COSMIC: TARGET-CONDITIONED CROSS-MODAL INTERACTION AND CLUSTER-GUIDED MULTI-SLICE ALIGNMENT FOR SPATIAL MULTI-OMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-slice spatial multi-omics integration seeks coherent tissue representations that preserve complementary molecular signals and biological heterogeneity. However, cross-modal fusion and cross-slice alignment are tightly coupled: reliable correspondence requires robust multimodal semantics, while learning such semantics is itself confounded by slice-specific shifts. Existing methods largely optimize one side first, propagating fusion noise or overcorrecting noncorresponding tissue regions. To address this limitation, we propose COSMIC, an end-to-end framework that couples target-conditioned consensus learning with cluster-guided conditional alignment in a shared latent space. Multimodal semantics guide cross-slice correspondence and are jointly refined by alignment, while target-conditioned cross-modal interaction forms a reliability-weighted consensus that preserves modality-private information. Based on the resulting representation, cluster-guided conditional alignment uses soft semantic assignments and confidence filtering to guide the alignment of well-supported slice-specific semantic centers with their global semantic counterparts. This conditional strategy avoids indiscriminate matching of entire slice distributions and reduces overcorrection of noncorresponding tissue structures. Through comprehensive benchmarks across multiple species, COSMIC achieves highly competitive performance while maintaining a favorable balance between biological structure preservation and slice-effect removal, thereby facilitating comparative analyses across heterogeneous spatial multi-omics tissue sections.

1 INTRODUCTION

Spatial omics technologies have progressed from measuring single transcriptomic layers to profiling multiple molecular modalities while preserving spatial coordinates(Liu et al., 2020), enabling characterization of tissue organization and local microenvironments(Lv et al., 2024). However, biological studies often involve multiple tissue sections(Zeira et al., 2022) from different individuals, anatomical regions, or experimental batches(Zhou et al., 2023). Effective integration must therefore not only fuse heterogeneous omics modalities within each section(Zhou et al., 2025) but also learn comparable representations across sections(Long et al., 2022), supporting cross-sample spatial domain identification and microenvironment discovery.

Existing methods generally address multimodal fusion and cross-slice alignment separately. Most spatial multi-omics models primarily integrate modalities within individual sections, whereas conventional multi-slice integration methods mainly target single-omics data or align sections independently of multimodal fusion(Bienroth et al., 2025). Although a few recent approaches extend integration to multiple spatial multi-omics samples(Lou et al., 2025), they do not explicitly use multimodal consensus semantics to determine which tissue regions should be aligned across sections.

Bridging this gap is challenging because cross-modal fusion and cross-slice alignment are interdependent. Different modalities provide complementary but heterogeneous descriptions of the same microenvironment(Tang et al., 2024), so indiscriminate aggregation may propagate technical noise or obscure modality-specific variation(Cai et al., 2026). Across sections, these uncertainties are compounded by the entanglement of technical shifts with genuine differences in tissue composition

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.

and spatial architecture(Xu et al., 2023). The central challenge is therefore to jointly learn reliable multimodal semantics and distinguish true cross-section correspondence from modality- and section-specific variation (Liu et al., 2023b).

To address this challenge, we propose COSMIC, an end-to-end framework for multi-slice spatial multi-omics integration and clustering. COSMIC learns multimodal consensus semantics to identify biologically corresponding regions and restricts cross-slice alignment to reliable structures. It first employs target-conditioned cross-modal interaction and shared-private disentanglement to capture complementary and modality-specific information in a joint representation. It then uses cluster-guided conditional alignment to derive soft semantic assignments and selectively calibrate high-confidence slice-specific centers against their global counterparts. Joint optimization allows multimodal semantics to guide alignment while reducing overcorrection.

• We propose COSMIC, an end-to-end framework that couples multimodal consensus learning with cross-slice alignment in a shared latent space, allowing multimodal semantics to guide correspondence estimation while alignment regularizes representation learning.

• We develop target-conditioned cross-modal interaction, where each modality selectively retrieves complementary evidence from the others and drives the subsequent synthesis of a reliability-weighted shared representation while retaining modality-private information.

• We introduce cluster-guided conditional alignment, which uses soft semantic assignments and within-slice confidence to align only well-supported slice-specific centers with their global counterparts, avoiding indiscriminate global matching.

2 Related Work

2.1 Spatial Multi-omics Integration

Spatial multi-omics integration jointly models multiple molecular layers, such as transcriptomic and proteomic measurements, at shared tissue locations to obtain a more comprehensive representation of the tissue microenvironment. Existing methods typically use graph neural networks, attention mechanisms, or autoencoders to map heterogeneous omics data into a shared latent space (Huang et al., 2025; Yu et al., 2025). SpatialGlue (Long et al., 2024) uses a dual-attention graph neural network to fuse cross-modal correspondences; GROVER (Xiao et al., 2026) combines mixture-of-experts routing with graph convolution for adaptive feature integration; and MISO (Coleman et al., 2025) incorporates histological images into joint multi-omics modeling. These methods demonstrate that multimodal information can improve spatial domain identification and cellular-state characterization (Mo et al., 2026). However, most focus on single-slice settings and rarely address slice-specific shifts, uncertain local correspondence, or semantic consistency across slices. Extending spatial multi-omics integration to multi-slice settings is therefore necessary for stable joint analysis across tissue sections.

2.2 Multi-slice Spatial Multi-omics Integration and Alignment

Multi-slice spatial multi-omics integration must establish a comparable representation space across tissue sections while preserving tissue architecture and biological heterogeneity (Guo et al., 2023; Duan et al., 2024). Early methods, including MOFA+ (Argelaguet et al., 2020), MultiVI (Ashuach et al., 2023), and TotalVI (Gayoso et al., 2021), primarily characterized molecular heterogeneity in non-spatial single-cell data and therefore lacked the spatial context needed to resolve tissue architecture. Recent spatial approaches, such as PRECAST (Liu et al., 2023a), GRASS (Gui et al., 2025), and spCLUE (Wang et al., 2025), incorporate spatial information for multi-slice alignment and domain identification, but remain limited to single-omics data and cannot model complex cross-modal interactions. More recently, SMART-MS (Du et al., 2026) and PRESENT (Li et al., 2026) have explored multi-slice or multi-sample spatial multi-omics integration. Nevertheless, existing methods still struggle to jointly ensure cross-slice semantic consistency, complementary multimodal interaction, and tissue heterogeneity preservation: global alignment may mix semantically distinct regions, whereas local anchors are sensitive to imbalanced slice sizes and noise. COSMIC addresses these challenges through contextual cross-modal interaction and cluster-conditional center alignment, enabling adaptive fusion and alignment of multi-slice spatial multi-omics data.

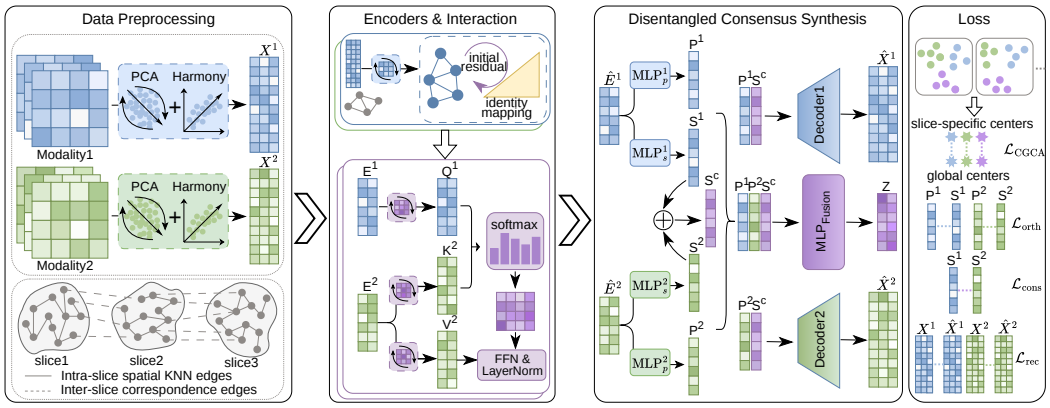

Figure 1: The framework of the proposed COSMIC, which couples target-conditioned consensus learning with cluster-guided conditional alignment: multi-slice graph encoding captures intra-slice topology and inter-slice correspondences, target-conditioned cross-modal interaction extracts complementary semantics that undergo disentangled consensus synthesis into shared and private subspaces to form the unified representation $\mathbf{Z}$, and soft semantic assignments with confidence-aware center construction enable conditional cross-slice calibration between local and global centers, with all modules jointly optimized under a unified objective.

3 METHOD

Given a multi-slice spatial multi-omics dataset comprising B tissue sections and N spots, each spot i has spatial coordinates and a slice label, denoted by $\mathcal{S} = \{(x_i, y_i, b_i)\}_{i=1}^{N}$, each spot also has observed features from M omics modalities, denoted by $\mathcal{X} = \{\mathbf{x}_i^{(m)} \in \mathbb{R}^{D_m}\}_{i=1,m=1}^{N,M}$. COSMIC aims to learn a unified low-dimensional representation:

$$\mathbf{Z} = \Phi(\mathcal{X}, \mathcal{S}) \in \mathbb{R}^{N \times d_h}, \tag{1}$$

where $\mathbf{Z}$ preserves spatial topology and complementary multimodal semantics while reducing slice-specific variation.

As illustrated in Figure 1, COSMIC comprises two coupled mechanisms. First, target-conditioned cross-modal interaction and disentangled consensus synthesis derive topology-aware multimodal representations. Second, cluster-guided conditional alignment uses the learned soft semantics to selectively calibrate well-supported slice-specific centers against their global semantic centers. The two mechanisms are jointly optimized in a shared latent space, allowing multimodal semantics to guide cross-slice alignment while alignment regularizes representation learning.

3.1 TARGET-CONDITIONED CONSENSUS LEARNING

Target-conditioned consensus learning transforms the multimodal observations into a unified latent representation through three sequential operations. It first encodes each modality over a shared multi-slice topology, then performs target-conditioned cross-modal interaction, and finally decomposes and synthesizes the disentangled shared and private components into a multimodal consensus representation.

Multi-Slice Graph Encoding. To capture spatial continuity in complex microenvironments and mitigate batch effects, COSMIC constructs a unified joint multi-slice topology graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Within each slice, the model constructs a KNN spatial graph from the two-dimensional coordinates:

$$A_{ij}^{\text{intra}} = \mathbb{I}\left[j \in \mathcal{N}_{k_s}^{\text{sp}}(i) \text{ or } i \in \mathcal{N}_{k_s}^{\text{sp}}(j)\right], \tag{2}$$

where $\mathcal{N}_{k_s}^{\text{sp}}(i)$ denotes the k_s spatial neighbors of spot i within the same slice.

For $i \in \mathcal{I}_b$ and $j \in \mathcal{I}_{b'}$ from different slices, the inter-slice MNN adjacency is defined as:

$$A_{ij}^{\text{mnn}} = \mathbb{I}\left[j \in \mathcal{N}_{k_b}^{\text{rep}}(i; \mathcal{I}_{b'}) \text{ and } i \in \mathcal{N}_{k_{b'}}^{\text{rep}}(j; \mathcal{I}_b)\right], \tag{3}$$

where $\mathcal{N}^{\mathrm{rep}}$ denotes the neighborhood set in the representation space used for cross-slice matching.

To avoid matching bias caused by a fixed number of MNN neighbors when slice sizes are imbalanced, COSMIC introduces an adaptive scaling factor $\rho_{b,b'} = \min(|\mathcal{I}_b|, |\mathcal{I}_{b'}|)/\max(|\mathcal{I}_b|, |\mathcal{I}_{b'}|)$. It then uses the spatial geometric features $[\mathbf{s}_i, \mathbf{s}_j, \mathbf{s}_i - \mathbf{s}_j]$ to filter anomalous connections, producing a high-confidence inter-slice anchor graph $\mathbf{A}_{\mathrm{inter}}$. The final joint adjacency matrix is:

$$\mathbf{A} = \mathbf{A}^{\mathrm{intra}} + w_g \mathbf{A}^{\mathrm{inter}}. \tag{4}$$

The same topology is used to encode all modalities. For modality m, a multilayer graph convolutional encoder with initial residual connections and identity mapping propagates $\mathbf{X}^{(m)}$ over $\mathbf{A}$. The encoder produces a topology-aware representation $\mathbf{E}^{(m)} \in \mathbb{R}^{N \times d_h}$. Consequently, all modality representations share the same graph structure while retaining modality-specific feature transformations.

Target-Conditioned Cross-Modal Interaction. To capture complementary dependencies across modalities, COSMIC uses a Target-Conditioned Cross-Modal Interaction (TCCI) mechanism. For modality m, its topology-aware representation serves as the target query, while the remaining modalities are aggregated into a cross-modal context:

$$\mathbf{C}^{(m)} = \frac{1}{M-1} \sum_{r \neq m} \mathbf{E}^{(r)}. \tag{5}$$

The target representation $\mathbf{E}^{(m)}$ is used as the query, whereas the context $\mathbf{C}^{(m)}$ aggregates complementary information from the remaining modalities and provides the key and value features:

$$\mathbf{O}^{(m)} = \mathrm{MHCA}(\mathbf{E}^{(m)}, \mathbf{C}^{(m)}, \mathbf{C}^{(m)}). \tag{6}$$

Specifically, for attention head h, the model computes scaled dot-product attention using independent parameter matrices, enabling adaptive extraction of cross-modal information:

$$\mathrm{head}_h^{(m)} = \mathrm{Softmax}\left(\frac{\mathbf{Q}_h^{(m)} \left(\mathbf{K}_h^{(m)}\right)^\top}{\sqrt{d_k}} \right) \mathbf{V}_h^{(m)}, \tag{7}$$

where $\mathbf{Q}_h^{(m)} = \mathbf{E}^{(m)} \mathbf{W}_Q^{(h)}$, $\mathbf{K}_h^{(m)} = \mathbf{C}^{(m)} \mathbf{W}_K^{(h)}$, and $\mathbf{V}_h^{(m)} = \mathbf{C}^{(m)} \mathbf{W}_V^{(h)}$ denote the query, key, and value projections for attention head h, respectively.

The head outputs are concatenated and linearly projected to form the interaction result $\mathbf{O}^{(m)}$. We then use two residual updates with an FFN and layer normalization to integrate cross-modal context while retaining the target-modality representation:

$$\widetilde{\mathbf{E}}^{(m)} = \mathrm{LN}\left(\mathbf{E}^{(m)} + \mathbf{O}^{(m)}\right), \qquad \widehat{\mathbf{E}}^{(m)} = \mathrm{LN}\left(\widetilde{\mathbf{E}}^{(m)} + \mathrm{FFN}\left(\widetilde{\mathbf{E}}^{(m)}\right)\right). \tag{8}$$

This process preserves an independent output branch for each modality, yielding the context-enhanced representations $\{\widehat{\mathbf{E}}^{(m)}\}_{m=1}^{M}$.

Disentangled Consensus Synthesis. To prevent multimodal commonality from becoming entangled with modality-specific technical noise, independent projection heads are used to explicitly decompose each context-enhanced representation into a shared subspace $\mathbf{S}^{(m)}$ and a private subspace $\mathbf{P}^{(m)}$:

$$\mathbf{S}^{(m)} = \phi_{\mathrm{s}}^{(m)}(\widehat{\mathbf{E}}^{(m)}), \quad \mathbf{P}^{(m)} = \phi_{\mathrm{p}}^{(m)}(\widehat{\mathbf{E}}^{(m)}). \tag{9}$$

Here, $\mathbf{S}^{(m)} \in \mathbb{R}^{N \times d_h}$ captures biological semantics shared across modalities, whereas $\mathbf{P}^{(m)} \in \mathbb{R}^{N \times d_h}$ preserves technical or molecular attributes specific to that modality. An orthogonality constraint reduces information redundancy between shared and private representations, while a consistency constraint brings the shared semantics of different modalities closer together.

The modality-specific shared representations $\{\mathbf{S}^{(m)}\}_{m=1}^{M}$ are adaptively weighted through a learnable channel-wise gating mechanism to form the consensus backbone $\mathbf{S}^c$. The consensus backbone is then concatenated with all private representations as $\mathbf{Z}^c = [\mathbf{S}^c \parallel \mathbf{P}^{(1)} \parallel \cdots \parallel \mathbf{P}^{(M)}]$ and mapped through a nonlinear fusion network to obtain $\mathbf{Z} \in \mathbb{R}^{N \times d_h}$. To preserve the input information during disentanglement, each modality is reconstructed from $\mathbf{U}^{(m)} = [\mathbf{S}^c \parallel \mathbf{P}^{(m)}]$ using a modality-specific decoder, with the reconstruction loss $\mathcal{L}_{\text{rec}} = \sum_{m=1}^{M} \|\mathbf{X}^{(m)} - \widehat{\mathbf{X}}^{(m)}\|_F^2$.

3.2 Cluster-Guided Conditional Alignment

Cluster-Guided Conditional Alignment (CGCA) operates on the joint representation $\mathbf{Z}$. It first estimates soft semantic assignments, constructs global and slice-specific centers from these assignments, and then calibrates valid slice-specific centers toward their corresponding global centers.

Soft Semantic Assignment. A global alignment objective can overcorrect the data by forcing semantically distinct tissue regions to mix across slices. COSMIC instead applies a clustering projection head $g_{\text{clu}}(\cdot)$ to the learned joint representation $\mathbf{Z}$ to obtain soft semantic assignments:

$$\mathbf{Q} = \text{Softmax}\left(g_{\text{clu}}(\mathbf{Z})\right) \in \mathbb{R}^{N \times K}, \tag{10}$$

where K is the number of latent semantic clusters and Q_{ik} denotes the probability that spot i belongs to cluster k. In contrast to hard labels, $\mathbf{Q}$ captures gradual transitions between tissue states and provides uncertainty-aware weights for subsequent center estimation.

Confidence-Aware Center Construction. Let $\bar{\mathbf{z}}_i = \mathbf{z}_i / \|\mathbf{z}_i\|_2$ denote the ℓ_2-normalized representation of spot i. COSMIC first estimates a global semantic center for cluster k by aggregating all spots across slices. For slice b, the corresponding slice-specific center is defined jointly as:

$$\mu_k = \text{Norm}\left(\frac{\sum_{i=1}^{N} Q_{ik}\bar{\mathbf{z}}_i}{\sum_{i=1}^{N} Q_{ik} + \epsilon}\right), \qquad \mu_{b,k} = \text{Norm}\left(\frac{\sum_{i\in\mathcal{I}_b} Q_{ik}\bar{\mathbf{z}}_i}{\sum_{i\in\mathcal{I}_b} Q_{ik} + \epsilon}\right). \tag{11}$$

Here, $\text{Norm}(\cdot)$ denotes ℓ_2 normalization and ϵ is a numerical stabilizer.

Not every slice–cluster pair provides a reliable alignment signal. COSMIC therefore retains a pair (b, k) only if its cumulative assignment mass $\sum_{i\in\mathcal{I}_b} Q_{ik}$ exceeds a predefined threshold η. The resulting valid set Ω excludes sparse, rare, or uncertain slice-specific clusters from the alignment objective.

Conditional Cross-Slice Calibration. For each valid pair $(b, k) \in \Omega$, COSMIC calibrates the local center $\boldsymbol{\mu}_{b,k}$ toward the corresponding global semantic anchor $\boldsymbol{\mu}_k$ using cosine distance:

$$\mathcal{L}_{\text{CGCA}} = \frac{1}{|\Omega|} \sum_{(b,k)\in\Omega} \left(1 - \cos(\mu_{b,k}, \text{sg}(\mu_k))\right), \tag{12}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operation. The global center is thus treated as a stable semantic anchor in this alignment branch, while each slice-specific center is calibrated toward it. In addition, $\mathbf{Q}$ is detached when used as the weighting matrix in this loss, preventing alignment gradients from directly distorting the semantic assignment itself.

This conditional calibration removes slice-specific shifts only within well-supported semantic structures. It therefore avoids indiscriminate distribution matching and reduces the risk of overcorrecting genuinely distinct tissue microenvironments.

3.3 Joint Optimization

COSMIC jointly optimizes multimodal reconstruction ($\mathcal{L}_{\text{rec}}$), cluster-guided conditional alignment ($\mathcal{L}_{\text{CGCA}}$), shared-private disentanglement ($\mathcal{L}_{\text{orth}}$), and consensus consistency ($\mathcal{L}_{\text{cons}}$). To bal-

Table 1: Performance comparison of COSMIC and baselines on the HLN and HT datasets. Bold indicates the best result and underline denotes the second-best. Iso. denotes isolated labels, ASW denotes average silhouette width, and Graph conn. denotes graph connectivity.

Metric	SMART-MS	stMSA	SpaMosaic	PRESENT	SpaMode	stClinic	stMixer	COSMIC
HLN dataset								
Iso. labels	0.519	0.487	0.443	**0.527**	0.500	0.499	0.430	0.505
K-means NMI	0.350	0.235	0.366	0.328	0.223	0.340	0.183	**0.381**
K-means ARI	0.212	0.132	0.216	0.192	0.123	0.218	0.090	**0.256**
ASW label	**0.501**	0.485	0.495	0.495	0.494	0.466	0.410	0.501
cLISI	**0.921**	0.898	0.908	0.920	0.856	0.907	0.853	0.918
ASW batch	0.944	0.869	0.913	**0.946**	0.946	0.888	0.882	0.939
iLISI	0.726	0.163	**0.794**	0.677	0.600	0.273	0.787	0.755
kBET	0.742	0.018	**0.843**	0.686	0.249	0.281	0.420	0.788
Graph conn.	0.720	0.440	0.673	**0.759**	0.506	0.576	0.412	0.708
Bio. conservation	0.501	0.447	0.486	0.492	0.439	0.486	0.393	**0.512**
Batch correction	0.783	0.373	**0.806**	0.767	0.575	0.505	0.625	0.798
Total	0.614	0.417	0.614	0.602	0.494	0.493	0.486	**0.626**
HT dataset								
Iso. labels	0.551	0.491	0.562	0.560	0.532	0.547	0.524	**0.565**
K-means NMI	0.233	0.145	0.241	0.226	0.233	0.243	0.199	**0.278**
K-means ARI	0.121	0.066	0.126	0.119	0.135	0.153	0.093	**0.177**
ASW label	0.496	0.481	0.494	0.516	0.508	0.526	0.453	**0.533**
cLISI	0.867	0.842	0.872	**0.873**	0.835	0.872	0.807	0.870
ASW batch	0.936	0.861	0.920	**0.965**	0.959	0.945	0.917	0.949
iLISI	0.717	0.166	**0.822**	0.677	0.411	0.737	0.803	0.711
kBET	0.695	0.002	**0.781**	0.589	0.057	0.576	0.663	0.700
Graph conn.	0.879	0.574	0.836	**0.915**	0.838	0.622	0.809	0.912
Bio. conservation	0.454	0.405	0.459	0.459	0.449	0.468	0.415	**0.485**
Batch correction	0.807	0.401	**0.840**	0.787	0.566	0.720	0.798	0.818
Total	0.595	0.403	0.611	0.590	0.496	0.569	0.568	**0.618**

ance objectives with different numerical scales and convergence rates, we employ homoscedastic-uncertainty-based adaptive weighting:

$$\mathcal{L}_{\text{total}} = \sum_{i=1}^{4} \left[\exp(-s_i)\mathcal{L}_i + s_i\right], \tag{13}$$

where $\{s_i\}_{i=1}^{4}$ are learnable parameters. This dynamic weighting stabilizes end-to-end training, allowing multimodal semantics to guide robust cross-slice alignment without sacrificing biological heterogeneity.

4 Experiments

4.1 Experimental Setups

Datasets. We evaluated COSMIC on four multi-slice spatial multi-omics benchmarks: human lymph node (HLN) and human tonsil (HT) (Yan et al., 2026), mouse postnatal brain (MPB) (Zhang et al., 2023), and mouse thymus (MT) (Liao et al., 2023). Each dataset contains three tissue sections. Together, they cover human and mouse tissues, annotated and unannotated settings, RNA-ADT and RNA-ATAC modalities, and three spatial profiling platforms. Dataset and preprocessing details are provided in the Appendix.

Table 2: Performance comparison of COSMIC and baselines on the MT and MPB datasets. Bold indicates the best result and underline denotes the second-best.

Metric	SMART-MS	stMSA	SpaMosaic	PRESENT	SpaMode	stClinic	stMixer	COSMIC
MT dataset								
cLISI	0.935	0.951	0.994	0.933	0.818	**1.000**	0.999	0.985
ASW cluster	0.566	0.569	0.664	0.562	0.527	**0.719**	0.663	0.605
Moran's I	0.423	0.314	0.421	0.370	0.273	**0.471**	0.250	0.439
ASW slice	0.498	0.497	0.497	0.499	0.499	**0.503**	0.493	0.498
iLISI	0.810	0.679	0.822	0.743	0.821	0.283	**0.855**	0.783
Domain conservation	0.641	0.611	0.693	0.622	0.539	**0.730**	0.637	0.676
Batch correction	0.654	0.588	0.660	0.621	0.660	0.393	**0.674**	0.641
Total	0.646	0.602	**0.680**	0.621	0.588	0.595	0.652	0.662
MPB dataset								
cLISI	0.951	0.948	0.993	0.943	0.966	**1.000**	0.984	0.981
ASW cluster	0.565	0.551	0.645	0.565	0.568	**0.785**	0.616	0.582
Moran's I	0.339	0.215	0.655	0.253	0.369	**0.658**	0.106	0.575
ASW slice	0.502	0.525	0.483	0.502	0.512	**0.541**	0.493	0.496
iLISI	0.418	0.372	0.317	0.446	0.047	0.058	0.460	**0.509**
Domain conservation	0.618	0.571	0.764	0.587	0.634	**0.814**	0.569	0.713
Batch correction	0.460	0.449	0.400	0.474	0.280	0.300	0.477	**0.503**
Total	0.555	0.522	0.619	0.542	0.492	0.608	0.532	**0.628**

Baselines. We compared COSMIC with seven representative spatial multi-omics and multi-slice methods: SMART-MS (Du et al., 2026), stMSA (Shu et al., 2025), SpaMosaic (Yan et al., 2026), PRESENT (Li et al., 2026), SpaMode (Zheng et al., 2026), stClinic (Zuo et al., 2025), and stMixer (Yang et al., 2026). All baselines were run using their official implementations with the default hyperparameters recommended in the original papers. Results are reported as means over 15 runs.

Metrics. Annotated and unannotated datasets were evaluated separately, with higher scores indicating better performance. For HLN and HT, we used the `scib-metrics` Benchmarker (Luecken et al., 2022); Bio Conservation averages isolated-label, K-means NMI/ARI, label silhouette, and cLISI, whereas Batch Correction averages batch silhouette, iLISI, kBET, and graph connectivity. For MT and MPB, Domain Conservation averages cLISI, Moran's I, and ASWcluster, whereas Batch Correction averages iLISI and ASWslice. Total is the $0.6{:}0.4$ weighted sum of the conservation and Batch Correction scores. Metric definitions are provided in the Appendix.

4.2 Quantitative Experimental Results

Tables 1 and 2 quantitatively compare COSMIC with seven baseline methods along two complementary dimensions: biological or structural conservation and batch-effect removal. Overall performance was measured by their weighted Total score. COSMIC ranked first on HLN (0.626), HT (0.618), and MPB (0.628) and second on MT (0.662), achieving the highest mean Total score across the four benchmarks (0.634). This pattern was consistent across labeled and unlabeled datasets and across RNA-ADT and RNA-chromatin accessibility modalities, indicating that COSMIC effectively balances structure preservation with cross-slice alignment. On the annotated HLN and HT datasets, COSMIC achieved the best Total scores of 0.626 and 0.618, respectively, driven by stronger biological conservation while retaining competitive batch correction. On the unlabeled benchmarks, it ranked first on the size-imbalanced MPB dataset and second on MT, maintaining a favorable balance between domain conservation and cross-slice mixing. These results demonstrate consistently strong integration performance without systematically sacrificing tissue structure for batch removal.

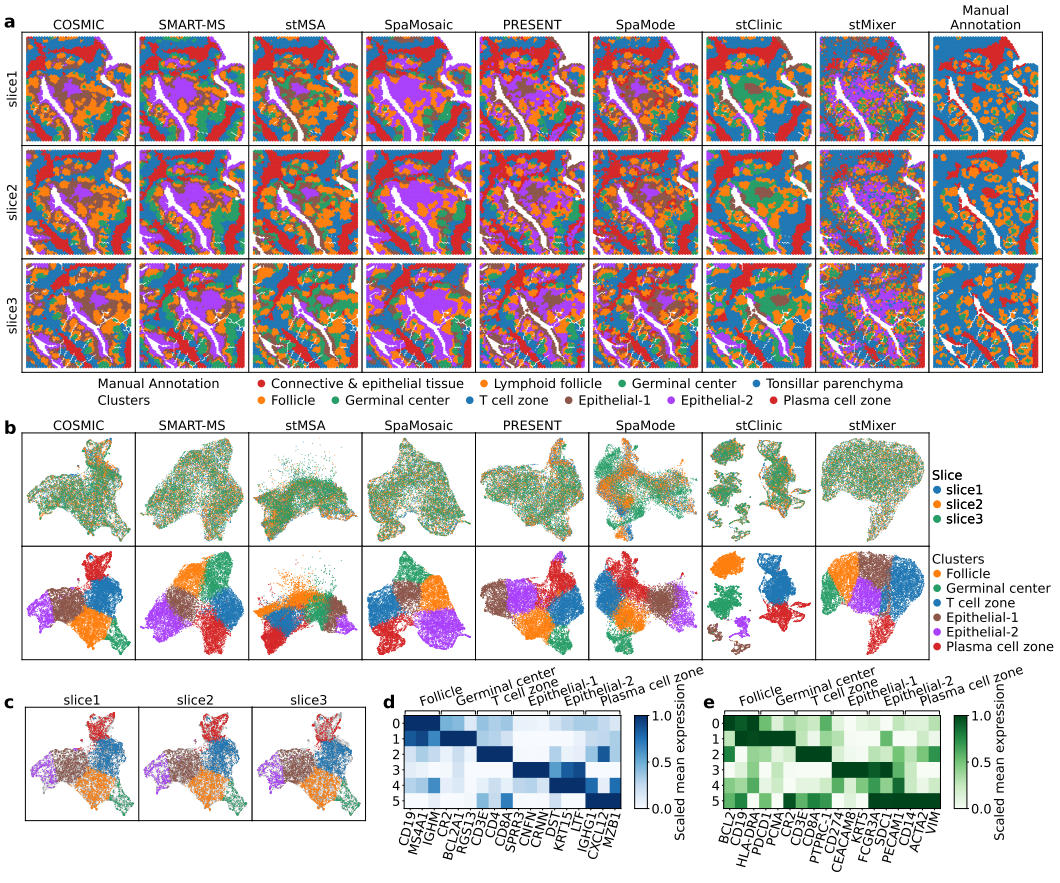

Figure 2: Spatial Domain Analysis on the HT dataset: (a) manual annotations and spatial-domain identification results across three human tonsil slices using baseline methods and COSMIC; (b) UMAP visualizations of the integrated representations produced by COSMIC and baseline methods, colored by slice identity and inferred domain labels, respectively; (c) COSMIC-derived spatial domains visualized separately in each slice, using the same cluster annotations as in (b); (d,e) cluster-wise RNA and ADT marker-expression heatmaps supporting the biological identities of the COSMIC-derived domains.

4.3 Spatial Domain Analysis

The human tonsil contains highly heterogeneous and tightly interleaved compartments, providing a stringent benchmark for evaluating section integration and spatial structure preservation. As shown in Figure 2a, COSMIC recovers the major follicular, germinal-center, and T-cell-rich regions with high spatial coherence. In contrast, stMixer produces noisy, scattered labels, and SpaMode exhibits fragmented regional boundaries. UMAP visualizations (Figure 2b,c) further show that COSMIC substantially mitigates section-associated variation, achieving broad mixing across all three slices while maintaining compact, distinguishable clusters and comparable domain representation across individual sections. By comparison, SpaMode retains prominent section-associated artifacts in the latent space, and stMSA exhibits diffuse and poorly separated clusters. COSMIC identifies six molecularly distinguishable domains, whose annotations are supported primarily by domain-specific RNA programs and complemented by ADT enrichment profiles (Figure 2d,e). Specifically, the follicular domain exhibits enrichment of the B-cell markers *CD19* and *MS4A1*, whereas the T-cell zone is characterized by elevated *CD3E* and *CD8A*. Together with the observed spatial coherence and cross-slice consistency, these molecular signatures demonstrate that COSMIC produces an integrated yet biologically interpretable representation of the tonsil sections. Additional HLN qualitative results are detailed in the Appendix.

4.4 Ablation Studies

We assessed the contributions of Target-Conditioned Cross-Modal Interaction and Cluster-Guided Conditional Alignment using two variants, each removing one mechanism (Table 3). COSMIC w/o TCCI removes the target-conditioned cross-modal interaction process, whereas COSMIC w/o CGCA removes the cluster-guided conditional alignment loss. On HT, removing TCCI reduced Bio Conservation and Total from 0.618 to 0.600, while removing CGCA primarily weakened Batch Correction, lowering Total to 0.566. On MPB, the two variants mainly degraded Domain Conservation and Batch Correction, respectively, yielding Total scores of 0.609 and 0.614 versus 0.628 for the full model. These results indicate complementary roles: TCCI preserves biological and spatial structure by extracting complementary cross-modal semantics and mitigating modality-specific noise, whereas CGCA improves cross-slice mixing and the conservation–correction balance. Detailed metrics are provided in the Appendix.

Table 3: Ablation results on the HT and MPB datasets. Bold indicates the best result.

HT dataset			
Metric	w/o TCCI	w/o CGCA	Full
Bio. conservation	0.464	0.479	**0.485**
Batch correction	0.805	0.696	**0.818**
Total	0.600	0.566	**0.618**
MPB dataset			
Metric	w/o TCCI	w/o CGCA	Full
Domain conservation	0.682	0.711	**0.713**
Batch correction	0.500	0.469	**0.503**
Total	0.609	0.614	**0.628**

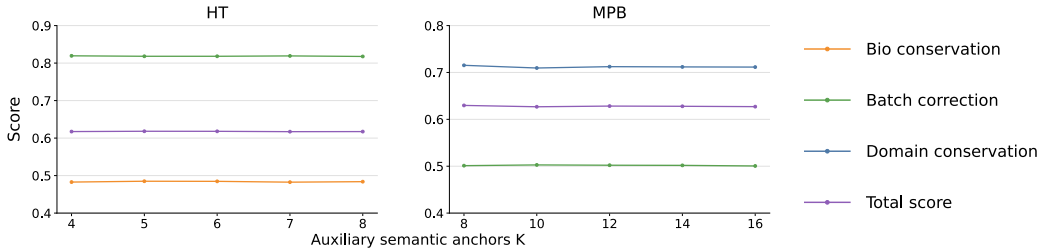

Figure 3: Parameter sensitivity analysis on the HT and MPB datasets.

4.5 Parameter Sensitivity Experiments

To evaluate the robustness of COSMIC in practice, we examined its sensitivity to the number of auxiliary semantic anchors, K, in the cluster-guided conditional alignment module (Figure 3). Here, K determines the dimensionality of the soft semantic assignment matrix Q and the granularity of confidence-aware center construction; it is not the final number of ground-truth cell-type clusters. On HT, the Total score remained between 0.617 and 0.618 for $K = 4$–8; both the best value ($K = 5$) and the default value ($K = 6$) achieved 0.618. On MPB, scores ranged from 0.627 to 0.630 for $K = 8$–16; $K = 8$ performed best (0.630), while the default value ($K = 12$) achieved 0.628. These small variations indicate robustness to the predefined number of semantic centers without requiring an exact match to the true number of cell types, supporting empirical selection of K without extensive tuning. Detailed metrics are provided in the Appendix.

5 Conclusion

We presented COSMIC for multi-slice spatial multi-omics integration. By coupling target-conditioned cross-modal interaction with cluster-guided conditional alignment, it uses reliable shared semantics for selective cross-slice alignment. Across four benchmarks spanning species, platforms, annotation settings, and modalities, COSMIC balanced tissue-structure preservation with slice-effect removal. These results suggest that robust correspondence can emerge from learned consensus while selective alignment preserves meaningful heterogeneity.

AI USE STATEMENT

Generative AI tools were used solely for grammar and language polishing. All content was verified by the authors.

REFERENCES

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

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

Denis Bienroth, Natalie Charitakis, Dillon Wong, Yunhan C Zhang, Sabrina Jaeger-Honz, Jialin Ding, Kevin I Watt, Julian Stolper, Hazel Chambers-Smith, Duncan MacGregor, et al. Automated integration of multi-slice spatial transcriptomics data in 2d and 3d using vr-omics. *Genome Biology*, 26(1):182, 2025.

Guangchang Cai, Fuqun Chen, Kepei Wen, Ying Li, and Le Ou-Yang. Identifying spatial domains from spatial multi-omics data using consistent and specific deep subspace learning. *Information Fusion*, 125:103428, 2026.

Kyle Coleman, Amelia Schroeder, Melanie Loth, Daiwei Zhang, Jeong Hwan Park, Ji-Youn Sung, Niklas Blank, Alexis J Cowan, Xuyu Qian, Jianfeng Chen, et al. Resolving tissue complexity by multimodal spatial omics modeling with miso. *Nature methods*, 22(3):530–538, 2025.

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

Bin Duan, Shaoqi Chen, Xiaojie Cheng, and Qi Liu. Multi-slice spatial transcriptome domain analysis with spado. *Genome biology*, 25(1):73, 2024.

Adam Gayoso, Zoë 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.

Yang Gui, Zhaorui Tan, Yan Xu, and Chunzhong Li. Heterogeneous graph contrastive learning for integration and alignment of spatial transcriptomics data. *Briefings in Bioinformatics*, 26(5): bbaf497, 2025.

Tiantian Guo, Zhiyuan Yuan, Yan Pan, Jiakang Wang, Fengling Chen, Michael Q Zhang, and Xiangyu Li. Spiral: integrating and aligning spatially resolved transcriptomics data across different experiments, conditions, and technologies. *Genome Biology*, 24(1):241, 2023.

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

Zhen Li, Xuejian Cui, Xiaoyang Chen, Zijing Gao, Yuyao Liu, Yan Pan, Shengquan Chen, Hairong Lv, Lei Zhai, and Rui Jiang. Cross-modality representation and multi-sample integration of spatially resolved omics data. *Briefings in Bioinformatics*, 27(3):bbag214, 2026.

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

Wei Liu, Xu Liao, Ziye Luo, Yi Yang, Mai Chan Lau, Yuling Jiao, Xingjie Shi, Weiwei Zhai, Hongkai Ji, Joe Yeong, et al. Probabilistic embedding, clustering, and alignment for integrating spatial transcriptomics data with precast. *Nature communications*, 14(1):296, 2023a.

Xinhao Liu, Ron Zeira, and Benjamin J Raphael. Partial alignment of multislice spatially resolved transcriptomics data. *Genome Research*, 33(7):1124–1132, 2023b.

Yang Liu, Mingyu Yang, Yanxiang Deng, Graham Su, Archibald Enninful, Cindy C Guo, Toma Tebaldi, Di Zhang, Dongjoo Kim, Zhiliang Bai, et al. High-spatial-resolution multi-omics sequencing via deterministic barcoding in tissue. *Cell*, 183(6):1665–1681, 2020.

Yahui Long, Kok Siong Ang, Mengwei Li, Kian Long Kelvin Chong, Raman Sethi, Chengwei Zhong, Hang Xu, Zhiwei Ong, Karishma Sachaphibulkij, Ao Chen, et al. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with graphst. *BioRxiv*, pp. 2022–08, 2022.

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

Yikai Lou, Xuan Li, Qixing Yang, Hao Dai, Kaiyue Ma, and Chunman Zuo. Vector-guided graph learning for spatial multi-slice multi-omics alignment. *Cell Reports Methods*, 5(12), 2025.

Malte D Luecken, Maren Büttner, Kridsadakorn Chaichoompu, Anna Danese, Marta Interlandi, Michaela F Müller, Daniel C Strobl, Luke Zappia, Martin Dugas, Maria Colomé-Tatché, et al. Benchmarking atlas-level data integration in single-cell genomics. *Nature methods*, 19(1):41–50, 2022.

Tongxuan Lv, Yong Zhang, Junlin Liu, Qiang Kang, and Lin Liu. Multi-omics integration for both single-cell and spatially resolved data based on dual-path graph attention auto-encoder. *Briefings in bioinformatics*, 25(5):bbae450, 2024.

Chenfeng Mo, Xiufen Zou, and Suoqin Jin. Multisp deciphers tissue structure and multicellular communication from spatial multi-omics data. *Cell Genomics*, 2026.

Han Shu, Jing Chen, Chang Xu, Jialu Hu, Yongtian Wang, Jiajie Peng, Qinghua Jiang, Xuequn Shang, and Tao Wang. Efficient integration of spatial omics data for joint domain detection, matching, and alignment with stmsa. *Genome Research*, 35(10):2285–2299, 2025.

Zhenchao Tang, Jiehui Huang, Guanxing Chen, and Calvin Yu-Chian Chen. Comprehensive view embedding learning for single-cell multimodal integration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15292–15300, 2024.

Xiang Wang, Wei Vivian Li, and Hongwei Li. spclue: a contrastive learning approach to unified spatial transcriptomics analysis across single-slice and multi-slice data. *Genome Biology*, 26(1): 177, 2025.

Yongjun Xiao, Dian Meng, Xinlei Huang, Yanran Liu, Shiwei Ruan, Ziyue Qiao, and Xubin Zheng. Grover: Graph-guided representation of omics and vision with expert regulation for adaptive spatial multi-omics fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 1311–1318, 2026.

Hao Xu, Shuyan Wang, Minghao Fang, Songwen Luo, Chunpeng Chen, Siyuan Wan, Rirui Wang, Meifang Tang, Tian Xue, Bin Li, et al. Spacel: deep learning-based characterization of spatial transcriptome architectures. *Nature Communications*, 14(1):7603, 2023.

Xuhua Yan, Zhaoyu Fang, Kok Siong Ang, Lynn van Olst, Alex Edwards, Thomas Watson, Ruiqing Zheng, Di Zhang, Rong Fan, Min Li, et al. Mosaic integration of spatial multi-omics with spamosaic. *Nature Genetics*, pp. 1–12, 2026.

Qixing Yang, Yan Wang, Luonan Chen, and Chunman Zuo. stmixer for scalable mosaic integration and label transfer in spatial histology and multi-omics. *Advanced Science*, pp. e75905, 2026.

Na Yu, Wenrui Li, Xue Sun, Jing Hu, Qi Zou, Zhiping Liu, Daoliang Zhang, Wei Zhang, and Rui Gao. sofusion: facilitating tissue structure identification via spatial multi-omics data fusion. *Briefings in Bioinformatics*, 26(5):bbaf513, 2025.

Ron Zeira, Max Land, Alexander Strzalkowski, and Benjamin J Raphael. Alignment and integration of spatial transcriptomics data. *Nature Methods*, 19(5):567–575, 2022.

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

Xubin Zheng, Xinlei Huang, Xiang Zhou, Dian Meng, Ziyue Qiao, Zhiyuan Yuan, and Lixin Cheng. Spamode: A broadly applicable framework for deciphering spatial multi-omics using multimodal mixture of disentangled experts. *Advanced Science*, pp. e75478, 2026.

Xiang Zhou, Kangning Dong, and Shihua Zhang. Integrating spatial transcriptomics data across different conditions, technologies and developmental stages. *Nature Computational Science*, 3 (10):894–906, 2023.

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

Chunman Zuo, Junjie Xia, Yupeng Xu, Ying Xu, Pingting Gao, Jing Zhang, Yan Wang, and Luonan Chen. stclinic dissects clinically relevant niches by integrating spatial multi-slice multi-omics data in dynamic graphs. *Nature Communications*, 16(1):5317, 2025.

A Appendix

B Dataset Information and Preprocessing

B.1 Dataset Information

To comprehensively evaluate the robustness and broad applicability of COSMIC for multi-slice spatial multi-omics integration, we conducted extensive experiments on four representative benchmark datasets. These datasets span different species and tissue systems, encompassing real-world scenarios both with expert spatial-domain annotations and without labels. They also cover three widely used spatial profiling platforms—10x Visium, spatial ATAC-RNA-seq, and Stereo-CITE-seq—as well as different combinations of modalities, providing a comprehensive evaluation setting:

Human lymph node (HLN) comprises three consecutive sections profiled using the 10x Genomics CytAssist Visium platform. The three sections contain 3,484, 3,359, and 3,408 spots, respectively. Across all sections, the measurement dimensions are strictly identical, comprising 18,085 genes for the RNA modality and 31 proteins for the antibody-derived tags.

Human tonsil (HT) similarly comprises three consecutive sections profiled using the 10x Genomics CytAssist Visium platform, containing 4,326, 4,519, and 4,521 spots, respectively. While the transcriptomic feature dimension remains consistent at 18,085 genes across all sections, the ADT measurement dimensions naturally vary. Specifically, the first two slices measure 31 proteins, whereas the third slice captures 35 proteins.

Mouse postnatal brain (MPB) comprises three coronal brain sections profiled using spatial ATAC-RNA-seq, containing 2,372, 2,497, and 9,215 spots, respectively. Evaluated as an unlabeled dataset, it captures a complex joint modality comprising transcriptomic and epigenetic (chromatin accessibility) measurements. The feature dimensions are highly consistent across all slices, detecting 18,107 genes for the RNA modality and 161,514 peaks for the ATAC modality.

Mouse thymus (MT) comprises three RNA-protein sections profiled using the high-resolution Stereo-CITE-seq platform. Evaluated without spatial labels, the three slices contain 4,253, 4,646, and 4,228 spots, respectively. The sections share a consistent ADT measurement dimension of 19 proteins. Conversely, the transcriptomic modality exhibits natural variation in the number of detected genes across sections: 23,529 genes for slice 1, 23,960 genes for slice 2, and 23,221 genes for slice 3.

B.2 Data Preprocessing

We applied modality-specific preprocessing pipelines to the raw datasets, scaling all final feature matrices to a range of $[0, 1]$ using Min-Max scaling to ensure stable optimization for the COSMIC framework.For the transcriptomic (RNA) modality, we selected the top 3,000 highly variable genes (HVGs), followed by library-size normalization and log-transformation ($\log(1 + x)$). For the protein (ADT) modality, surface protein counts were normalized using the Centered Log-Ratio (CLR) transformation. For the epigenomic (ATAC) modality, the raw chromatin accessibility matrix was binarized, filtered for rare peaks, and normalized using Term Frequency-Inverse Document Frequency (TF-IDF). We then employed Latent Semantic Indexing (LSI) to extract a 50-dimensional latent representation, discarding the first component to mitigate technical variations linked to sequencing depth.Additionally, to account for technical variations across multiple tissue sections, we applied Harmony to the PCA-reduced RNA features to explicitly remove slice-specific batch effects.

C Evaluation Metrics

Let $\mathbf{Z} = \{\mathbf{z}_i\}_{i=1}^N$ denote the learned embedding, y_i the biological annotation or inferred cluster label, and b_i the slice identity of spot i. All reported metrics are scaled so that larger values indicate better performance, unless otherwise specified.

C.1 Metrics for Annotated Datasets

For the annotated HLN and HT datasets, we evaluate biological conservation using isolated-label, Kmeans NMI, Kmeans ARI, label silhouette, and cLISI. Batch correction is evaluated using batch silhouette, iLISI, kBET, and graph connectivity.

Isolated labels. Let $\mathcal{I}$ denote the set of biological labels occurring in the smallest number of slices. The Isolated labels score evaluates how well rare cell types are clustered, computing the label-wise average of rescaled silhouette coefficients:

$$\text{Isolated labels} = \frac{1}{|\mathcal{I}|} \sum_{c \in \mathcal{I}} \frac{1}{N_c} \sum_{i: y_i = c} \frac{s_i^{\text{label}} + 1}{2}, \tag{14}$$

where N_c is the number of spots with label c.

KMeans NMI. KMeans is applied to $\mathbf{Z}$ with a fixed number of clusters. Let Y denote the biological annotations and $\widehat{Y}$ the resulting KMeans assignments. Normalized mutual information is computed using arithmetic normalization:

$$\text{KMeans NMI}(Y, \widehat{Y}) = \frac{2I(Y; \widehat{Y})}{H(Y) + H(\widehat{Y})}, \tag{15}$$

where $I(\cdot;\cdot)$ is mutual information and $H(\cdot)$ is Shannon entropy.

KMeans ARI. Let n_{uv} be the number of spots shared by true class u and predicted cluster v, with $a_u = \sum_v n_{uv}$, $b_v = \sum_u n_{uv}$, and $n = \sum_{u,v} n_{uv}$. To prevent visual overflow, we decompose the adjusted Rand index formula into intermediate terms:

$$\text{RI} = \sum_{u,v} \binom{n_{uv}}{2} \tag{16}$$

$$\mathbb{E}[\text{RI}] = \frac{\sum_u \binom{a_u}{2} \sum_v \binom{b_v}{2}}{\binom{n}{2}} \tag{17}$$

$$\max(\text{RI}) = \frac{1}{2}\left[\sum_u \binom{a_u}{2} + \sum_v \binom{b_v}{2}\right] \tag{18}$$

$$\text{KMeans ARI} = \frac{\text{RI} - \mathbb{E}[\text{RI}]}{\max(\text{RI}) - \mathbb{E}[\text{RI}]}. \tag{19}$$

Silhouette label. For spot i, let a_i denote the mean distance from $\mathbf{z}_i$ to other spots with the same label, and b_i^{sil} denote its mean distance to the nearest different cluster. The standard silhouette coefficient is $s_i = (b_i^{\text{sil}} - a_i)/\max\{a_i, b_i^{\text{sil}}\}$. Silhouette label evaluates the overall separation of biological annotations, rescaled to $[0, 1]$:

$$\text{Silhouette label} = \frac{1}{N}\sum_{i=1}^{N}\frac{s_i^{\text{label}} + 1}{2}. \tag{20}$$

cLISI. For spot i, let p_{ic} denote the local proportion of category c in its embedding neighborhood. The base local inverse Simpson index is $\text{LISI}_i = (\sum_{c=1}^{C} p_{ic}^2)^{-1}$. When the categories correspond to biological labels (C being the number of biological labels), the scaled cell-type LISI score is:

$$\text{cLISI} = \frac{C - \text{median}_i(\text{LISI}_i^{\text{label}})}{C - 1}. \tag{21}$$

Silhouette batch. Silhouette batch measures slice mixing within each biological label. For each spot, a silhouette coefficient is computed using slice identity while restricting the comparison to spots with the same biological label. The score is:

$$\text{Silhouette batch} = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{N_c}\sum_{i:y_i=c}\left(1 - \left|s_i^{\text{slice}}\right|\right). \tag{22}$$

iLISI. Following the LISI definition, when the categories correspond to slices (B being the number of slices), the scaled integration LISI score measures local batch mixing:

$$\text{iLISI} = \frac{\text{median}_i(\text{LISI}_i^{\text{slice}}) - 1}{B - 1}. \tag{23}$$

KBET. Within each biological label, KBET compares the observed slice composition of each neighborhood with the overall slice composition of that label via a χ^2 test. With significance level $\alpha = 0.05$, the acceptance rate A_c for label c is computed, and the final score is averaged over valid labels:

$$\text{KBET} = \frac{1}{|\mathcal{C}_{\text{valid}}|}\sum_{c\in\mathcal{C}_{\text{valid}}} A_c. \tag{24}$$

Graph connectivity. For biological label c, let G_c denote the subgraph induced by spots with that label, and let $\text{LCC}(G_c)$ denote its largest connected component. Graph connectivity is:

$$\text{Graph connectivity} = \frac{1}{C}\sum_{c=1}^{C}\frac{|\,\text{LCC}(G_c)|}{|G_c|}. \tag{25}$$

Aggregate scores. The Bio conservation score is the arithmetic mean of its five component metrics. We split the formula to fit margin constraints:

$$\begin{aligned}\text{Bio conservation} = \frac{1}{5}\big(&\text{Isolated labels} + \text{KMeans NMI}\\ &+ \text{KMeans ARI} + \text{cLISI}\\ &+ \text{Silhouette label}\big).\end{aligned} \tag{26}$$

The Batch correction score is similarly averaged over its four metrics:

$$\begin{aligned}\text{Batch correction} = \frac{1}{4}\big(&\text{Silhouette batch} + \text{iLISI}\\ &+ \text{KBET} + \text{Graph connectivity}\big).\end{aligned} \tag{27}$$

The overall score is computed as:

$$\text{Total} = 0.6\,\text{Bio conservation} + 0.4\,\text{Batch correction}. \tag{28}$$

C.2 METRICS FOR UNANNOTATED DATASETS

For the unannotated MT and MPB datasets, each learned embedding is first partitioned using KMeans with a dataset-specific fixed number of clusters. The resulting assignments $\widehat{y}_i \in \{1, \ldots, K\}$ are used as inferred spatial-domain labels. Metrics are presented in the order of the corresponding benchmark tables.

cLISI. The cLISI score follows the definition established above, but uses the KMeans assignments $\widehat{y}_i$ in place of manual biological annotations to evaluate the local purity of unsupervised domains. The neighborhood size is 90.

ASW cluster. The silhouette coefficient is computed using the inferred KMeans labels and Euclidean distances in the learned embedding. It is rescaled to $[0, 1]$ as:

$$\text{ASW cluster} = \frac{1}{N} \sum_{i=1}^{N} \frac{s_i^{\text{cluster}} + 1}{2}. \tag{29}$$

Moran's I. Within each slice, a spatial K-nearest-neighbor graph is constructed from the physical coordinates. For inferred domain c, define the binary indicator $x_{ic} = \mathbb{I}(\widehat{y}_i = c)$. Moran's I for domain c in slice b is computed as follows (where $\bar{x}_{bc}$ is the mean indicator and w_{ij} is the spatial adjacency):

$$I_{b,c} = \frac{N_b}{W_b} \frac{\sum_{i=1}^{N_b} \sum_{j=1}^{N_b} w_{ij}(x_{ic} - \bar{x}_{bc})(x_{jc} - \bar{x}_{bc})}{\sum_{i=1}^{N_b}(x_{ic} - \bar{x}_{bc})^2}. \tag{30}$$

The score is averaged across inferred domains and then across slices:

$$\text{Moran's I} = \frac{1}{B} \sum_{b=1}^{B} \left(\frac{1}{K} \sum_{c=1}^{K} I_{b,c} \right). \tag{31}$$

ASW slice. For each spot i, let s_i^{slice} denote the silhouette coefficient computed using slice identities as labels. Because a larger silhouette coefficient indicates stronger slice separation, we reverse and rescale it to obtain a higher-is-better slice-mixing score:

$$\text{ASW slice} = \frac{1}{N} \sum_{i=1}^{N} \frac{1 - s_i^{\text{slice}}}{2}. \tag{32}$$

A larger value therefore indicates weaker slice-associated separation and better cross-slice mixing.

iLISI. Similar to cLISI, the iLISI metric evaluates local batch mixing based on the unannotated embeddings, maintaining the exact formulation defined in the annotated metric section.

Aggregate scores. The Domain Conservation score evaluates spatial coherence and cluster purity:

$$\begin{aligned} \text{Domain Conservation} = \frac{1}{3}\big(&\text{cLISI} + \text{ASW cluster} \\ &+ \text{Moran's I}\big). \end{aligned} \tag{33}$$

The Batch correction score evaluates cross-slice integration:

$$\text{Batch correction} = \frac{1}{2}\left(\text{iLISI} + \text{ASW slice}\right). \tag{34}$$

The overall score balances both aspects:

$$\text{Total} = 0.6\,\text{Domain Conservation} + 0.4\,\text{Batch correction}. \tag{35}$$

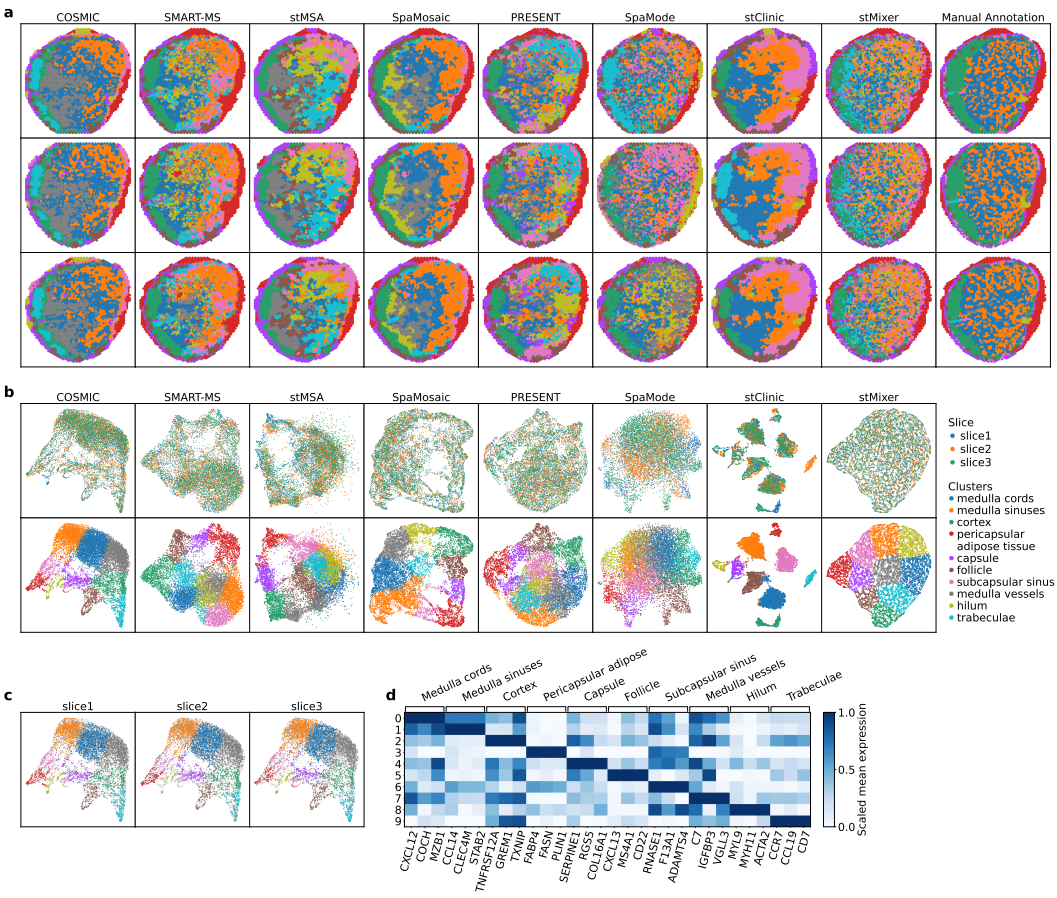

Figure 4: Qualitative comparison on the HLN dataset. (a) Manual annotations and spatial-domain identification results across three human lymph node slices using baseline methods and COSMIC. (b) UMAP visualizations of the integrated representations produced by COSMIC and baseline methods, colored by slice identity and inferred domain labels, respectively. (c) COSMIC-derived spatial domains visualized separately in each slice. (d) Cluster-wise RNA marker-expression heatmap supporting the biological identities of the COSMIC-derived domains.

D Additional Qualitative Results on HLN

Figure 4 presents additional qualitative evaluations on the highly heterogeneous human lymph node (HLN) dataset. As shown in Figure 4a, COSMIC successfully recovers complex spatial domains with high spatial coherence, whereas baseline methods often produce noisy or fragmented boundaries. UMAP visualizations (Figure 4b,c) further demonstrate that COSMIC effectively mitigates slice-associated variations, achieving broad integration across all three slices while maintaining distinguishable biological clusters. Finally, the biological validity of these inferred domains is confirmed by distinct molecular signatures in the cluster-wise RNA marker-expression heatmap (Figure 4d), demonstrating that the COSMIC framework preserves true biological variation while eliminating technical artifacts.

E Complete Ablation Results

To supplement the summarized ablation results presented in the main text, Tables 4 and 5 detail the comprehensive performance of COSMIC and its ablated variants on the HT and MPB datasets, respectively. While the main manuscript focuses on the aggregated macro-metrics, these tables provide the full breakdown of all fine-grained evaluation metrics. This includes specific measurements for clustering accuracy , local and global batch mixing , and spatial coherence . The detailed metric

breakdown further demonstrates that the complete COSMIC framework—integrating key components such as Target-Conditioned Cross-Modal Interaction (TCCI) and Cluster-Guided Conditional Alignment (CGCA)—achieves the most robust and balanced performance across all dimensions of multi-slice spatial multi-omics integration.

Table 4: Ablation study of COSMIC on the HT dataset. Bold indicates the best result, and underline denotes the second-best.

Metrics	Full	w/o TCCI	w/o CGCA
	HT dataset		
Isolated labels	<u>0.565</u>	**0.565**	0.562
KMeans NMI	**0.278**	0.235	<u>0.268</u>
KMeans ARI	**0.177**	0.125	<u>0.164</u>
Silhouette label	**0.533**	0.524	<u>0.529</u>
cLISI	**0.870**	0.868	**0.870**
Silhouette batch	**0.949**	<u>0.946</u>	0.945
iLISI	**0.711**	<u>0.679</u>	0.471
KBET	**0.700**	<u>0.695</u>	0.459
Graph connectivity	**0.912**	0.901	<u>0.909</u>
Bio conservation	**0.485**	0.464	<u>0.479</u>
Batch correction	**0.818**	<u>0.805</u>	0.696
Total	**0.618**	<u>0.600</u>	0.566

Table 5: Ablation study of COSMIC on the MPB dataset. Bold indicates the best result, and underline denotes the second-best.

Metrics	COSMIC	w/o TCCI	w/o CGCA
	MPB dataset		
cLISI	<u>0.981</u>	0.979	**0.981**
ASW cluster	0.582	**0.587**	<u>0.584</u>
Moran's I	**0.575**	0.480	<u>0.569</u>
ASW slice	0.496	**0.507**	<u>0.498</u>
iLISI	**0.509**	<u>0.493</u>	0.440
Domain Conservation	**0.713**	0.682	<u>0.711</u>
Batch correction	**0.503**	<u>0.500</u>	0.469
Total	**0.628**	0.609	<u>0.614</u>

F Complete Parameter Sensitivity Results

To supplement the summarized parameter sensitivity analysis presented in the main text, Tables **??** and **??** detail the performance of COSMIC under varying numbers of auxiliary semantic anchors K on the HT and MPB datasets, respectively. While the main manuscript focuses on the aggregated macro-metrics, these tables provide a complete breakdown of the fine-grained evaluation metrics. Here, K determines the dimensionality of the soft assignment matrix produced by the clustering projection head and, consequently, the number of soft prototypes used to construct confidence-filtered slice-specific and global semantic centers in the CGCA module. It does not represent the spatial neighborhood size, the number of clusters used for downstream evaluation, or the number of ground-truth biological domains. The results remain stable when K ranges from 4 to 8 on HT and from 8 to 16 on MPB, demonstrating that COSMIC is robust to moderate variations in the number of auxiliary semantic anchors and does not require K to exactly match the true number of biological domains.