}\left[\operatorname{tr}\left(\boldsymbol{G}_\ell \frac{\partial \boldsymbol{\Sigma}}{\partial \boldsymbol{\theta}_x}\right) + \left(\sum_{m=1}^{M} w_m \frac{\partial \mathcal{H}_m}{\partial \boldsymbol{r}_\ell}\right)^\top \frac{\partial \boldsymbol{\Sigma}}{\partial \boldsymbol{\theta}_x}\bar{\boldsymbol{r}}_\ell\right]. \tag{90}$$

The first term propagates the change of $\boldsymbol{\Sigma}$ through the posterior covariance and the KL term, and the second propagates it through the posterior mean. Equations (82–90) give the gradients for the full-covariance variational parameterization. The corresponding gradients for the sparse spatial inducing-variable parameterization are provided in Appendix A.10. Both parameterizations are optimized under the same weighted variational objective described in Section 2.4. All of them are available in closed form, so no Monte Carlo sampling or reparameterized gradient estimator is required.

A.10 Gradients for the Sparse Inducing-Variable Variational Parameterization

The sparse spatial inducing-variable variational parameterization uses the same mapping-GP lower bounds $\mathcal{H}_m$ as the fixed-point variational parameterization. Hence, the derivatives of $\mathcal{H}_m$ with respect to the latent posterior moments follow from equation (80). We only need to propagate these derivatives through the sparse spatial inducing-variable variational parameterization.

Whitened spatial variational parameters. Under the sparse spatial inducing-variable variational parameterization,

$$\boldsymbol{r}_\ell = \boldsymbol{B}^x \boldsymbol{\eta}_\ell, \qquad \boldsymbol{B}^x = \boldsymbol{K}^x_{s\bar{s}} (\boldsymbol{L}^x)^{-\top}, \tag{91}$$

and

$$\operatorname{diag}(\boldsymbol{V}_\ell) = \operatorname{diag}(\boldsymbol{\Sigma}) + (\boldsymbol{B}^x \circ \boldsymbol{B}^x)(\boldsymbol{\omega}_\ell - \mathbf{1}_J). \tag{92}$$

Using the whitened spatial KL in equation (75), the chain rule gives

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\eta}_\ell} = (\boldsymbol{B}^x)^\top \left(\sum_{m=1}^M w_m \frac{\partial \mathcal{H}_m}{\partial \boldsymbol{r}_\ell} \right) - \Big(\sum_{m=1}^M w_m \Big) \boldsymbol{\eta}_\ell. \tag{93}$$

Similarly,

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\omega}_\ell} = (\boldsymbol{B}^x \circ \boldsymbol{B}^x)^\top \left(\sum_{m=1}^M w_m \operatorname{diag}\Big(\frac{\partial \mathcal{H}_m}{\partial \boldsymbol{V}_\ell} \Big) \right) - \frac{1}{2} \Big(\sum_{m=1}^M w_m \Big) \left(\mathbf{1}_J - \boldsymbol{\omega}_\ell^{-1} \right), \tag{94}$$

where $\circ$ denotes the Hadamard product and the inverse of $\boldsymbol{\omega}_\ell$ is taken elementwise.

Spatial-kernel parameters. After whitening, $\mathcal{K}_X$ does not explicitly depend on $\boldsymbol{\theta}_x$. The spatial-kernel parameters affect the objective through $\boldsymbol{B}^x$ and $\operatorname{diag}(\boldsymbol{\Sigma})$. Applying the chain rule gives

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}_x} = \sum_{\ell=1}^L \Bigg[& \left(\sum_{m=1}^M w_m \frac{\partial \mathcal{H}_m}{\partial \boldsymbol{r}_\ell} \right)^\top \frac{\partial \boldsymbol{B}^x}{\partial \boldsymbol{\theta}_x} \boldsymbol{\eta}_\ell \\ & + \left(\sum_{m=1}^M w_m \operatorname{diag}\Big(\frac{\partial \mathcal{H}_m}{\partial \boldsymbol{V}_\ell} \Big) \right)^\top \left\{ \frac{\partial \operatorname{diag}(\boldsymbol{\Sigma})}{\partial \boldsymbol{\theta}_x} + 2 \left(\boldsymbol{B}^x \circ \frac{\partial \boldsymbol{B}^x}{\partial \boldsymbol{\theta}_x} \right) (\boldsymbol{\omega}_\ell - \mathbf{1}_J) \right\} \Bigg]. \end{aligned} \tag{95}$$

Thus, although whitening removes the explicit dependence of the spatial KL on $\boldsymbol{\theta}_x$, the spatial kernel remains coupled to the mapping lower bounds through the latent posterior moments.

Spatial inducing locations. The spatial inducing locations $\overline{\boldsymbol{S}} \in \mathbb{R}^{J \times P}$ enter the sparse posterior through $\boldsymbol{B}^x$, since

$$\boldsymbol{B}^x = \boldsymbol{K}^x_{s\bar{s}} (\boldsymbol{L}^x)^{-\top}, \qquad \boldsymbol{K}^x_{\bar{s}\bar{s}} = \boldsymbol{L}^x (\boldsymbol{L}^x)^\top. \tag{96}$$

Because $\boldsymbol{\Sigma} = k_x(\boldsymbol{S}, \boldsymbol{S})$ and the whitened spatial KL are independent of $\overline{\boldsymbol{S}}$, for the p-th coordinate of the j-th spatial inducing location,

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \bar{s}_{jp}} = \sum_{\ell=1}^L \Bigg[& \left(\sum_{m=1}^M w_m \frac{\partial \mathcal{H}_m}{\partial \boldsymbol{r}_\ell} \right)^\top \frac{\partial \boldsymbol{B}^x}{\partial \bar{s}_{jp}} \boldsymbol{\eta}_\ell \\ & + 2 \left(\sum_{m=1}^M w_m \operatorname{diag}\Big(\frac{\partial \mathcal{H}_m}{\partial \boldsymbol{V}_\ell} \Big) \right)^\top \left(\boldsymbol{B}^x \circ \frac{\partial \boldsymbol{B}^x}{\partial \bar{s}_{jp}} \right) (\boldsymbol{\omega}_\ell - \mathbf{1}_J) \Bigg]. \end{aligned} \tag{97}$$

The required derivatives of $\boldsymbol{B}^x$ follow from the derivatives of $\boldsymbol{K}^x_{s\bar{s}}$ and $\boldsymbol{K}^x_{\bar{s}\bar{s}}$ with respect to the spatial kernel parameters or inducing locations.

The mapping-kernel parameters, shared mapping inducing inputs $\boldsymbol{Z}$, and noise parameters have the same gradients as in Appendix A.9, since the corresponding expressions depend on the spatial variational parameterization only through the Ψ-statistics. All gradients remain available in closed form, so no Monte Carlo sampling or reparameterized gradient estimator is required.

A.11 Time and Memory Complexity

Table S6 accounts for one evaluation of $\mathcal{L}$ and its gradients under the two spatial variational parameterizations.

SpaLVM with the fixed-point variational parameterization. The dominant cost is the $O(N^3)$ factorization associated with the full spatial covariance. The overall time complexity is

$$O\Big(LN^3 + MNT^2L + \textstyle\sum_m NTD_m \Big), \tag{98}$$

and the memory requirement is

$$O\Big(LN^2 + NT + \textstyle\sum_m ND_m\Big). \tag{99}$$

SpaLVM with the sparse spatial inducing-variable variational parameterization. Let $J \ll N$ denote the number of spatial inducing locations. The spatial covariance computations are replaced by operations involving the $J \times J$ inducing covariance and the $N \times J$ cross-covariance. The overall time complexity is

$$O\Big(J^3 + NJ^2 + LNJ + MNT^2L + \textstyle\sum_m NTD_m\Big), \tag{100}$$

and the memory requirement is

$$O\Big(NJ + J^2 + LN + LJ + NT + \textstyle\sum_m ND_m\Big). \tag{101}$$

The mapping-GP terms are unchanged between the two parameterizations, while the sparse spatial inducing-variable parameterization replaces the $O(LN^3)$ spatial cost by computations involving $J \ll N$.

B Related Work

Our work is closely related to algorithms developed for modeling spatial multi-omics data. Graph neural network (GNN)-based methods, including SpatialGlue (Long et al., 2024), COSMOS (Zhou et al., 2025), and MISO (Coleman et al., 2025), use graph convolution and attention architectures to capture cell neighborhood relationships. SpatialGlue (Long et al., 2024) uses graph convolutional networks (GCNs) to incorporate cell expressions and spatial information for latent embedding learning, followed by dual-attention aggregation to integrate two cellular modalities. COSMOS (Zhou et al., 2025) encodes each modality with a distinct GCN and merges their spatial representations using the Weighted Nearest Neighbor (WNN) algorithm (Hao et al., 2021), followed by spatially regularized Deep Graph Infomax (DGI) (Velickovic et al., 2018) contrastive learning scheme to refine the unified embeddings for downstream tasks such as spatial domain segmentation, low-dimensional visualization, and trajectory mapping. MISO (Coleman et al., 2025) learns modality-specific low-dimensional representations using multilayer perceptrons (MLPs) and captures cross-modal relationships by computing outer products between pairs of modality embeddings. These modality-specific features and interaction vectors are then concatenated into a unified embedding that captures information across all omics modalities. However, GNN-based methods rely on pre-built spatial graphs that are independent of model learning. Consequently, any artifactual bias introduced during graph construction inevitably propagates to downstream tasks, resulting in suboptimal performance (Tian et al., 2024).

To overcome the decoupling between spatial graph construction and model learning in GNN-based methods, SpaVAE (Tian et al., 2024) incorporates spatial dependencies directly into a probabilistic framework through Gaussian processes. Specifically, SpaVAE (Tian et al., 2024) is a Gaussian Process Prior Variational Autoencoder (Casale et al., 2018, GPPVAE) primarily designed to model single-modality spatially resolved transcriptomics (SRT) data. Similar to non-spatial VAE frameworks such as scVI (single-modality) (Lopez et al., 2018) and TotalVI (dual-modality) (Gayoso et al., 2021), SpaVAE can be extended to joint dual-modality analysis ($M = 2$) by incorporating additional network branches and modifying the objective functions.

Current deep learning-based methods are limited to at most two modalities ($M = 2$). Extending these models to multiple modalities is challenging in unsupervised settings (e.g., dimensionality reduction, visualization, and spatial domain detection), as it requires redesigning architectures, modifying loss functions, and tuning hyperparameters (Chen et al., 2025). Moreover, some methods rely heavily on preliminary dimensionality reduction rather than directly modeling expression measurements. For example, MISO and SpatialGlue use principal components (PCs) derived from traditional PCA on normalized expression values as model inputs, potentially degrading performance due to the information loss inherent in linear dimensionality reduction.

Recently proposed statistical methods, MEFISTO (Velten et al., 2022) and SMOPCA (Chen et al., 2025), can naturally integrate multiple modalities ($M \geq 1$) while modeling the original expression

matrices. These methods follow two distinct development lines: MEFISTO builds upon non-spatial single-cell multimodal methods, specifically MOFA (Argelaguet et al., 2018) and its extension MOFA+ (Argelaguet et al., 2020), whereas SMOPCA extends SpatialPCA (Shang & Zhou, 2022) to generalize single-modality spatially resolved transcriptomics (SRT) modeling to multi-omics data. Like SpaVAE, both MEFISTO and SMOPCA incorporate Gaussian process priors to capture spatial dependencies among neighboring cells, thereby avoiding the limitations associated with the explicit graph construction required by GNN-based methods. However, previous benchmarks (Long et al., 2024; Chen et al., 2025) have highlighted MEFISTO's notable limitations in accurately identifying spatial domains. Furthermore, MEFISTO incurs substantially higher computational costs (e.g., $\geq$ 24 hours) than SMOPCA and other methods (Chen et al., 2025), limiting its scalability to large datasets. The linear nature of SMOPCA and MEFISTO limits their capacity for capturing complex nonlinear patterns, potentially compromising performance in downstream analyses.

Our work builds on a line of previous studies based on Gaussian Process Latent Variable Models (GPLVMs). Lawrence (2003) introduced the GPLVM as a principled nonlinear alternative to linear models and developed a point-estimation approach for learning latent representations. Subsequent works extended the GPLVM framework to scalable and fully Bayesian inference, including the sparse GPLVM (Lawrence, 2007) and the Bayesian GPLVM (Titsias & Lawrence, 2010). GPLVMs have since been applied to the analysis of scRNA-seq data (Campbell & Yau, 2015; Ahmed et al., 2019; Verma & Engelhardt, 2020; Lalchand et al., 2022; Zhao et al., 2024). More recently, this framework has been extended to spatial transcriptomics (single modality) by augmenting the latent space with the spatial locations of cells as fixed covariates (semi-supervised Gaussian process latent variable model, ssGPLVM) (Okamoto et al.). In contrast to ssGPLVM, we employ GP priors defined over cell coordinates to model the spatial dependencies among neighborhood cells and conducted fully Bayesian inference. We further extend it to jointly modeling multi-modal readouts in a unified variational framework. We developed a sparse, structured variational inference scheme that uses inducing points for both the spatial GPs and modality-specific mapping GPs, reducing computational time by approximately 10-fold and enabling faster, more scalable inference. As a result, the proposed method is well suited to jointly capture spatial and cross-modal dependencies, which improves the efficiency of spatial domain detection in spatial multi-omics studies.

C Additional Implementation and Experimental Details

This appendix records the information needed to reproduce Section 3: the compute environment and model configuration, the baseline settings, the metric definitions, the two generative processes used in the simulation studies, and the preprocessing steps applied to the real datasets.

C.1 Implementation

SpaLVM was implemented in Python 3.10 using the following packages: PyTorch (version 2.6.0), NumPy (version 2.2.6), SciPy (version 1.15.3), scikit-learn (version 1.7.2), AnnData (version 0.11.4), and Scanpy (version 1.11.5).

SpaLVM models spatial dependence through the covariance matrix of the Gaussian process prior over the latent representation. Following Chen et al. (2025) and Tian et al. (2024), we use the well-known Matérn kernel with its smoothness parameter fixed at $3/2$ as the default spatial covariance function k_x to define $\mathbf{\Sigma}$, which is shared across all L latent dimensions. Mathematically, for $1 \leq i, j \leq N$,

$$(\mathbf{\Sigma})_{ij} = k_x(\boldsymbol{s}_i, \boldsymbol{s}_j) = \sigma_x^2 \left(1 + \frac{\sqrt{3}}{\ell_x} d(\boldsymbol{s}_i, \boldsymbol{s}_j)\right) \exp\left(-\frac{\sqrt{3}}{\ell_x} d(\boldsymbol{s}_i, \boldsymbol{s}_j)\right), \tag{102}$$

where $d(\boldsymbol{s}_i, \boldsymbol{s}_j)$ denotes the Euclidean distance between the spatial coordinates of cells i and j, and $\boldsymbol{\theta}_x = \{\sigma_x^2, \ell_x\}$ denotes the spatial-kernel parameters, with σ_x^2 the signal variance of the spatial kernel and ℓ_x its length-scale parameter. The spatial kernel parameters are jointly learned with the other model parameters by maximizing the variational objective during training.

Each modality additionally employs an independent mapping GP with an ARD-RBF kernel, as defined in equation (16). The modality-specific ARD parameters allow different latent dimensions to contribute differently to each omics modality.

Unless otherwise stated, the same SpaLVM configuration was used across all datasets. Following the conventions adopted in existing methods, such as SpaVAE (default: 20 latent factors), SpatialPCA (default: 10 PCs), and SMOPCA (default: 20 PCs), we set the dimensionality of the latent space, L, to 20 by default in SpaLVM. We used $T = 64$ inducing inputs for each mapping GP. For the sparse spatial GP approximation, we set the number of spatial inducing locations to $J = 512$ for simulated datasets and $J = 1024$ for real datasets. Kernel signal variances were initialized to 1, and length-scales to 0.5; mapping-GP inducing inputs were initialized from $\mathcal{N}(0, 0.1^2\boldsymbol{I})$, whereas spatial inducing locations $\overline{\boldsymbol{S}}$ were initialized as the K-means centroids of the spatial coordinates. All parameters were optimized jointly using Adam for 4000 iterations. The learning rate was set to 1×10^{-3}. Although the formulation in Section 2.4 supports different modality weights, following previous studies (Shang & Zhou, 2022; Tian et al., 2024; Chen et al., 2025), we use equal modality weights in all experiments (i.e., $w_m = 1$ for every modality). K-means clustering (`sklearn.cluster.KMeans`) is then applied to the learned low-dimensional representations to obtain the final clustering labels.

The optimized parameters depend on the variational parameterization: $\{\bar{\boldsymbol{r}}_\ell, \boldsymbol{\lambda}_\ell\}_{\ell=1}^L$ under the Fixed-Point Variational Parameterization, and $\{\boldsymbol{\eta}_\ell, \boldsymbol{\omega}_\ell\}_{\ell=1}^L$ under the Sparse Spatial Inducing-Variable Variational Parameterization. The mapping-GP inducing inputs $\boldsymbol{Z}$, spatial inducing locations $\overline{\boldsymbol{S}}$, kernel parameters, and modality-specific noise variances are optimized jointly in both cases.

C.2 Runtime

Table S7 reports the training time of the related methods on the real spatial datasets. All methods were evaluated on the same machine, equipped with 120 GB of memory, 10 CPU cores, and two Nvidia Tesla V100 GPUs. PCA and SpatialPCA used only CPUs, whereas SpaLVM and the deep learning-based methods (e.g., scMDC, Multigrate, COSMOS, MISO, SpaVAE, and SpatialGlue) utilized both CPUs and GPUs for model training.

Methods such as MISO and SpatialGlue generally achieved shorter training times because they operate on pre-computed principal components (PCs), rather than directly on the original high-dimensional expression matrices. In particular, they first applied classical PCA for dimensionality reduction and uses the resulting low-dimensional PCs as model inputs. Consequently, its subsequent model training involves substantially fewer input dimensions and requires considerably less computation. Thus, the shorter training time of MISO and SpatialGlue is expected and is not directly comparable to the computational cost of SpaLVM, which is trained directly on the original high-dimensional expression data.

Overall, despite being trained directly on the original high-dimensional expression data, SpaLVM remained computationally efficient, being better than SpatialPCA and approximately 10 times faster than SpaVAE across the two real datasets.

C.3 Evaluation Metrics

For Simulation Studies I and II (in Sections 3 and 3.2) and the MISAR-seq application (Section 3.4), reference labels are available, and we assess clustering performance using three widely-used metrics: Adjusted Mutual Information (AMI), Adjusted Rand Index (ARI), and Normalized Mutual Information (NMI).

Let $U = \{U_1, U_2, \cdots, U_{C_U}\}$ and $V = \{V_1, V_2, \cdots, V_{C_V}\}$ be cluster assignments for n data points with C_U and C_V clusters, respectively. Normalized Mutual Information (NMI) is defined as the mutual information between two cluster assignments, normalized by the average entropy of the two assignments:

$$NMI = \frac{I(U,V)}{\frac{1}{2}\left[H(U) + H(V)\right]},$$

where $I(U,V) = \sum_{p=1}^{C_U}\sum_{q=1}^{C_V}\frac{|U_p \cap V_q|}{n}\log\frac{n|U_p \cap V_q|}{|U_p| \times |V_q|}$ represents the mutual information between U and V; $H(U) = -\sum_{p=1}^{C_U}\frac{|U_p|}{n}\log\frac{|U_p|}{n}$ and $H(V) = -\sum_{q=1}^{C_V}\frac{|V_q|}{n}\log\frac{|V_q|}{n}$ are the entropies. Simi-

larly, Adjusted Mutual Information (AMI) is defined as follows:

$$AMI = \frac{I(U,V) - E\{I(U,V)\}}{\frac{1}{2}\left[H(U) + H(V)\right] - E\{I(U,V)\}},$$

where the extra component $E\{I(U,V)\}$ is the expected mutual information between two random clusters. Finally, Adjusted Rand Index (ARI) was calculated as follows:

$$ARI = \frac{\binom{n}{2}(a+d) - [(a+b)(a+c)+(c+d)(b+d)]}{\binom{n}{2}^2 - [(a+b)(a+c)+(c+d)(b+d)]},$$

where a is the number of pairs of objects in the same group in both U and V; b is the number of pairs in the same group in U but different groups in V; c is the number of pairs in different groups in U but the same group in V; and d is the number of pairs in different groups in both U and V.

When the predicted clusters perfectly match the ground truth, AMI, ARI and NMI reach their optimal value 1. These metrics were computed using the scikit-learn library in Python, via the functions: *normalized_mutual_info_score*, *adjusted_mutual_info_score*, and *adjusted_rand_score*.

To evaluate real datasets without ground-truth labels, we used the Moran's I score and local inverse Simpson's index (LISI) score to quantify the spatial smoothness of the identified domains. Moran's I measures global spatial autocorrelation of the learned latent representations:

$$I = \frac{N}{W} \cdot \frac{\sum_i \sum_j \left[w_{ij}(x_i - \bar{x})(x_j - \bar{x})\right]}{\sum_i (x_i - \bar{x})^2},$$

where x_i and x_j denote the values of a given latent dimension at spots i and j, respectively, $\bar{x}$ is the mean value of that latent dimension, N is the number of spots, w_{ij} is the element of the spatial weight matrix with $w_{ii} = 0$, and $W = \sum_i \sum_j w_{ij}$ is the sum of all spatial weights. Following Chen et al. (2025), we constructed the spatial weight matrix using a k-nearest-neighbor (k-NN) graph based on the spatial coordinates, with $k = 5$. Moran's I was calculated separately for each latent dimension using *squidpy.gr.spatial_autocorr* from the Squidpy package (Palla et al., 2022), and the reported score was the mean Moran's I across all latent dimensions.

The local inverse Simpson's index (LISI) score proposed by Korsunsky et al. (2019) combines the perplexity based neighborhood construction and the Inverse Simpson's Index, which reflects the effective number of clusters in each cell's neighborhood. Mathematically,

$$S = \frac{1}{\sum_{b=1}^{B} p(b)^2},$$

where $p(b)$ is the probability that a cell in the local neighborhood belongs to cluster label b, and B is the total number of clusters. The LISI score is calculated using the *compute_lisi* function from the LISI R package with default parameters (Korsunsky et al., 2019).

C.4 Data Generation Procedure for Simulation Study I

Note that the symbols introduced in this subsection describe only the data-generating process and are local to it; they are not related to the model notations used in Section 2. The code used to generate the simulated data for each scenario will be available at github after publication.

This simulation procedure was first developed by Townes & Engelhardt (2023) and later improved by Long et al. (2024) and Liu et al. (2025) to benchmark spatial multi-omics methods. Specifically, two count matrices were generated at $n = 1{,}296$ locations on a 36×36 grid, with cell i having coordinate $\boldsymbol{s}_i \in \mathbb{R}^2$ shared by both modalities. The transcriptomic-like modality $\mathbf{Y}_1 \in \mathbb{N}^{D_1 \times n}$ has $D_1 = 1{,}000$ features and the proteomic-like modality $\mathbf{Y}_2 \in \mathbb{N}^{D_2 \times n}$ has $D_2 = 100$ features. Spatial patterns, nonspatial co-activation, and a uniform background are combined to produce the count rate of each modality.

Nine canonical ggblocks patterns (Townes & Engelhardt, 2023) together with the background are tiled on the grid, and the two modalities draw on partly overlapping pattern subsets. Each modality m ($m = 1$ or 2) carries a one-hot encoded spatial indicator $\mathbf{F}_m \in \{0,1\}^{L_m \times n}$ over its L_m patterns, where an all-zero column marks a background spot. To evaluate varying degrees of shared structure,

we established three scenarios (Scenarios 1–3) by altering the allocation of private versus shared spatial patterns while keeping the total pattern pool fixed at nine. Consequently, the reference partition remains identical across scenarios, and only the overlap varies. Specifically, Scenario 1 assigns seven shared patterns and one private pattern per modality ($L_m = 8$), Scenario 2 assigns five shared and two private ($L_m = 7$), and Scenario 3 assigns three shared and three private ($L_m = 6$).

Each modality additionally carries $H = 3$ nonspatial components $\mathbf{U}_k \in \{0,1\}^{H\times n}$, with entries drawn from a $\mathrm{Bernoulli}(0.05)$ distribution. Every feature j is randomly assigned one spatial pattern and one nonspatial component. The sparse loading matrices $\mathbf{W}_m$ and $\mathbf{V}_m$ contain weights of $\rho\,\mu_{\mathrm{expr}} = 11$ and $(1-\rho)\,\mu_{\mathrm{expr}} = 9$, respectively, where $\mu_{\mathrm{expr}} = 20$ and $\rho = 0.55$ as suggested by Townes & Engelhardt (2023). Finally, a background mean of $\mu_{\mathrm{bkg}} = 0.2$ was assigned for all cells (Townes & Engelhardt, 2023; Liu et al., 2025).

Putting these elements together, the expected rate $\lambda_{m,j}(\mathbf{s}_i)$ for feature j at cell i is given by

$$\lambda_{m,j}(\mathbf{s}_i) = \mu_{\mathrm{bkg}} + \sum_{l=1}^{L_k} (\mathbf{W}_m)_{j,l}\, \mathbf{f}_{m,l}(\mathbf{s}_i) + \sum_{h=1}^{H} (\mathbf{V}_m)_{j,h}\, \mathbf{u}_{m,h}(\mathbf{s}_i),$$

where $\mathbf{f}_{m,l}(\mathbf{s}_i)$ and $\mathbf{u}_{m,h}(\mathbf{s}_i)$ are the corresponding elements of $\mathbf{F}_m$ and $\mathbf{U}_m$, respectively. The count data were simulated using a negative binomial distribution with shape parameter $r = 10$ (Townes & Engelhardt, 2023). For transcriptomic-like modality (modality 1), we sampled $\mathbf{Y}_{1,j}(\mathbf{s}_i) \sim \mathrm{NB}\big(r,\, r/(\lambda_{1,j}(\mathbf{s}_i) + r)\big)$ and introduced a zero-inflation fraction of $\pi = 0.5$ to match the dropout seen in real spatial transcriptomics data. For proteomic-like modality (modality 2), counts were simulated analogously but without zero inflation, consistent with the lower sparsity of protein assays (Townes & Engelhardt, 2023). Ten replicates per scenario were generated with distinct random seeds. We applied SpaLVM and competing methods to the data generated as described above to obtain latent factors and predict clustering labels.

C.5 Data Generation Procedure for Simulation Study II

Here, we mapped real single-cell multi-omics data onto the fixed spatial patterns of Simulation Study I, generating $n = 1,296$ spots arranged in a 36×36 grid across ten distinct spatial domains. For each real single-cell multi-omics dataset, the ten most abundant annotated cell types are selected and mapped one-to-one to the ten domains according to the relative abundance of cells and the number of spots in each domain, ensuring that each domain is assigned a sufficient number of cells. For each matched cell type–domain pair, real cells were sampled uniformly at random from the corresponding cell type population and assigned to the spots within the domain until all spots were filled. This procedure preserves the real multi-omics profiles of individual cells while imposing a predefined domain layout as the reference partition for evaluation. For each dataset, ten replicates were generated using distinct random seeds, with cells independently redrawn for every spot in each replicate.

Under this simulation setting, Scenarios 4 and 5 are designed to evaluate the method on different types of single-cell multi-omics data used in previous studies (Chen et al., 2025; Lin et al., 2022; Do et al., 2021; Gayoso et al., 2021). Specifically, Scenario 4 uses CITE-seq data, which jointly measure gene expression and surface protein profiles, across five datasets: PBMC, SLN111D1, SLN111D2, SLN208D1, and SLN208D2 (Chen et al., 2025; Lin et al., 2022; Do et al., 2021; Gayoso et al., 2021). In contrast, Scenario 5 uses SMAGE-seq data, which jointly profile gene expression and chromatin accessibility (ATAC), across two datasets, PBMC3k and PBMC10k (Chen et al., 2025; Lin et al., 2022) (See Table S1 for more information).

C.6 Preprocessing Procedure for Real Spatial Transcriptome–Proteome Data

We evaluated SpaLVM on a real human tonsil tissue (spatial-CITE-seq dataset) provided by Liu et al. (2023). The dataset was downloaded from `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE213264`. Building upon CITE-seq, the spatial CITE-seq developed by Liu et al. (2023) enables co-mapping of the genome-wide transcriptome and high-plex surface protein abundance while preserving spatial coordinates at near-cellular resolution. Following the preprocessing protocol of Chen et al. (2025), we identified spatially variable genes (SVGs) using SPARK with default parameters (Sun et al., 2020), applying a false discovery rate (FDR) threshold

of ≤ 0.05. To construct the input of the model, we retained all the available protein data and only filtered the mRNA data. Guided by Liu et al. (2023) and Tian et al. (2024), we applied each method to identify 7 spatial clusters.

C.7 Preprocessing Procedure for Real Spatial Epigenomics–Transcriptomics Data

To illustrate spatial epigenomics and transcriptomics integration, we applied SpaLVM to a benchmark MISAR-seq dataset of mouse embryonic (E15.5) brain (Jiang et al., 2023; Tian et al., 2024; Chen et al., 2025; Jiang et al., 2026). The MISAR-seq data and spot coordinates were downloaded from `https://drive.google.com/drive/folders/1xIUkhB-W2O3wHca0yfMaKAVXdVTBuf5N`. Jiang et al. (2023) developed MISAR-seq to achieve simultaneous, spatially resolved high-throughput profiling of chromatin accessibility and gene expression. Applying MISAR-seq to the mouse fetal brain, Jiang et al. (2023) characterized spatial tissue domains and elucidated the spatiotemporal regulatory logics during brain development. The dataset includes 12 expert-annotated spatial domains (Jiang et al., 2023; 2026) that can be utilized to benchmark multi-modal integration methods (Figure S26).

Following previous studies (Chen et al., 2025; Jiang et al., 2026), ATAC reads were mapped to gene regions, and the peak matrix was collapsed into a gene activity matrix, adhering to the established protocol from the Satija lab. The resulting gene activity matrix was subjected to the same preprocessing and normalization procedures as the mRNA data. Like spatial transcriptomics data, we utilized the Seurat R package (with default parameters) and selected top 2,000 highly variable genes (HVGs).

D SUPPLEMENTARY TABLES

Table S1: **Summary of the experimental datasets.** Note that the original authors did not provide ground-truth labels for the human tonsil dataset.

Dataset	Data type	Download Link	#Cell	#Groups
PBMC	CITE-seq	https://support.10xgenomics.com/single-cell-gene-expression/datasets	3,762	16
SLN111D1	CITE-seq	https://github.com/YosefLab/totalVI_reproducibility	8,853	28
SLN111D2	CITE-seq	https://github.com/YosefLab/totalVI_reproducibility	6,949	28
SLN208D1	CITE-seq	https://github.com/YosefLab/totalVI_reproducibility	8,371	28
SLN208D2	CITE-seq	https://github.com/YosefLab/totalVI_reproducibility	6,678	28
PBMC3k	SMAGE-seq	https://www.10xgenomics.com/resources/datasets	2,585	14
PBMC10k	SMAGE-seq	https://www.10xgenomics.com/resources/datasets	11,019	12
Human Tonsil Data	Spatial-CITE-seq	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE213264	2,491	–
Mouse Embryonic Brain (E15.5) Data	MISAR-seq	https://drive.google.com/drive/folders/1xlUkhB-W2O3wHca0yfMaKAVXdVTBuf5N	1,949	12

Table S2: **Summary of the competing methods.**

Method	Applicable Data Type	Download Link
PCA	Non-spatial single-modality	`https://scikit-learn.org`
SpatialPCA	Spatial single-modality	`https://github.com/shangll123/SpatialPCA`
scMDC	Non-spatial dual-modality	`https://github.com/xianglin226/scMDC`
Multigrate	Non-spatial dual-modality	`https://github.com/theislab/multigrate`
COSMOS	Spatial dual-modality	`https://cosmos-tutorials.readthedocs.io/en/latest/index.html`
MISO	Spatial dual-modality	`https://github.com/kpcoleman/miso`
SMOPCA	Spatial multi-modality	`https://github.com/cmhimself/SMOPCA`
SpaVAE	Spatial dual-modality	`https://github.com/ttgump/spaVAE`
SpatialGlue	Spatial dual-modality	`https://spatialglue-tutorials.readthedocs.io/en/latest/index.html`

Table S3: **Ablation study of SpaLVM.** ARI values are reported as mean $\pm$ standard deviation over 10 runs. SpaLVM-M1 and SpaLVM-M2 use only the first modality or the second modality, respectively, whereas SpaLVM w/o Spatial Prior replaces the spatial covariance $\boldsymbol{\Sigma}$ with the identity matrix $\boldsymbol{I}_N$.

DATASET	SpaLVM	SpaLVM-M1	SpaLVM-M2	SpaLVM w/o Spatial Prior
Scenario 1	$\mathbf{0.9994 \pm 0.0010}$	0.8094 ± 0.0056	0.8504 ± 0.0048	0.7649 ± 0.0510
Scenario 2	$\mathbf{0.9993 \pm 0.0011}$	0.5977 ± 0.0068	0.6151 ± 0.0141	0.7360 ± 0.0567
Scenario 3	$\mathbf{0.9984 \pm 0.0015}$	0.3699 ± 0.0382	0.5362 ± 0.0120	0.7208 ± 0.0490
Scenario 4 (PBMC)	$\mathbf{0.8955 \pm 0.0452}$	0.8184 ± 0.0626	0.5949 ± 0.0216	0.3570 ± 0.0292
Scenario 4 (SLN111D1)	$\mathbf{0.9492 \pm 0.0115}$	0.9101 ± 0.0277	0.4939 ± 0.0421	0.5115 ± 0.0633
Scenario 4 (SLN111D2)	$\mathbf{0.9109 \pm 0.0322}$	0.8588 ± 0.1141	0.4731 ± 0.0105	0.4852 ± 0.0418
Scenario 4 (SLN208D1)	$\mathbf{0.9411 \pm 0.0189}$	0.8992 ± 0.0215	0.4897 ± 0.0454	0.4854 ± 0.0287
Scenario 4 (SLN208D2)	$\mathbf{0.9358 \pm 0.0214}$	0.9285 ± 0.0182	0.5012 ± 0.0309	0.4848 ± 0.0493
Scenario 5 (PBMC3k)	$\mathbf{0.5711 \pm 0.0288}$	0.3845 ± 0.0443	0.5526 ± 0.0588	0.3549 ± 0.0155
Scenario 5 (PBMC10k)	$\mathbf{0.5886 \pm 0.0394}$	0.4810 ± 0.0540	0.5293 ± 0.0854	0.3350 ± 0.0199
Mouse Embryonic Brain Data	$\mathbf{0.4786 \pm 0.0109}$	0.4148 ± 0.0056	0.3976 ± 0.0049	0.4124 ± 0.0025

Table S4: **Ablation study of SpaLVM.** AMI values are reported as mean ± standard deviation over 10 runs. SpaLVM-M1 and SpaLVM-M2 use only the first modality or the second modality, respectively, whereas SpaLVM w/o Spatial Prior replaces the spatial covariance $\mathbf{\Sigma}$ with the identity matrix $\boldsymbol{I}_N$.

DATASET	SpaLVM	SpaLVM-M1	SpaLVM-M2	SpaLVM w/o Spatial Prior
Scenario 1	$\mathbf{0.9993 \pm 0.0012}$	0.9062 ± 0.0088	0.8985 ± 0.0065	0.8353 ± 0.0323
Scenario 2	$\mathbf{0.9992 \pm 0.0012}$	0.7976 ± 0.0110	0.7646 ± 0.0175	0.7815 ± 0.0346
Scenario 3	$\mathbf{0.9982 \pm 0.0017}$	0.6486 ± 0.0136	0.6979 ± 0.0067	0.7721 ± 0.0383
Scenario 4 (PBMC)	$\mathbf{0.9040 \pm 0.0258}$	0.8666 ± 0.0283	0.7999 ± 0.0183	0.6360 ± 0.0275
Scenario 4 (SLN111D1)	$\mathbf{0.9499 \pm 0.0085}$	0.9125 ± 0.0210	0.7696 ± 0.0215	0.6567 ± 0.0168
Scenario 4 (SLN111D2)	$\mathbf{0.9351 \pm 0.0173}$	0.9085 ± 0.0376	0.7648 ± 0.0107	0.6534 ± 0.0197
Scenario 4 (SLN208D1)	$\mathbf{0.9459 \pm 0.0144}$	0.9111 ± 0.0153	0.7414 ± 0.0243	0.6394 ± 0.0098
Scenario 4 (SLN208D2)	$\mathbf{0.9471 \pm 0.0120}$	0.9342 ± 0.0144	0.7909 ± 0.0141	0.6499 ± 0.0177
Scenario 5 (PBMC3k)	$\mathbf{0.7802 \pm 0.0283}$	0.6342 ± 0.0387	0.7664 ± 0.0271	0.5654 ± 0.0147
Scenario 5 (PBMC10k)	$\mathbf{0.7688 \pm 0.0271}$	0.6900 ± 0.0384	0.7355 ± 0.0593	0.5532 ± 0.0130
Mouse Embryonic Brain Data	$\mathbf{0.6225 \pm 0.0083}$	0.5651 ± 0.0035	0.5379 ± 0.0046	0.4736 ± 0.0019

Table S5: **Ablation study of SpaLVM.** NMI values are reported as mean ± standard deviation over 10 runs. SpaLVM-M1 and SpaLVM-M2 use only the first modality or the second modality, respectively, whereas SpaLVM w/o Spatial Prior replaces the spatial covariance $\mathbf{\Sigma}$ with the identity matrix $\boldsymbol{I}_N$.

DATASET	SpaLVM	SpaLVM-M1	SpaLVM-M2	SpaLVM w/o Spatial Prior
Scenario 1	$\mathbf{0.9993 \pm 0.0012}$	0.9078 ± 0.0086	0.9002 ± 0.0064	0.8380 ± 0.0318
Scenario 2	$\mathbf{0.9993 \pm 0.0012}$	0.8012 ± 0.0108	0.7687 ± 0.0172	0.7851 ± 0.0340
Scenario 3	$\mathbf{0.9983 \pm 0.0017}$	0.6544 ± 0.0135	0.7033 ± 0.0066	0.7759 ± 0.0376
Scenario 4 (PBMC)	$\mathbf{0.9056 \pm 0.0254}$	0.8688 ± 0.0279	0.8030 ± 0.0180	0.6419 ± 0.0270
Scenario 4 (SLN111D1)	$\mathbf{0.9507 \pm 0.0084}$	0.9140 ± 0.0207	0.7732 ± 0.0212	0.6626 ± 0.0167
Scenario 4 (SLN111D2)	$\mathbf{0.9362 \pm 0.0170}$	0.9100 ± 0.0370	0.7685 ± 0.0105	0.6592 ± 0.0194
Scenario 4 (SLN208D1)	$\mathbf{0.9468 \pm 0.0142}$	0.9126 ± 0.0150	0.7455 ± 0.0240	0.6455 ± 0.0097
Scenario 4 (SLN208D2)	$\mathbf{0.9480 \pm 0.0118}$	0.9352 ± 0.0141	0.7942 ± 0.0139	0.6556 ± 0.0174
Scenario 5 (PBMC3k)	$\mathbf{0.7838 \pm 0.0278}$	0.6403 ± 0.0380	0.7702 ± 0.0267	0.5728 ± 0.0143
Scenario 5 (PBMC10k)	$\mathbf{0.7725 \pm 0.0266}$	0.6951 ± 0.0378	0.7398 ± 0.0584	0.5606 ± 0.0129
Mouse Embryonic Brain Data	$\mathbf{0.6281 \pm 0.0082}$	0.5716 ± 0.0034	0.5445 ± 0.0046	0.4813 ± 0.0019

Table S6: **Leading-order cost of one objective and gradient evaluation.** N is the number of spatial observations, L the latent dimension, J the number of spatial inducing locations, T the number of mapping inducing inputs, M the number of modalities, and D_m the feature dimension of modality m. SpaLVM* denotes SpaLVM with the fixed-point variational parameterization, whereas SpaLVM denotes SpaLVM with the sparse spatial inducing-variable variational parameterization.

Quantity	SpaLVM*		SpaLVM	
	Time	Memory	Time	Memory
$\{\boldsymbol{r}_\ell, \boldsymbol{V}_\ell\}_{\ell=1}^L$	$O(LN^3)$	$O(LN^2)$	–	–
$\boldsymbol{K}_{\bar{s}\bar{s}}^x$, $\boldsymbol{L}^x$, and $\boldsymbol{B}^x$	–	–	$O(J^3 + NJ^2)$	$O(NJ + J^2)$
$\{\boldsymbol{r}_\ell, \mathrm{diag}(\boldsymbol{V}_\ell)\}_{\ell=1}^L$	–	–	$O(LNJ)$	$O(LN + LJ)$
$\{\psi_0^{(m)}, \mathbf{\Psi}_1^{(m)}, \mathbf{\Psi}_2^{(m)}\}_{m=1}^M$	$O(MNT^2L)$	$O(NT + T^2)$	$O(MNT^2L)$	$O(NT + T^2)$
$\{\mathcal{H}_m\}_{m=1}^M$	$O(\sum_m NTD_m)$	$O(\sum_m ND_m)$	$O(\sum_m NTD_m)$	$O(\sum_m ND_m)$

Table S7: **Runtime comparison of the most related algorithms on real spatial datasets, with "s" for seconds, "m" for minutes and "h" for hours.** The results were obtained on a machine with 120 GB of memory, 10 CPU cores and 2 Nvidia Tesla V100 GPUs. PCA and SpatialPCA only utilized CPUs for computation, while SpaLVM, SpaLVM*, scMDC, Multigrate, COSMOS, MISO, SpaVAE and SpatialGlue used GPUs during training. SpaLVM* denotes SpaLVM with the fixed-point variational parameterization, whereas SpaLVM denotes SpaLVM with the sparse spatial inducing-variable variational parameterization.

Method	Human tonsil data	Mouse embryonic E15.5 brain data
SpaLVM	3.26m	3.06m
SpaLVM*	63.16m	39.19m
PCA	0.1s	0.4s
SpatialPCA	5.5m	13.0m
scMDC	1.7m	1.9m
Multigrate	35.1s	40.0s
COSMOS	53.2s	1.3m
MISO	27.2s	38.6s
SMOPCA	48.7s	1.2m
SpaVAE	20.4m	47.7m
SpatialGlue	25.5s	57.1s

E SUPPLEMENTARY FIGURES

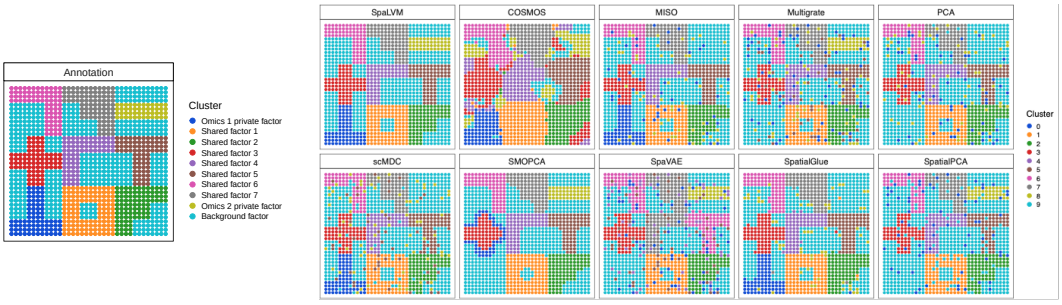

Figure S1: **Simulation Study I, spatial domain detection results for Scenario 1.** (left) Simulated cell spatial locations and matching ground-truth patterns. Each modality has one private pattern, and colors represent the ground-truth labels. (right) Example clustering results from SpaLVM and the competing methods on a simulated dataset. Colors represent the predicted labels.

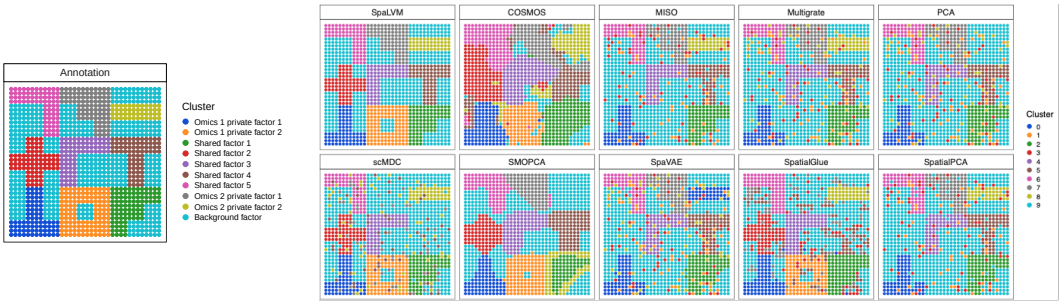

Figure S2: **Simulation Study I, spatial domain detection results for Scenario 2.** (left) Simulated cell spatial locations and matching ground-truth patterns. Each modality has two private patterns, and colors represent the ground-truth labels. (right) Example clustering results from SpaLVM and the competing methods on a simulated dataset. Colors represent the predicted labels.

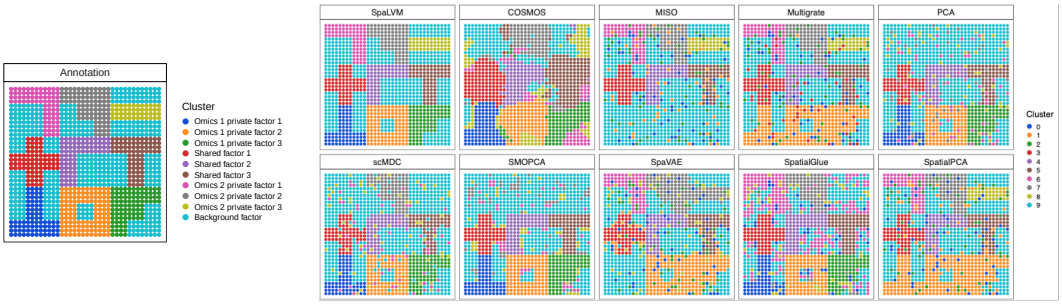

Figure S3: **Simulation Study I, spatial domain detection results for Scenario 3.** (left) Simulated cell spatial locations and matching ground-truth patterns. Each modality has three private patterns, and colors represent the ground-truth labels. (right) Example clustering results from SpaLVM and the competing methods on a simulated dataset. Colors represent the predicted labels.

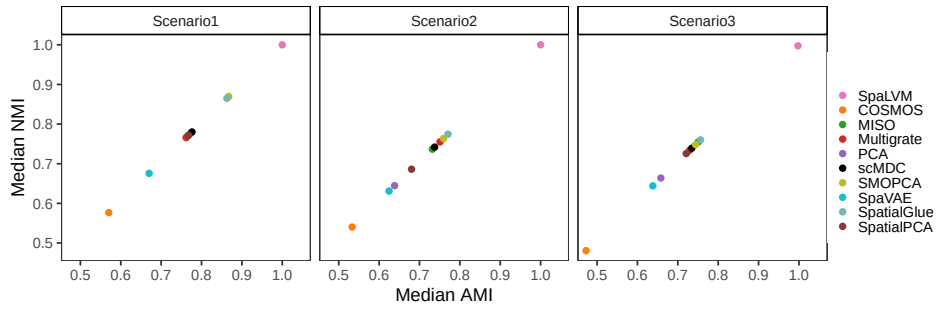

Figure S4: **Simulation Study I, clustering performance jointly quantified by NMI and AMI.** Clustering performance of SpaLVM and competing methods on simulated datasets (Scenario 1-3), jointly quantified by median NMI and AMI. Clustering performance is illustrated in a two-dimensional manner with median NMI as the Y axis and median AMI as the X axis. For each Scenario, we repeat the experiments 10 times with different random seeds for generating the data and the median metric value is calculated.

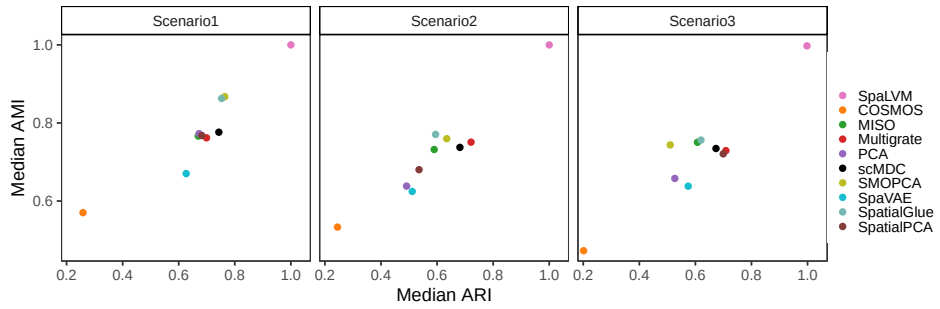

Figure S5: **Simulation Study I, clustering performance jointly quantified by AMI and ARI.** Clustering performance of SpaLVM and competing methods on simulated datasets (Scenario 1-3), jointly quantified by median AMI and ARI. Clustering performance is illustrated in a two-dimensional manner with median AMI as the Y axis and median ARI as the X axis. For each Scenario, we repeat the experiments 10 times with different random seeds for generating the data and the median metric value is calculated.

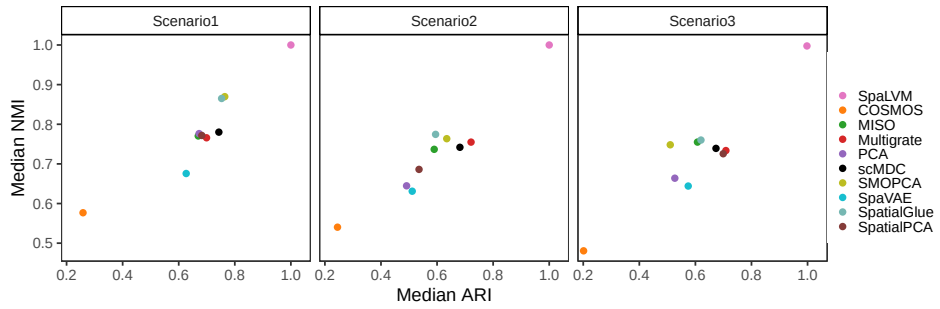

Figure S6: **Simulation Study I, clustering performance jointly quantified by NMI and ARI.** Clustering performance of SpaLVM and competing methods on simulated datasets (Scenario 1-3), jointly quantified by median NMI and ARI. Clustering performance is illustrated in a two-dimensional manner with median NMI as the Y axis and median ARI as the X axis. For each Scenario, we repeat the experiments 10 times with different random seeds for generating the data and the median metric value is calculated.

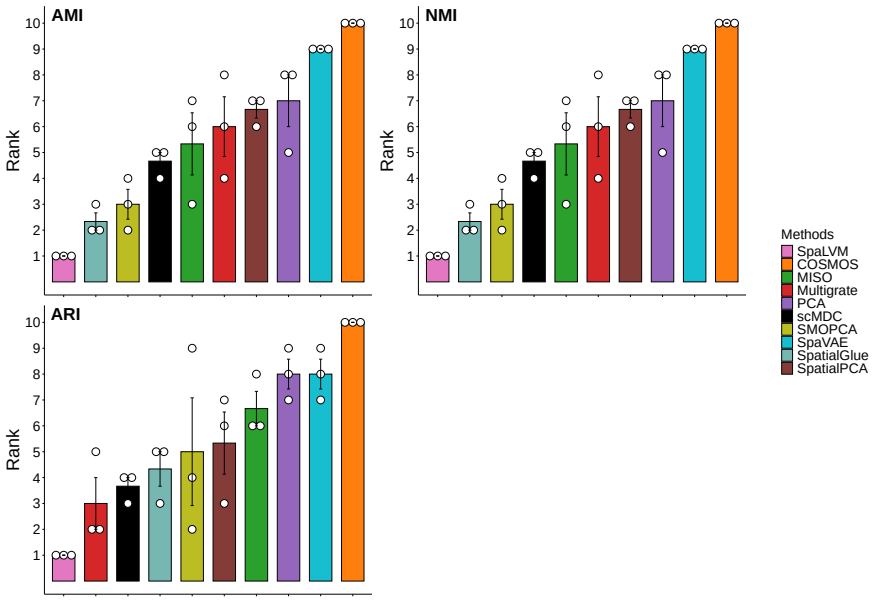

Figure S7: **Simulation Study I, method ranking.** Clustering performance of SpaLVM and competing methods on simulated datasets (Scenario 1-3), quantified by the averaged rank. We summarized the performance of each method by showing the averaged ranks based on median AMI, NMI and ARI values across the three simulation scenarios ($n = 3$). For each Scenario, we repeat the experiments 10 times with different random seeds for generating the data and the rank of each method is calculated based on the median metric values. A rank of a method in a scenario is shown by a dot and the standard errors are shown by the error bars.

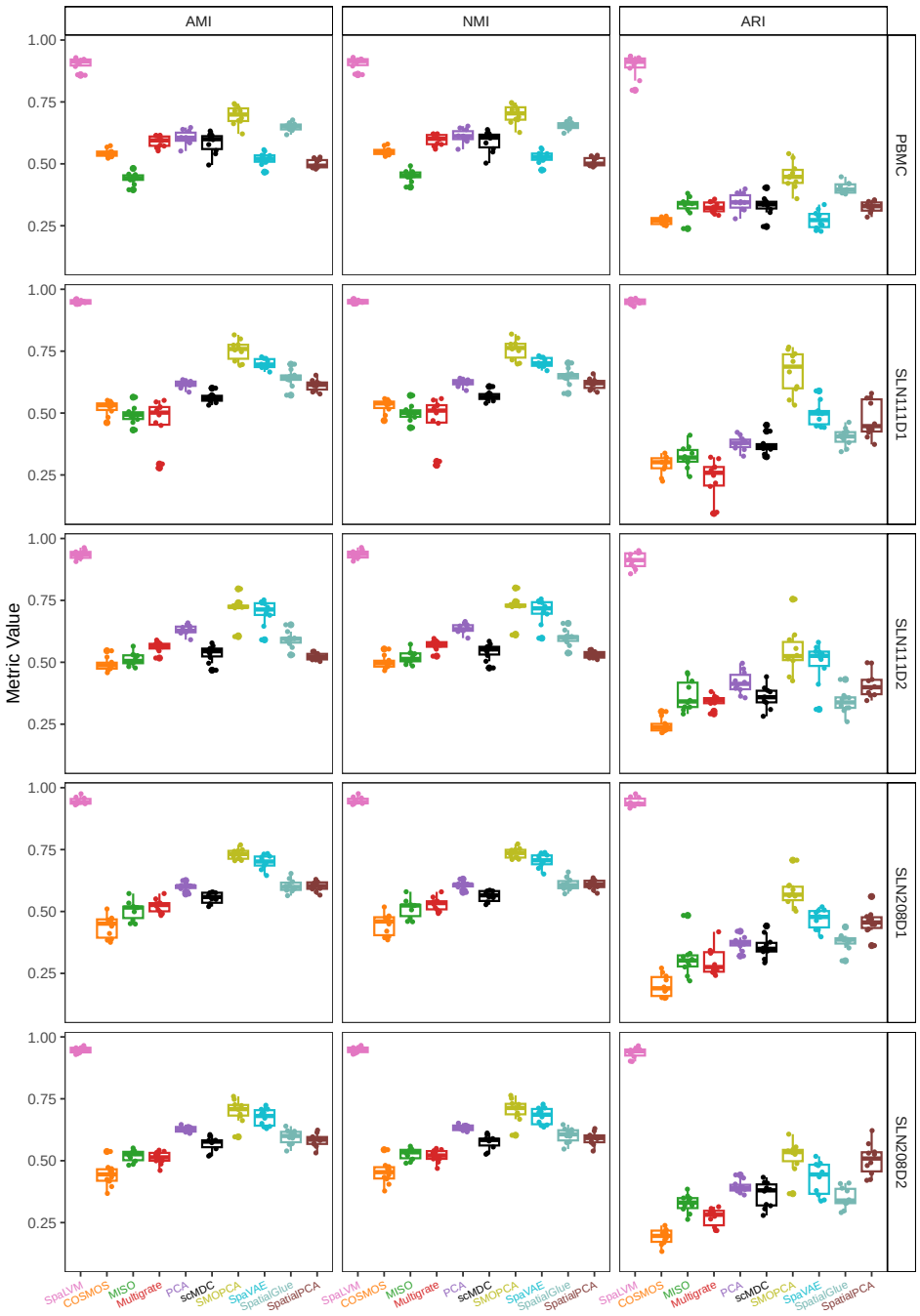

Figure S8: **Simulation Study II, CITE-seq benchmarks against multi-modality baselines.** Clustering performance on the five CITE-seq datasets for Scenario 4, measured by AMI, NMI, and ARI.

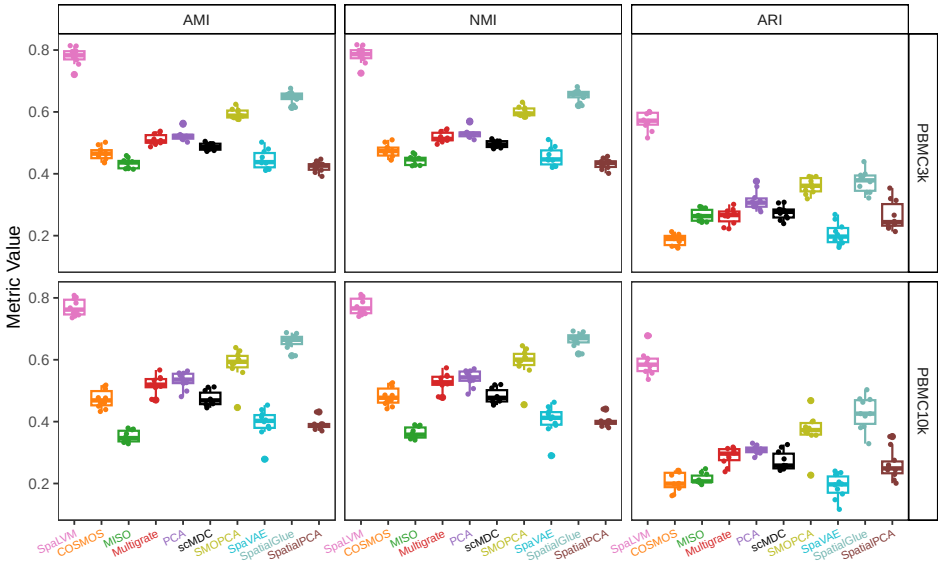

Figure S9: **Simulation Study II, SMAGE-seq benchmarks against multi-modality baselines.** Clustering performance on the two SMAGE-seq datasets for Scenario 5, measured by AMI, NMI, and ARI.

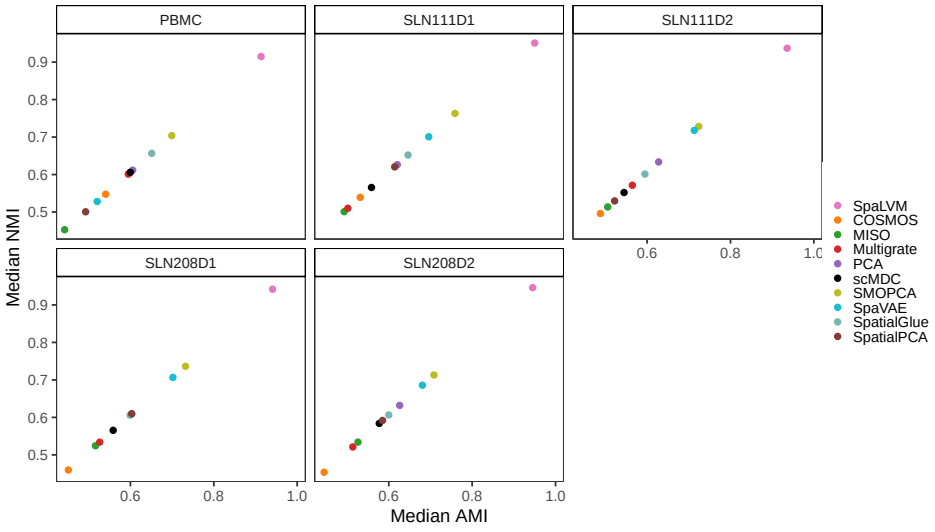

Figure S10: **Simulation Study II, CITE-seq clustering performance jointly quantified by NMI and AMI.** Clustering performance of SpaLVM and competing methods on five CITE-seq datasets with simulated spatial coordinates, jointly quantified by median NMI and AMI. Clustering performance is illustrated in a two-dimensional manner with median NMI as the Y axis and median AMI as the X axis. For each dataset, we repeat the experiments 10 times and the median metric value is used.

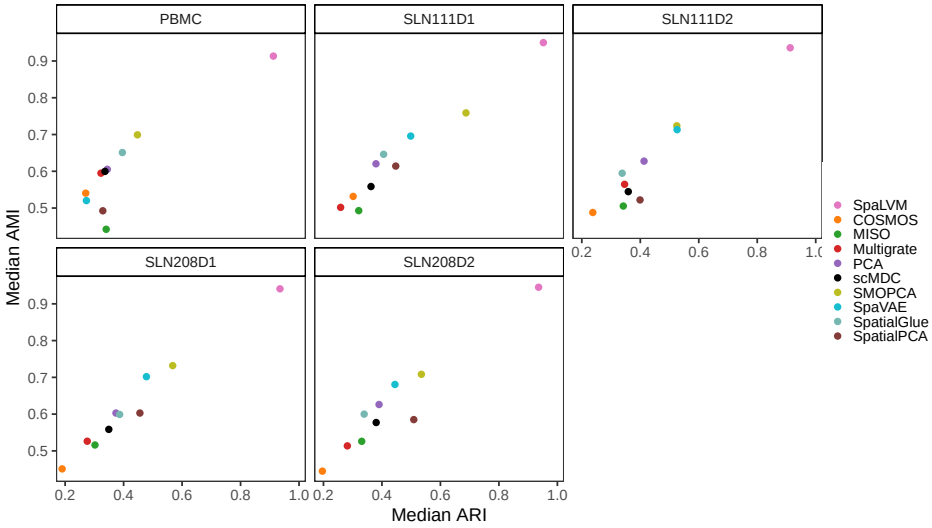

Figure S11: **Simulation Study II, CITE-seq clustering performance jointly quantified by AMI and ARI.** Clustering performance of SpaLVM and competing methods on five CITE-seq datasets with simulated spatial coordinates, jointly quantified by median AMI and ARI. Clustering performance is illustrated in a two-dimensional manner with median AMI as the Y axis and median ARI as the X axis. For each dataset, we repeat the experiments 10 times and the median metric value is used.

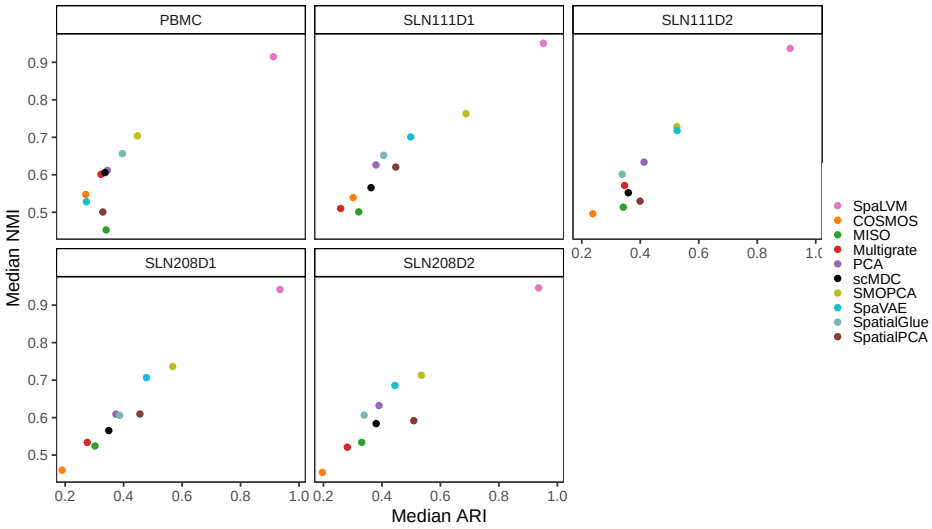

Figure S12: **Simulation Study II, CITE-seq clustering performance jointly quantified by NMI and ARI.** Clustering performance of SpaLVM and competing methods on five CITE-seq datasets with simulated spatial coordinates, jointly quantified by median NMI and ARI. Clustering performance is illustrated in a two-dimensional manner with median NMI as the Y axis and median ARI as the X axis. For each dataset, we repeat the experiments 10 times and the median metric value is used.

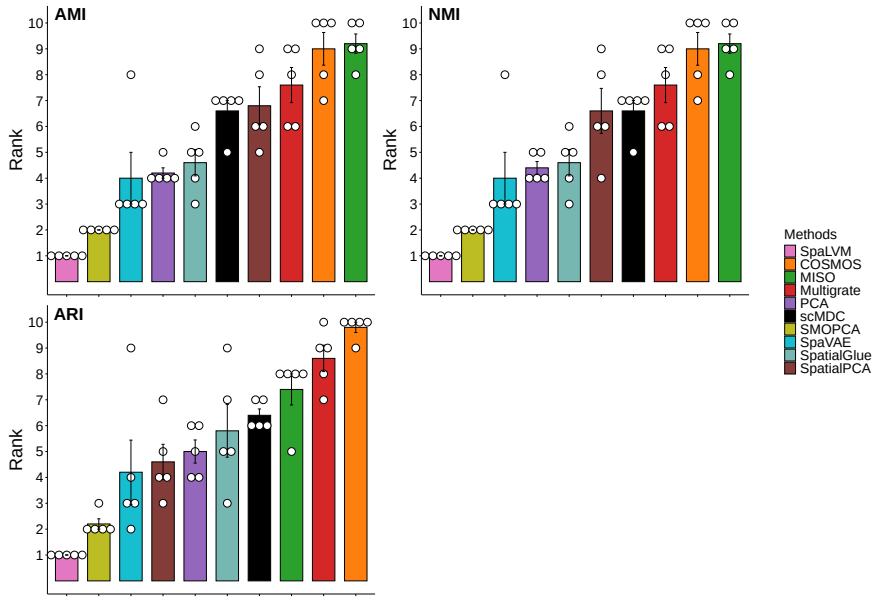

Figure S13: **Simulation Study II, CITE-seq method ranking.** Clustering performance of SpaLVM and competing methods on CITE-seq datasets with simulated spatial coordinates, quantified by the averaged rank. We summarized the performance of each method by showing the averaged ranks based on the median AMI, NMI and ARI values on the five CITE-seq datasets ($n = 5$). For each dataset, we repeat the experiments 10 times and the rank of each method is calculated based on the median metric values. A rank of a method in a dataset is shown by a dot and the standard errors are shown by the error bars.

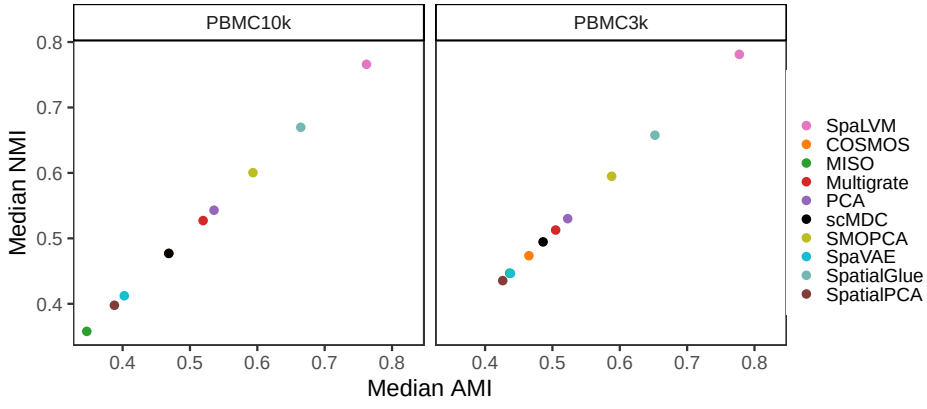

Figure S14: **Simulation Study II, SMAGE-seq clustering performance jointly quantified by NMI and AMI.** Clustering performance of SpaLVM and competing methods on two SMAGE-seq datasets with simulated spatial coordinates, jointly quantified by median NMI and AMI. Clustering performance is illustrated in a two-dimensional manner with median NMI as the Y axis and median AMI as the X axis. For each dataset, we repeat the experiments 10 times and the median metric value is used.

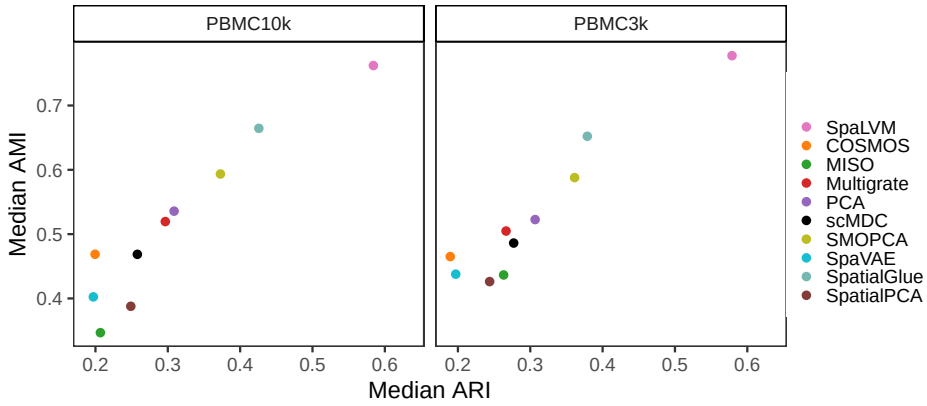

Figure S15: **Simulation Study II, SMAGE-seq clustering performance jointly quantified by AMI and ARI.** Clustering performance of SpaLVM and competing methods on two SMAGE-seq datasets with simulated spatial coordinates, jointly quantified by median AMI and ARI. Clustering performance is illustrated in a two-dimensional manner with median AMI as the Y axis and median ARI as the X axis. For each dataset, we repeat the experiments 10 times and the median metric value is used.

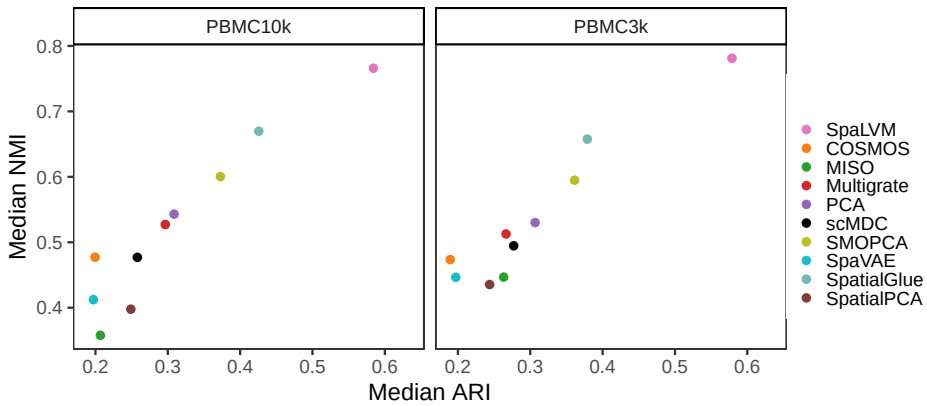

Figure S16: **Simulation Study II, SMAGE-seq clustering performance jointly quantified by NMI and ARI.** Clustering performance of SpaLVM and competing methods on two SMAGE-seq datasets with simulated spatial coordinates, jointly quantified by median NMI and ARI. Clustering performance is illustrated in a two-dimensional manner with median NMI as the Y axis and median ARI as the X axis. For each dataset, we repeat the experiments 10 times and the median metric value is used.

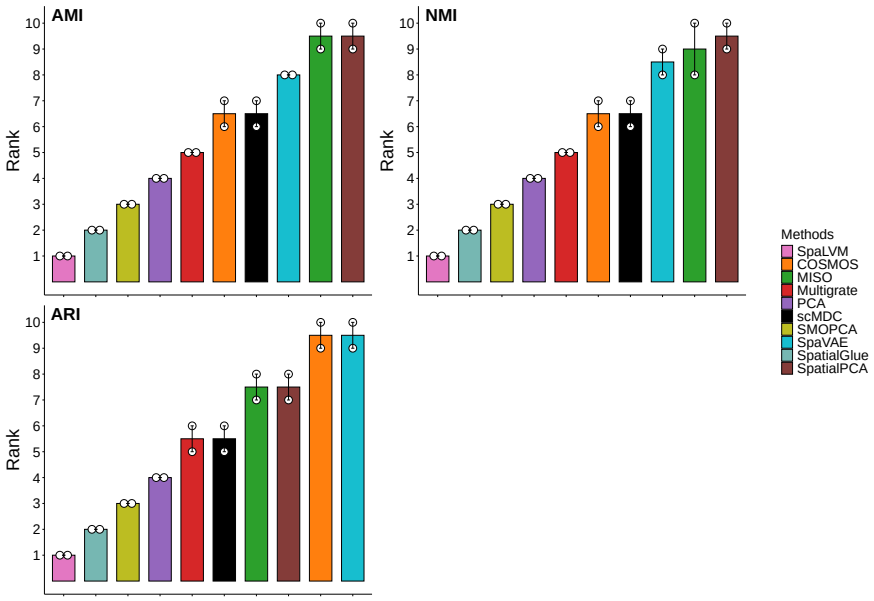

Figure S17: **Simulation Study II, SMAGE-seq method ranking.** Clustering performance of SpaLVM and competing methods on two SMAGE-seq datasets with simulated spatial coordinates, quantified by the averaged rank. We summarized the performance of each method by showing the averaged ranks based on the median AMI, NMI and ARI values on the two SMAGE-seq datasets ($n = 2$). For each dataset, we repeat the experiments 10 times and the rank of each method is calculated based on the median metric values. A rank of a method in a dataset is shown by a dot and the standard errors are shown by the error bars.

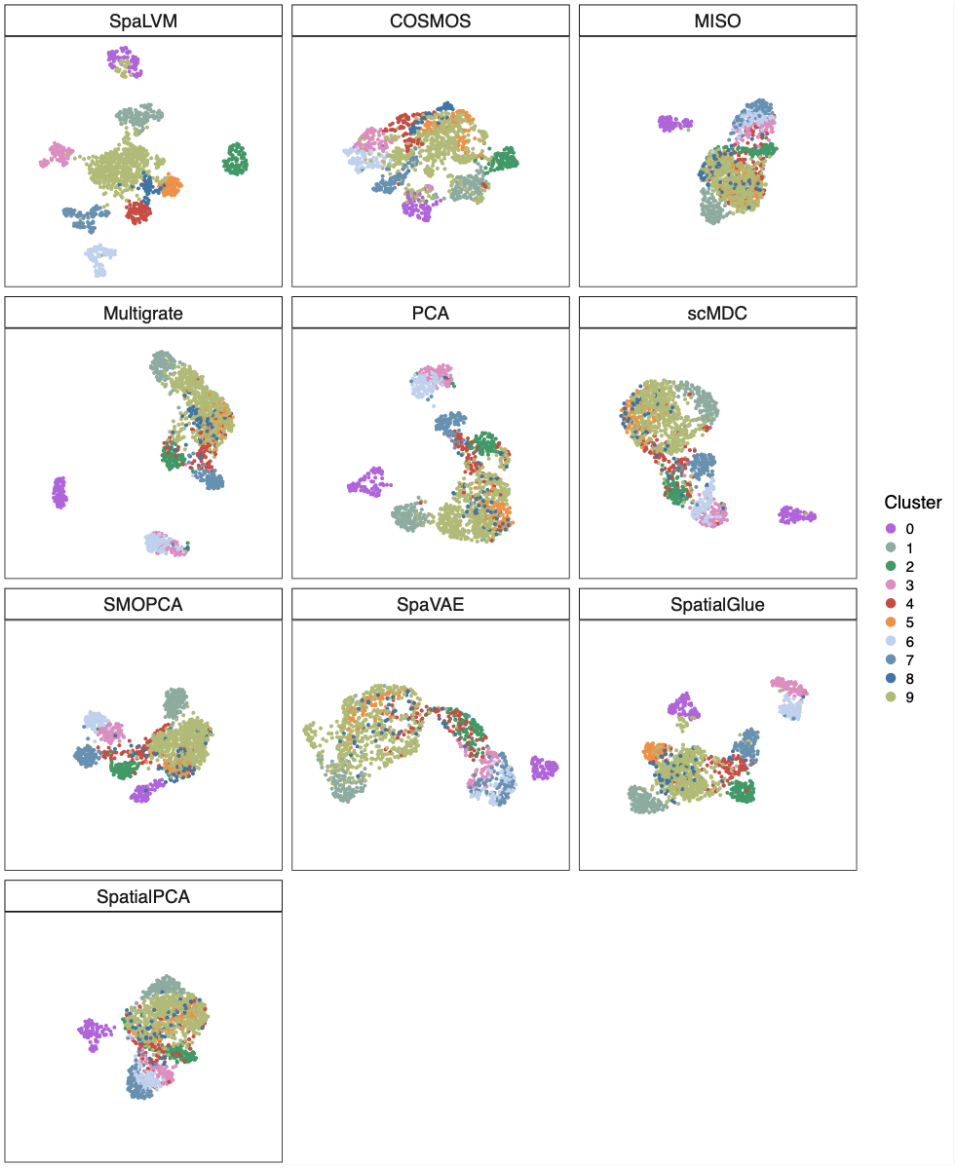

Figure S18: **Simulation Study II, PBMC CITE-seq latent representations.** Learned latent representations of SpaLVM and competing methods on the PBMC dataset with simulated spatial coordinates. The latent representations are visualized with UMAP plots. Colors represent ground-truth labels.

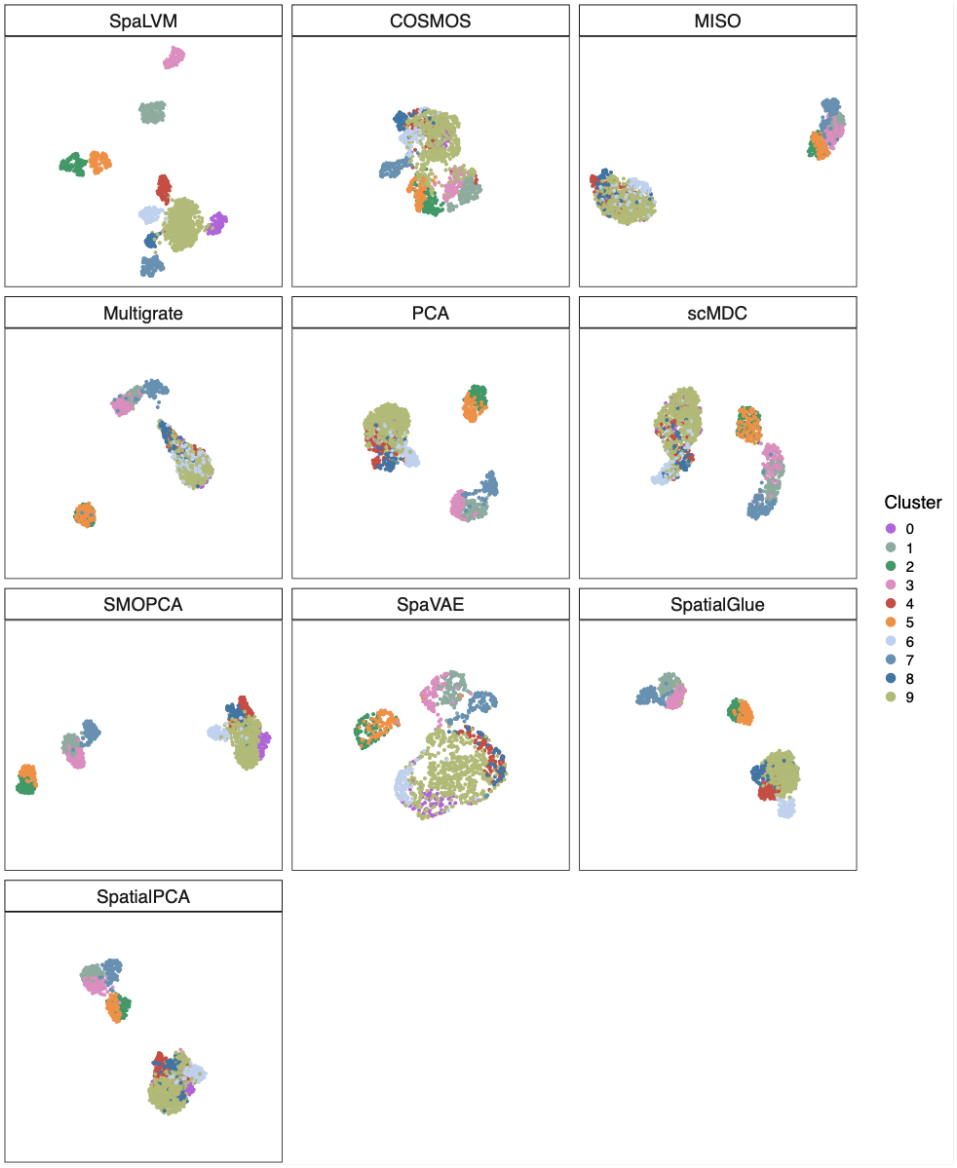

Figure S19: **Simulation Study II, SLN111D1 CITE-seq latent representations.** Learned latent representations of SpaLVM and competing methods on the SLN111D1 dataset with simulated spatial coordinates. The latent representations are visualized with UMAP plots. Colors represent ground-truth labels.

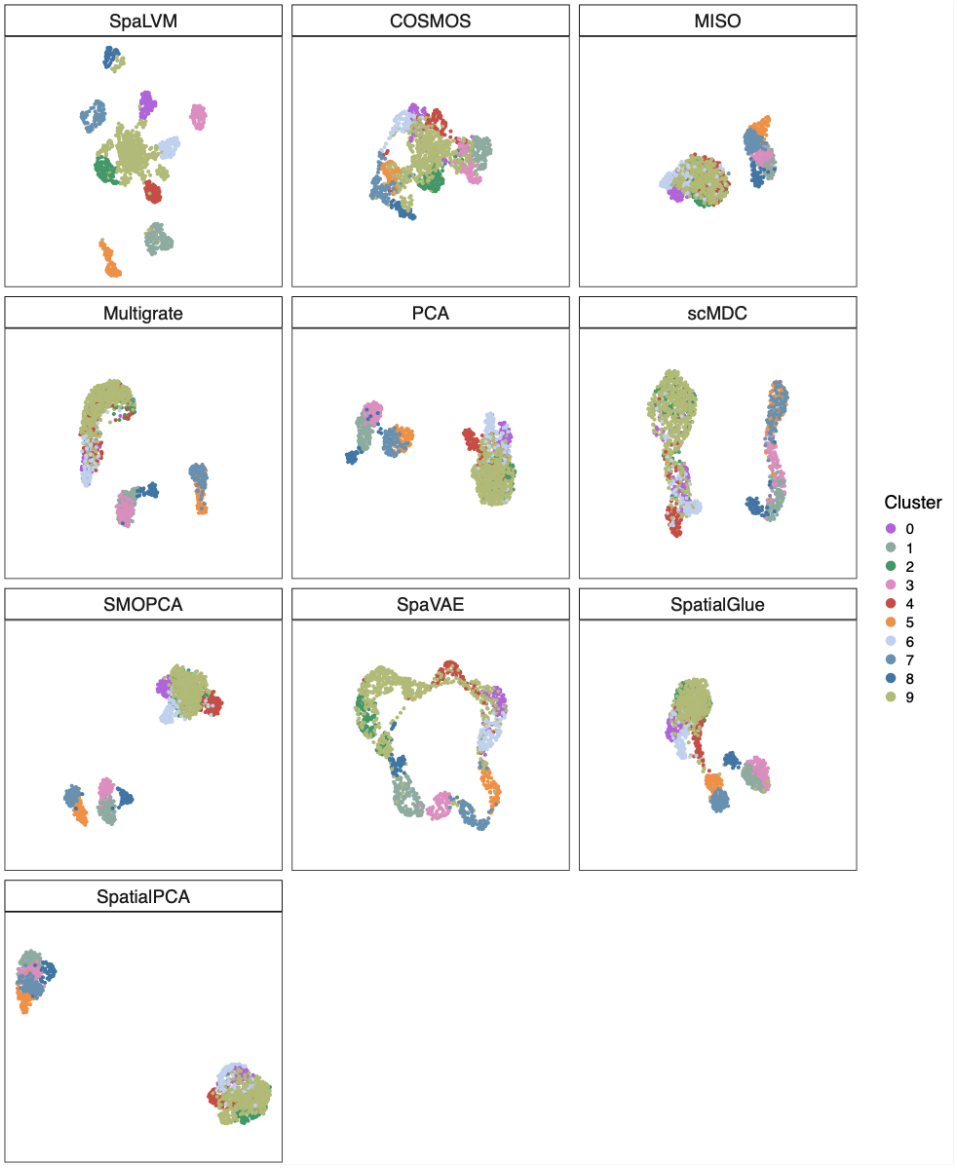

Figure S20: **Simulation Study II, SLN111D2 CITE-seq latent representations.** Learned latent representations of SpaLVM and competing methods on the SLN111D2 dataset with simulated spatial coordinates. The latent representations are visualized with UMAP plots. Colors represent ground-truth labels.

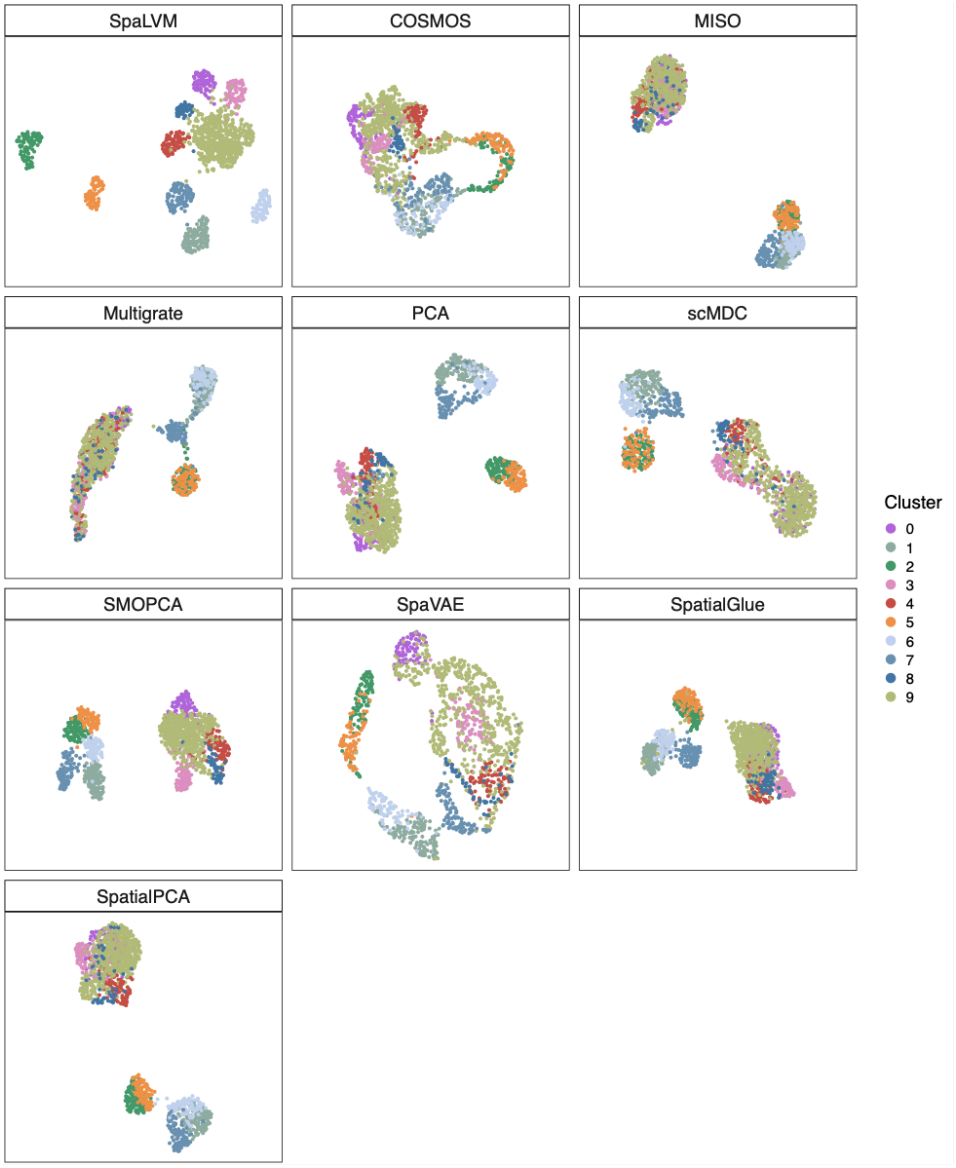

Figure S21: **Simulation Study II, SLN208D1 CITE-seq latent representations.** Learned latent representations of SpaLVM and competing methods on the SLN208D1 dataset with simulated spatial coordinates. The latent representations are visualized with UMAP plots. Colors represent ground-truth labels.

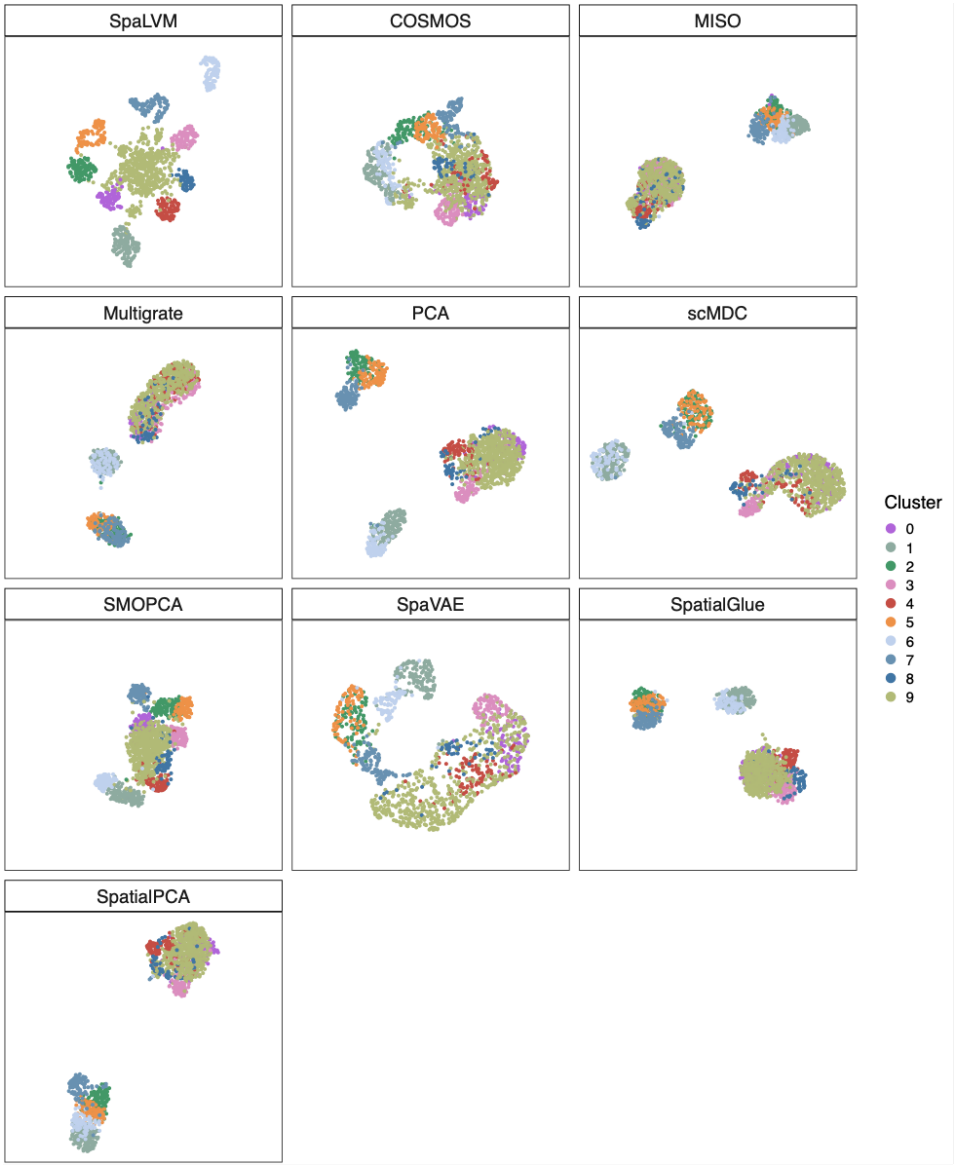

Figure S22: **Simulation Study II, SLN208D2 CITE-seq latent representations.** Learned latent representations of SpaLVM and competing methods on the SLN208D2 dataset with simulated spatial coordinates. The latent representations are visualized with UMAP plots. Colors represent ground-truth labels.



Figure S23: **Simulation Study II, PBMC3K SMAGE-seq latent representations.** Learned latent representations of SpaLVM and competing methods on the PBMC3K dataset with simulated spatial coordinates. The latent representations are visualized with UMAP plots. Colors represent ground-truth labels.

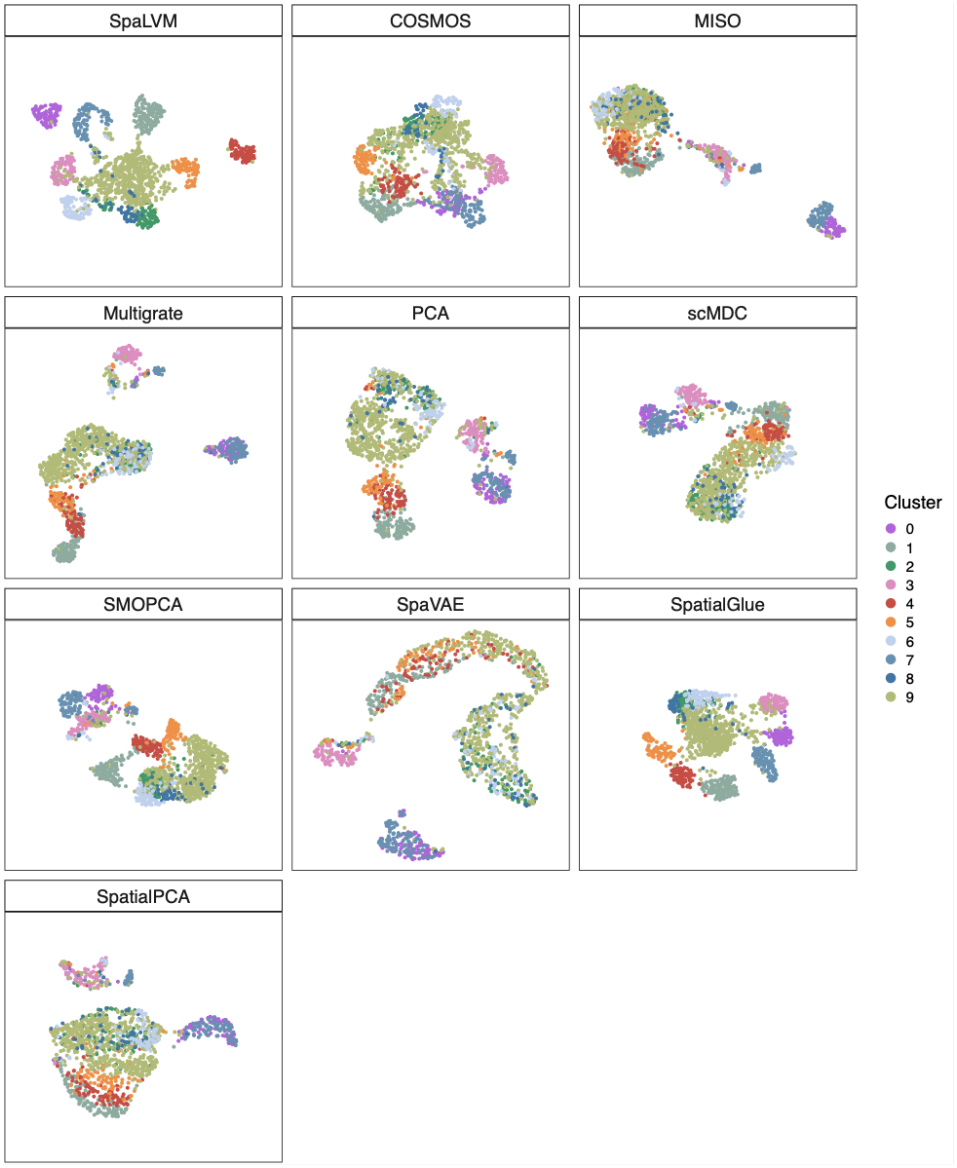

Figure S24: **Simulation Study II, PBMC10K SMAGE-seq latent representations.** Learned latent representations of SpaLVM and competing methods on the PBMC10K dataset with simulated spatial coordinates. The latent representations are visualized with UMAP plots. Colors represent ground-truth labels.

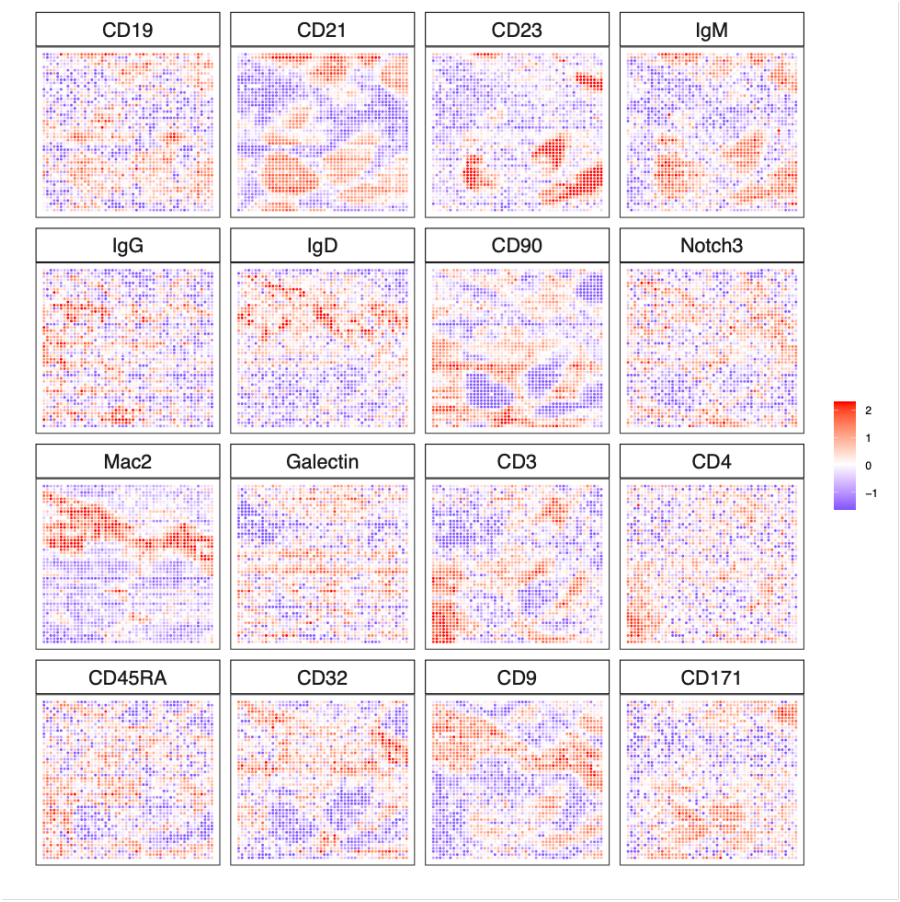

Figure S25: **Human tonsil data (spatial-CITE-seq) results, identified protein markers.** Protein expression levels of the identified differentially expressed proteins for the real spatial-CITE-seq dataset.

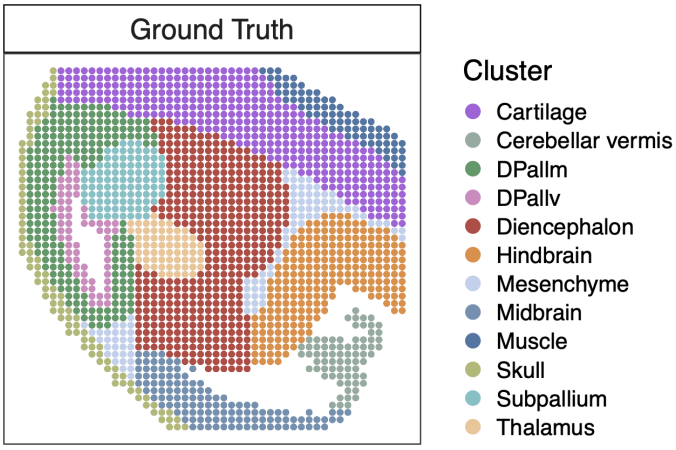

Figure S26: **Mouse embryonic brain (E15.5) data (MISAR-seq), expert annotation.** Manually annotated labels of the MISAR-seq mouse E15.5 brain dataset.

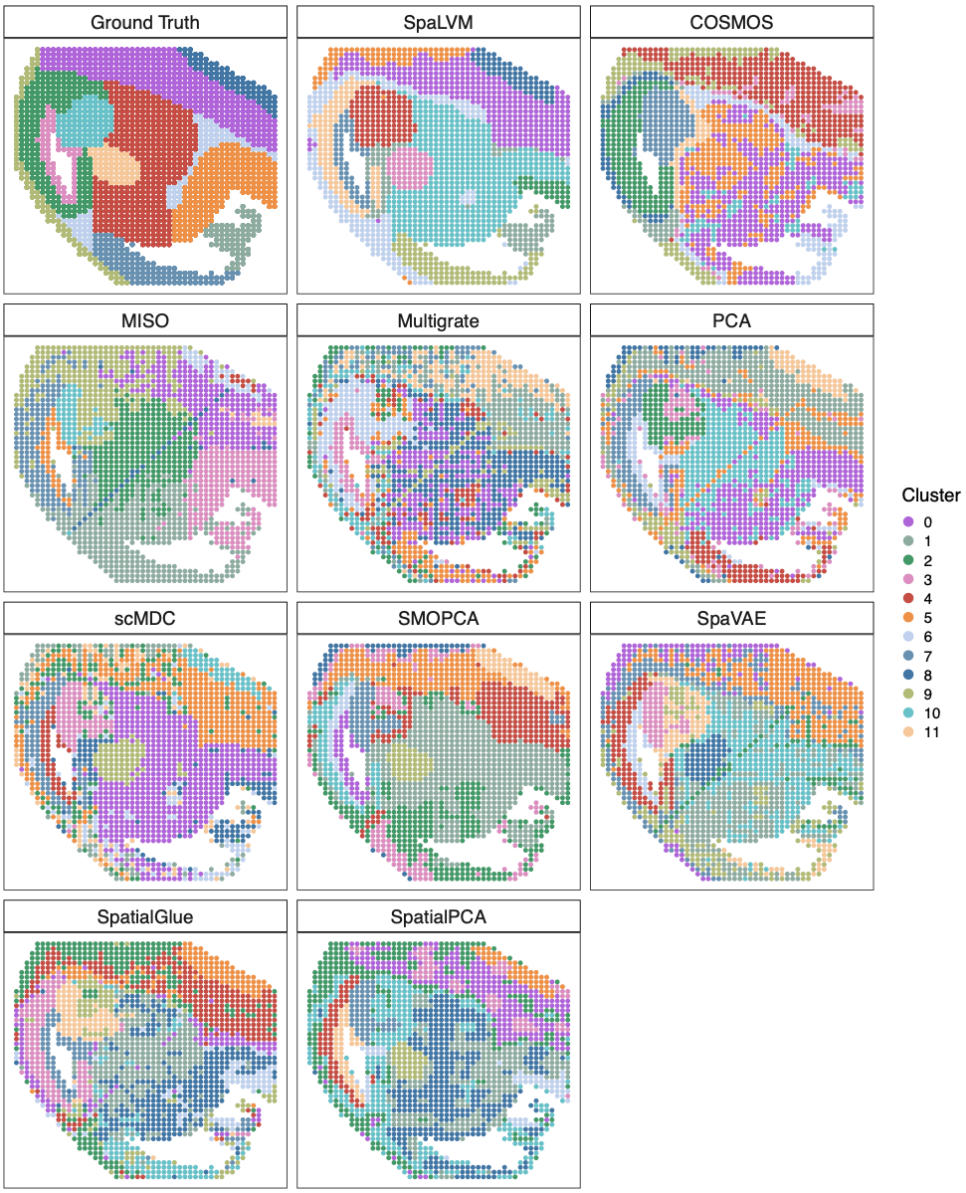

Figure S27: **Mouse embryonic brain (E15.5) data (MISAR-seq), identified spatial domains.** Spatial domains detected by SpaLVM and the competing methods on the MISAR-seq mouse E15.5 brain dataset. Colors denote the predicted cluster labels.



Figure S28: **Mouse embryonic brain (E15.5) data (MISAR-seq), learned latent representations.** Learned latent representations of different methods for real MISAR-seq data. The latent representations are visualized with UMAP and colors represent ground-truth labels provided by the original studies. Detailed ground-truth labels: 0: Cartilage, 1: Cerebellar vermis, 2: DPallm, 3: DPallv, 4: Diencephalon, 5: Hindbrain, 6: Mesenchyme, 7: Midbrain, 8: Muscle, 9: Skull, 10: Subpallium, 11: Thalamus.

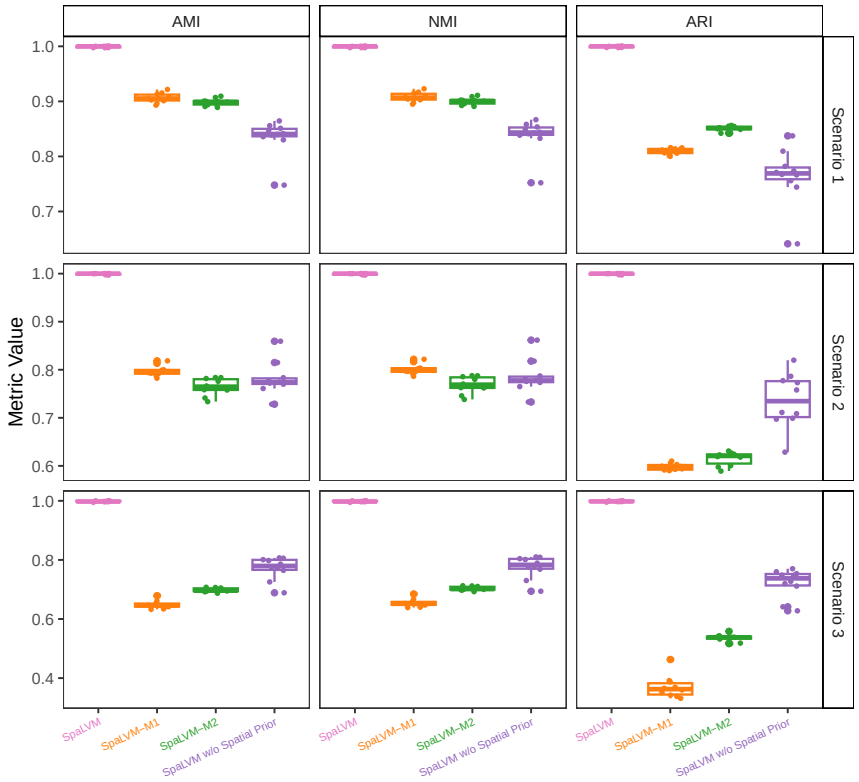

Figure S29: **Ablation results on Simulation Study I.** Boxplots of AMI, NMI, and ARI for SpaLVM, SpaLVM-M1, SpaLVM-M2, and SpaLVM w/o Spatial Prior on the Scenario 1, Scenario 2, and Scenario 3. SpaLVM-M1 and SpaLVM-M2 use only the first or the second modality, respectively, whereas SpaLVM w/o Spatial Prior replaces the spatial covariance $\boldsymbol{\Sigma}$ with the identity matrix $\boldsymbol{I}_N$. Each box summarizes results from ten runs.

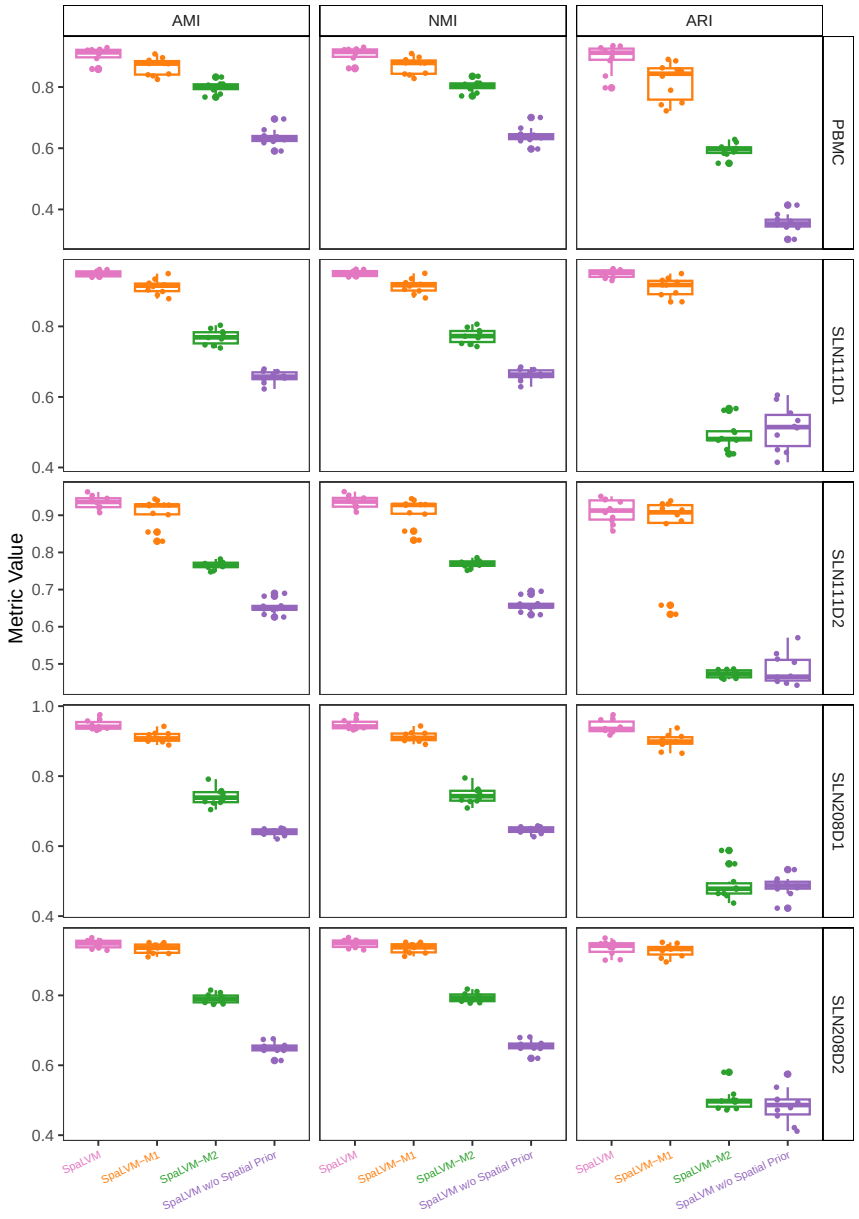

Figure S30: **Ablation results on the CITE-seq benchmarks (Scenario 4).** Boxplots of AMI, NMI, and ARI for SpaLVM, SpaLVM-M1, SpaLVM-M2, and SpaLVM w/o Spatial Prior on the PBMC, SLN111D1, SLN111D2, SLN208D1, and SLN208D2 datasets. SpaLVM-M1 and SpaLVM-M2 use only the first or the second modality, respectively, whereas SpaLVM w/o Spatial Prior replaces the spatial covariance $\boldsymbol{\Sigma}$ with the identity matrix $\boldsymbol{I}_N$. Each box summarizes results from ten runs.

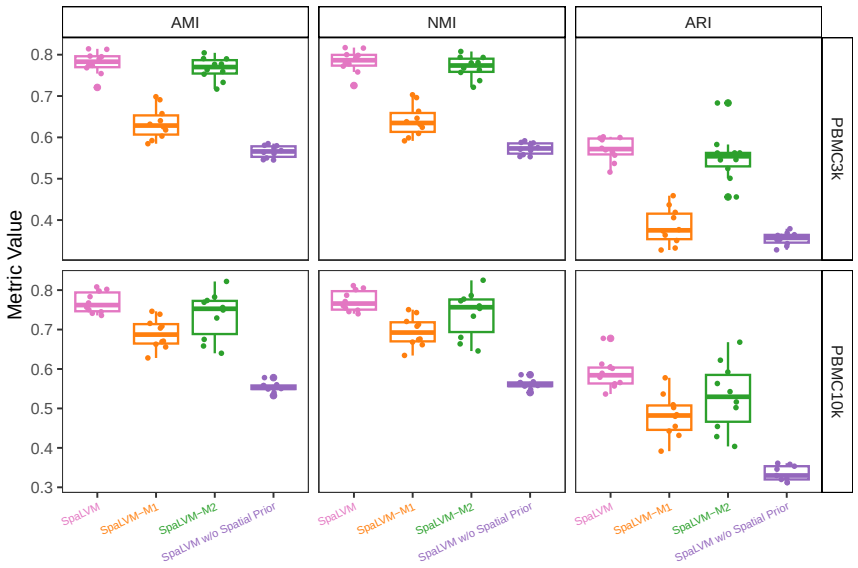

Figure S31: **Ablation results on the SMAGE-seq benchmarks (Scenario 5).** Boxplots of AMI, NMI, and ARI for SpaLVM, SpaLVM-M1, SpaLVM-M2, and SpaLVM w/o Spatial Prior on the PBMC3k and PBMC10k datasets. SpaLVM-M1 and SpaLVM-M2 use only the first or the second modality, respectively, whereas SpaLVM w/o Spatial Prior replaces the spatial covariance $\boldsymbol{\Sigma}$ with the identity matrix $\boldsymbol{I}_N$. Each box summarizes results from ten runs.

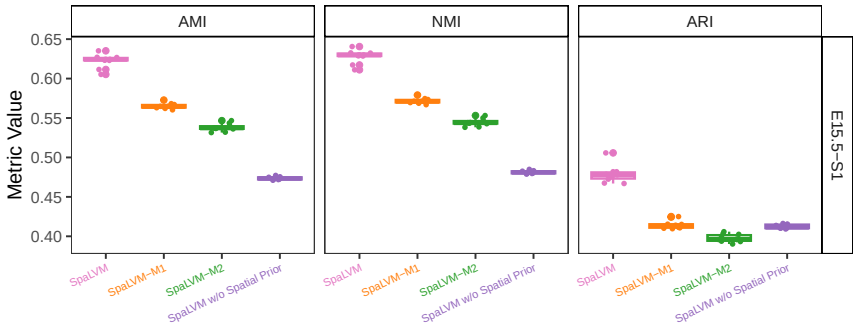

Figure S32: **Ablation results on the MISAR-seq dataset.** Boxplots of AMI, NMI, and ARI for SpaLVM, SpaLVM-M1, SpaLVM-M2, and SpaLVM w/o Spatial Prior on the mouse embryonic brain dataset. SpaLVM-M1 and SpaLVM-M2 use only the first or the second modality, respectively, whereas SpaLVM w/o Spatial Prior replaces the spatial covariance $\boldsymbol{\Sigma}$ with the identity matrix $\boldsymbol{I}_N$. Each box summarizes results from ten runs.