CONTEXT-AWARE SYNCHRONIZATION FOR SPATIAL MULTI-OMICS REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial multi-omics domain identification relies on unified representations that preserve multimodal molecular information while organizing domain-resolvable latent geometry. Existing geometry organization mechanisms mainly rely on explicit inter-location relations or shared semantic references, leaving location-specific multimodal and multiscale context without a direct role in geometry optimization. To address this gap, Context-aware Synchronization for Spatial Multi-omics Representation Learning (CSR) formulates latent geometry optimization as a context-regulated synchronization process on the hypersphere. CSR constructs location-specific contextual references from multimodal representations and multiscale graph contexts, and combines them with relational graph coupling to regulate representation evolution. The continuous synchronization dynamics monotonically decrease the associated objective, while contextual references impose location-dependent stationary conditions. Experiments on spatial multi-omics benchmarks demonstrate consistent improvements in spatial domain identification over representative baselines.

1 INTRODUCTION

Recent advances in spatial multi-omics technologies have enabled the simultaneous measurement of multiple molecular modalities within the same tissue context, providing complementary views of cellular states and tissue organization (Longo et al., 2021). By associating heterogeneous molecular measurements with spatial locations, spatial multi-omics data enable the characterization of cellular heterogeneity and spatially organized tissue structures (Dries et al., 2021). However, learning unified representations from spatial multi-omics data remains challenging because different modalities exhibit distinct measurement mechanisms, feature spaces, and noise characteristics (Geng et al., 2026). Therefore, a central computational challenge is to learn unified representations that preserve multimodal molecular information while organizing biologically meaningful latent representations.

Existing spatial multi-omics representation learning approaches can be viewed from two complementary perspectives: preserving multimodal molecular information and organizing the latent representation space. Many existing methods preserve multimodal information by learning modality-specific representations and integrating heterogeneous modalities through fusion, alignment, or reconstruction objectives (Long et al., 2024; Du et al., 2026; Huang et al., 2025). Meanwhile, the organization of latent geometry is often guided by relational constraints derived from spatial proximity, molecular similarity, or semantic references. For example, graph-based methods incorporate spatial relations into representation learning through graph construction and relational aggregation (Hu et al., 2021; Dong & Zhang, 2022; Long et al., 2023). In contrast, metric learning, contrastive learning, and prototype-based methods introduce explicit similarity or semantic constraints to organize the latent space (Du et al., 2026; Long et al., 2023; Huang et al., 2025; Zhang et al., 2026). From this perspective, existing methods preserve molecular information while organizing latent representations through different forms of relational or semantic guidance.

Despite these advances, existing geometry organization strategies primarily rely on explicitly constructed relations or references, such as spatial graphs, molecular similarities, or semantic prototypes. These mechanisms provide effective constraints for shaping the latent space, but they mainly characterize relations among locations or predefined references rather than directly modeling the intrinsic multimodal and multiscale context of each location. As a result, how to enable the continu-

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.

ous context of each spatial location to directly guide representation organization while maintaining relational consistency remains an open question.

To address this question, Context-aware Synchronization for Spatial Multi-omics Representation Learning (CSR) formulates representation learning as a context-regulated synchronization process. In CSR, location-specific contextual information directly participates in representation evolution. Specifically, CSR optimizes hyperspherical representations under graph-based relational constraints and location-specific contextual guidance. A location-specific contextual reference is constructed by integrating multimodal representations with multiscale graph contexts, providing location-dependent guidance throughout synchronization. The resulting synchronization dynamics are theoretically characterized through energy descent and context-regulated stationary states.

The contributions of this paper are summarized as follows:

- We formulate spatial multi-omics representation learning as a context-regulated synchronization problem on a hyperspherical representation space.
- We introduce a location-specific contextual reference that integrates multimodal representations with multiscale graph contexts to provide continuous guidance for representation organization.
- We provide theoretical analysis of the synchronization dynamics, establishing energy descent and characterizing context-regulated stationary states.

2 Related Work

Spatial Multi-omics Representation Learning. Spatial multi-omics methods integrate heterogeneous molecular measurements and spatial information to learn representations for spatial domain identification. We review existing mechanisms for organizing latent representations from two non-mutually exclusive perspectives: explicit geometry-shaping objectives and relation-conditioned representation learning.

Explicit Geometry-Shaping Objectives. Explicit geometric constraints guide the organization of latent representations through similarity- or cluster-based objectives. MISO (Coleman et al., 2025) employs neural spectral clustering to encourage similarity-based cluster structure, while SMART (Du et al., 2026) uses triplet-based metric learning to distinguish feature-similar and dissimilar spots. SpaMosaic (Yan et al., 2026) aligns representations of the same spot across modalities through contrastive learning while distinguishing different spots. PRAGA (Huang et al., 2025) introduces dynamic prototype contrastive learning to guide cluster formation, while ProtoPSR (Zhang et al., 2026) builds on PRAGA by further incorporating pairwise similarity regularization to enforce relational consistency.

Relation-Conditioned Representation Learning. Another line of work incorporates spatial and molecular relationships into the representation learning process. CellCharter (Varrone et al., 2024) incorporates multiscale spatial context through neighborhood feature aggregation, while SpatialGlue (Long et al., 2024) separately encodes spatial and feature graphs and adaptively fuses their representations through attention. COSMOS (Zhou et al., 2025) further incorporates spatial relationships through graph convolution and spatial regularization. Moving beyond fixed neighborhood structures, PRAGA (Huang et al., 2025) and ProtoPSR (Zhang et al., 2026) learn adaptive modality-specific graphs to capture latent molecular relationships.

Summary. Existing methods organize latent representations through explicit geometric constraints, relation-conditioned learning, or their combination. While these methods utilize molecular measurements from individual spatial locations, representation organization is largely guided by similarity relationships, neighborhood interactions, or shared cluster-level references. Consequently, the role of continuous, location-specific context in directly shaping latent representations remains less explicitly explored. This motivates exploring how continuous, location-specific context can actively shape latent representations.

3 Preliminaries

Spatial multi-omics and graph notation. Consider M omics modalities measured at N matched spatial locations. For the m-th modality, the feature matrix is denoted by $\mathbf{X}^{(m)} \in \mathbb{R}^{N \times D_m}$, where

D_m represents the number of retained omics features. The spatial coordinates are collected in $\mathbf{S} \in \mathbb{R}^{N\times 2}$, with $\mathbf{s}_i$ standing for the coordinate of the i-th location. When reference domain labels are available, $\mathbf{y} = [y_1, \ldots, y_N]^\top$ denotes the corresponding label vector, where y_i is the domain label of the i-th location. In the experimental setting of this paper, labels are used only for evaluation.

To characterize relations among the spatial locations, a graph is represented by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ contains the N locations, namely $N = |\mathcal{V}|$. A symmetric nonnegative relation matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$ encodes the relations between locations. The corresponding degree matrix and the graph Laplacian matrix are defined as $\mathbf{D} = \mathrm{diag}(\mathbf{A}\mathbf{1})$ and $\mathbf{L} = \mathbf{D} - \mathbf{A}$, respectively.

Graph smoothness. A smooth graph representation varies slowly across strongly related locations. For the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with relation matrix $\mathbf{A}$ and Laplacian $\mathbf{L}$ defined above, the smoothness of node representations $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_N]^\top \in \mathbb{R}^{N\times d}$ can be quantified by the Dirichlet energy Shuman et al. (2013), namely,

$$\mathcal{E}_{\mathrm{G}}(\mathbf{Z}; \mathbf{A}) = \mathrm{Tr}\left(\mathbf{Z}^\top \mathbf{L}\mathbf{Z}\right) = \frac{1}{2}\sum_{i,j} A_{ij}\|\mathbf{z}_i - \mathbf{z}_j\|_2^2. \tag{1}$$

Minimizing Eq. 1 penalizes representation differences between strongly related locations and therefore promotes graph smoothness. This formulation is employed to characterize the graph-dependent component of the inner representation objective.

Hyperspherical geometry. A unit hypersphere provides a fixed-norm state space for vector representations. The $(d-1)$-dimensional unit hypersphere is defined as $\mathbb{S}^{d-1} = \{\mathbf{z} \in \mathbb{R}^d : \|\mathbf{z}\|_2 = 1\}$. Unlike unconstrained Euclidean space, feasible local variations on the hypersphere must lie in the tangent space Boumal (2020). For $\mathbf{z} \in \mathbb{S}^{d-1}$, the tangent space is $T_{\mathbf{z}}\mathbb{S}^{d-1} = \{\mathbf{v} \in \mathbb{R}^d : \mathbf{z}^\top \mathbf{v} = 0\}$.

An arbitrary direction $\mathbf{v} \in \mathbb{R}^d$ can be mapped onto the tangent space via the orthogonal projection

$$\Pi_{\mathbf{z}}(\mathbf{v}) = \mathbf{v} - (\mathbf{z}^\top \mathbf{v})\mathbf{z}. \tag{2}$$

For a smooth scalar function F with Euclidean gradient $\nabla_{\mathbf{z}} F$ in the ambient space, its Riemannian gradient on the hypersphere is $\mathrm{grad}_{\mathbf{z}} F = \Pi_{\mathbf{z}}(\nabla_{\mathbf{z}} F)$. For N representation states, the joint feasible space is the product manifold $\mathcal{M} = (\mathbb{S}^{d-1})^N$. These geometric quantities provide the basis for formulating and analyzing the synchronization dynamics Miyato et al. (2025).

4 Analysis and Motivation

As discussed in the Introduction, representative spatial multi-omics methods shape latent geometry through relations among representations or shared semantic references. These mechanisms include spectral, triplet, contrastive, prototype-based, and pseudo-pairwise regularization, which are formalized here under the common view of *relation-guided embedding regularization*.

Relation-guided embedding regularization. The defining property is that the geometric constraint is specified through explicit relations among latent entities. Let $\mathcal{U} = \{\mathbf{u}_a\}_{a\in\mathcal{I}}$ stand for the latent entities involved in an explicit geometry-shaping term, and let $\mathcal{R}$ denote a collection of relations, where each $r \in \mathcal{R}$ specifies a set of related entity pairs (a, b). Under this formulation, the geometry-shaping term is defined as

$$\mathcal{L}_{\mathrm{rel}}(\mathcal{U}; \mathcal{R}) = \sum_{r\in\mathcal{R}} \ell_r\left(\{\kappa(\mathbf{u}_a, \mathbf{u}_b)\}_{(a,b)\in r}\right), \tag{3}$$

where $\kappa(\cdot,\cdot)$ represents a pairwise similarity or distance measure, and $\ell_r(\cdot)$ specifies the geometry constraint associated with relation r. This formulation isolates explicit geometry-shaping terms from general representation learning components such as encoding and reconstruction. Representative instances are provided in App. A.

The relational formulation in Eq. 3 induces the following constrained gradient structure.

Proposition 1 (Gradient constraint of relation-guided objectives)**.** *Assume that the pairwise function κ in Eq. 3 is differentiable and that its gradient with respect to each latent entity lies in the span of the involved entities. Then*

$$\nabla_{\mathbf{u}_i}\mathcal{L}_{\mathrm{rel}} \in \mathrm{span}\left(\{\mathbf{u}_i\} \cup \{\mathbf{u}_j : j \in \mathcal{N}_{\mathcal{R}}(i)\}\right), \tag{4}$$

where $\mathcal{N}_{\mathcal{R}}(i)$ denotes the latent entities explicitly related to $\mathbf{u}_i$.

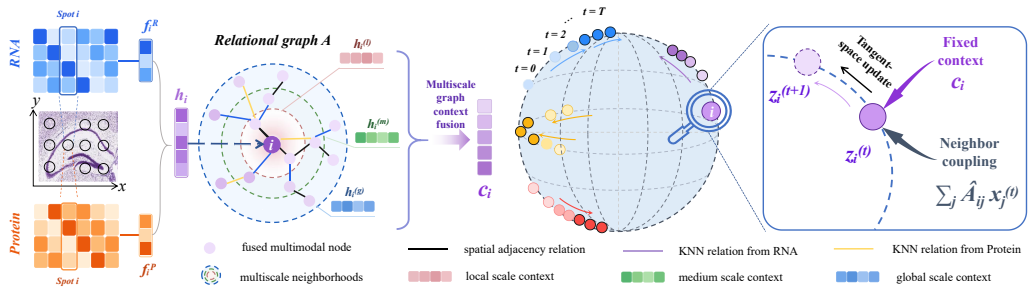

Figure 1: Illustration of CSR. The framework constructs multimodal representations and location-specific contextual references, followed by context-regulated hyperspherical synchronization.

Proposition 1 indicates that relation-guided objectives update latent representations through the entities explicitly involved in the specified relations. Hence, location-specific contextual information, although potentially encoded in latent representations, is not explicitly introduced as an independent component of the geometry optimization objective.

Location-specific contextual regularization. To introduce an explicit location-conditioned component into geometry optimization, each spatial location i is associated with a continuous contextual reference $\mathbf{c}_i \in \mathbb{R}^d$. The contextual reference is constructed from the multimodal representation and multiscale graph context of the corresponding location, providing a continuous location-specific signal rather than a shared semantic reference. Let $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_N]^\top$. During inner optimization, $\mathbf{C}$ is fixed with respect to the representation states.

Given hyperspherical states $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_N]^\top$, where $\mathbf{z}_i \in \mathbb{S}^{d-1}$, the inner representation objective combines graph smoothness and contextual regularization:

$$\begin{aligned} \mathcal{E}(\mathbf{Z}; \mathbf{A}, \mathbf{C}) &= \beta \mathcal{E}_{\mathrm{G}}(\mathbf{Z}; \mathbf{A}) - \alpha \sum_i \mathbf{c}_i^\top \mathbf{z}_i \\ &= \frac{\beta}{2} \sum_{i,j} A_{ij} \|\mathbf{z}_i - \mathbf{z}_j\|_2^2 - \alpha \sum_i \mathbf{c}_i^\top \mathbf{z}_i. \end{aligned} \tag{5}$$

The first term enforces graph smoothness among related locations, while the second term introduces an explicit location-specific optimization direction from contextual information. The contextual term extends relation-guided geometry organization by incorporating location-conditioned information into the representation optimization process. The resulting objective jointly models relational consistency and location-specific contextual information during representation optimization.

5 Context-aware Synchronization Representation Learning

Context-aware Synchronization Representation Learning (CSR) is a framework for learning hyperspherical representations under graph-based relational constraints and location-specific contextual information. CSR instantiates the objective in Eq. 5 by constructing the relational graph and contextual reference and optimizing the representation states through synchronization dynamics.

Overview. The overall framework is illustrated in Fig. 1. CSR consists of four main components. (1) Multimodal representation and relation graph construction obtain molecular representations and spatial relations. (2) Location-specific contextual reference construction generates contextual references from multimodal representations and multiscale graph responses. (3) Context-regulated hyperspherical synchronization updates representations under relational and contextual constraints. (4) Reconstruction-based joint optimization preserves modality-specific information by jointly optimizing the synchronization module and modality-specific decoders.

5.1 Multimodal Representation and Relation Graph Construction

This component learns multimodal representations from molecular measurements and constructs the relational graph used to define interactions among spatial locations.

Multimodal representation. For multiple molecular modalities, modality-specific encoders and a projection layer are applied to obtain the shared representation:

$$\mathbf{H} = f^{\mathrm{proj}}\left(f^{(1)}(\mathbf{X}^{(1)}) \| \cdots \| f^{(M)}(\mathbf{X}^{(M)})\right) \in \mathbb{R}^{N\times d}, \tag{6}$$

where $f^{(m)}(\cdot)$ denotes the modality-specific encoder for the m-th modality, $f^{\mathrm{proj}}(\cdot)$ denotes the learnable projection layer, and $\|$ indicates feature concatenation.

Relation graph construction. To define the interaction constraints in synchronization optimization, a relational graph is constructed by integrating spatial proximity and molecular similarity. Specifically, a spatial k-nearest-neighbor graph is obtained from the spatial coordinates, denoted as $\mathbf{A}^{\mathrm{spa}}$. For each modality, a molecular relation graph $\mathbf{A}^{\mathrm{feat},(m)}$ is constructed according to molecular profile correlation similarity. These complementary relations are combined as

$$\mathbf{A}^{\mathrm{raw}} = \mathbf{A}^{\mathrm{spa}} \vee \bigvee_{m=1}^{M} \mathbf{A}^{\mathrm{feat},(m)}, \tag{7}$$

where $\vee$ represents the union of binary relations. The resulting relation graph is further symmetrized and normalized to obtain the interaction matrix $\mathbf{A}$ used in synchronization optimization. The matrix $\mathbf{A}$ specifies which spatial locations participate in relational interactions, while the representation geometry is determined by the subsequent optimization dynamics.

5.2 Location-specific Contextual Reference Construction

The relational graph specifies interactions among spatial locations but does not provide a location-specific reference for representation optimization. To construct such a reference, multiscale graph context is extracted from multimodal representations through a graph scattering transform (GST) (Gao et al., 2019).

The GST captures multiscale graph context through diffusion-based wavelet operators. Given the fused representation $\mathbf{H}$ and relation matrix $\mathbf{A}$, the wavelet response at scale h is computed as

$$\mathbf{G}_h = \mathbf{P}^{2^H} \left|\mathbf{\Psi}_h \mathbf{H}\right|, \quad \mathbf{\Psi}_h = \mathbf{P}^{2^{h-1}} - \mathbf{P}^{2^h}, \tag{8}$$

where $\mathbf{P} = \frac{1}{2}(\mathbf{I} + \mathbf{D}^{-1}\mathbf{A})$ denotes the diffusion operator and $|\cdot|$ represents the modulus operation.

The contextual reference matrix is constructed by combining the multimodal representation with multiscale graph scattering responses:

$$\mathbf{C} = \mathrm{MLP}_{\mathrm{ctx}}\left(\mathbf{H} \| [\mathbf{G}_0 \| \cdots \| \mathbf{G}_H]\right) \in \mathbb{R}^{N\times d}, \tag{9}$$

where $\mathrm{MLP}_{\mathrm{ctx}}(\cdot)$ maps the concatenated features into the contextual reference space.

5.3 Context-Regulated Hyperspherical Synchronization

Given the relational matrix $\mathbf{A}$ and contextual reference matrix $\mathbf{C}$, CSR updates hyperspherical states by following the Riemannian gradient flow of the contextual objective in Eq. 5. After projection onto the tangent space of the hypersphere, the synchronization dynamics are defined as

$$\dot{\mathbf{z}}_i = \Pi_{\mathbf{z}_i}\left(\beta \sum_j A_{ij}\mathbf{z}_j + \alpha \mathbf{c}_i\right). \tag{10}$$

The graph and contextual terms respectively provide relational coupling and location-specific optimization directions.

For numerical optimization, the continuous dynamics are discretized using a projected Euler update:

$$\mathbf{z}_i^{(t+1)} = \mathrm{Norm}\left[\mathbf{z}_i^{(t)} + \eta \Pi_{\mathbf{z}_i^{(t)}}\left(\beta \sum_j A_{ij}\mathbf{z}_j^{(t)} + \alpha \mathbf{c}_i\right)\right], \tag{11}$$

where η is the integration step size. The normalization operation projects the updated states back onto $\mathbb{S}^{d-1}$.

The initial state is constructed as

$$\mathbf{Z}^{(0)} = \mathrm{RowNorm}(\mathbf{H} + \mathbf{C} + \epsilon), \quad \epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I}), \tag{12}$$

where the perturbation is sampled only at initialization. During each synchronization trajectory, $\mathbf{C}$ remains fixed while the representation states evolve according to Eq. 11.After T synchronization steps, the final representation is obtained as

$$\mathbf{Z}^{\star} = \mathbf{Z}^{(T)}. \tag{13}$$

where the perturbation is sampled only at initialization. The contextual reference remains fixed during each synchronization trajectory, while the representation states evolve according to Eq. 11.

5.3.1 Reconstruction-Based Joint Optimization

The synchronized representation is decoded into each molecular modality using modality-specific GCN decoders:

$$\widehat{\mathbf{X}}^{(m)} = g^{(m)}(\mathbf{Z}^{\star}, \mathbf{A}^{\mathrm{spa}}), \tag{14}$$

where $g^{(m)}(\cdot)$ denotes the GCN decoder for the m-th modality.

The reconstruction objective is defined as

$$\mathcal{L}_{\mathrm{rec}} = \sum_{m=1}^{M} \frac{1}{N} \left\| \mathbf{X}^{(m)} - \widehat{\mathbf{X}}^{(m)} \right\|_F^2. \tag{15}$$

The reconstruction objective preserves modality-specific information, while the representation geometry is determined by the context-regulated synchronization dynamics.

6 Theoretical Analysis

The theoretical properties of CSR are characterized from two aspects. First, the monotonicity of the continuous synchronization dynamics is established. Second, the effect of location-specific contextual references on synchronization states is characterized.

6.1 Energy Descent of Synchronization Dynamics

The synchronization dynamics induce a continuous optimization trajectory on the hyperspherical manifold. The following theorem characterizes the energy behavior along this trajectory.

Theorem 1 (Energy Descent)**.** *Assume that $\mathbf{A} = \mathbf{A}^{\top}$, $A_{ij} \geq 0$, $\beta > 0$, and $\alpha \geq 0$, with $\mathbf{C}$ fixed during synchronization. For any initial state $\mathbf{Z}^{(0)} \in (\mathbb{S}^{d-1})^N$, the continuous trajectory induced by Eq. 10 satisfies*

$$\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}t} = -\sum_{i=1}^{N} \left\| \mathrm{grad}_{\mathbf{z}_i} \mathcal{E} \right\|_2^2 \leq 0. \tag{16}$$

Theorem 1 guarantees that CSR performs objective-guided synchronization rather than heuristic state evolution. Since the objective is bounded below on the hyperspherical state space, the continuous trajectory converges to a finite energy value. The discrete update in Eq. 11 is a numerical approximation of this flow, and its monotonic decrease is not guaranteed for arbitrary step sizes. The proof is provided in App. C.1.

6.2 Effect of Location-Specific Context

The contextual reference introduces location-dependent constraints on the stationary states of synchronization. In particular, the condition for a fully synchronized representation to remain stationary depends on the contextual references.

Proposition 2 (Context-regulated synchronization condition)**.** *Assume $\alpha > 0$ and consider a fully synchronized state*

$$\mathbf{z}_1 = \cdots = \mathbf{z}_N = \mathbf{z}, \quad \mathbf{z} \in \mathbb{S}^{d-1}. \tag{17}$$

Table 1: Clustering results on the RNA and ADT datasets.

Methods	ARI	AMI	NMI	FMI	HOMO	COMPL	F1	Jaccard	sLISI
Human Tonsil Section 1									
COSMOS	10.16	18.06	18.14	43.61	19.76	16.76	43.25	27.59	84.31
SpatialGlue	17.69	28.50	28.57	48.33	30.72	26.70	48.51	32.02	67.83
ProtoPSR	24.58	29.17	29.24	55.08	30.32	28.23	55.03	37.96	69.23
SMART	36.46	35.86	35.92	61.87	**37.55**	34.42	61.77	44.69	71.09
Ours	**41.86**	**38.18**	**38.25**	**67.61**	37.47	**39.05**	**67.54**	**50.99**	73.33
Human Lymph Node A1									
COSMOS	18.65	33.15	33.58	34.50	37.48	30.42	33.97	20.46	93.06
SpatialGlue	27.25	39.03	39.43	41.61	43.50	36.05	40.99	25.78	83.94
ProtoPSR	30.08	40.86	41.26	44.80	43.97	38.87	44.63	28.72	88.49
SMART	35.23	42.79	43.18	48.78	**46.51**	40.29	48.54	32.05	87.75
Ours	**38.13**	**43.63**	**44.03**	**53.49**	44.28	**43.78**	**53.33**	**36.36**	88.99

This state is an equilibrium of Eq. 10 if and only if

$$\Pi_{\mathbf{z}}(\mathbf{c}_i) = 0, \quad i = 1, \ldots, N. \tag{18}$$

Equivalently, each contextual reference aligns with the synchronized direction, i.e., $\mathbf{c}_i \in \mathrm{span}\{\mathbf{z}\}$.

Proposition 2 demonstrates that location-specific context changes the stationary structure of graph synchronization. When representation states become identical, the graph coupling term vanishes after tangent projection and cannot provide further optimization directions. In contrast, the contextual reference introduces location-dependent directions that constrain the stationary states. Therefore, CSR does not merely smooth representations over relational graphs, but jointly optimizes relational consistency and contextual information. The proof is offered in App. C.2.

7 Experiment

Details of the datasets, preprocessing, baselines, evaluation metrics, and implementation are provided in Appendix C.3.

Overall Performance on Spatial Domain Identification. Tables 1–3 summarize the results across RNA+ADT, RNA+ATAC, and RNA+ATAC+ADT integration scenarios.Complete results for all baselines and metrics are provided in Appendix D. All metrics are reported on a 0–100 scale, with higher values indicating better performance; bold and underlined entries denote the best and second-best results. CSR consistently achieves strong performance across the three integration scenarios. On RNA+ADT datasets, it attains the best results on most reference-based metrics; for Human Lymph Node A1, ARI and FMI improve from 35.23% to 38.13% and from 48.78% to 53.49%, respectively. For RNA+ATAC integration, CSR yields substantial gains on both mouse brain datasets, including an ARI of 70.94% on Mouse Embryonic Brain E15.5. On the simulated tri-omics dataset, it achieves near-perfect performance with an ARI of 99.78% and a Jaccard score of 99.66%. Although CSR is slightly less competitive on sLISI, its stronger reference-based performance indicates that higher local spatial coherence does not necessarily correspond to more accurate recovery of annotated spatialdomains.

Qualitative Analysis of Spatial Representations. Figure 2 juxtaposes the spatial domains identified by CSR, SMART, SpatialGlue, and COSMOS with the reference annotations for Human Lymph Node A1. Reference cluster labels were obtained using the single-cell dataset GSE1441367 (Li et al., 2023). These labels reveal a heterogeneous tissue architecture. CSR closely matches the reference annotations by recovering the dominant cortical region, the interspersed medullary cords and sinuses, and key peripheral structures. In comparison, SMART shows deviations in several regions, SpatialGlue produces more fragmented medullary assignments, and COSMOS yields overly uniform spatial domains.

The analysis of Mouse Embryonic Brain E15.5 further combines spatial domain assignments with UMAP (Healy & McInnes, 2024) projections of the learned representations (Fig. 3). CSR most faithfully recovers the spatial organization of the reference domains (Fig. 3(b)). In particular, it

Table 2: Clustering results on the RNA and ATAC datasets.

Methods	ARI	AMI	NMI	FMI	HOMO	COMPL	F1	Jaccard	sLISI
Mouse Embryonic Brain E15.5									
COSMOS	33.84	49.32	50.05	43.55	53.37	47.11	42.65	27.10	89.88
SpatialGlue	44.96	58.81	59.41	53.36	62.99	56.21	52.56	35.65	90.58
ProtoPSR	50.76	63.46	63.99	58.42	67.72	60.65	57.60	40.45	92.12
SMART	61.39	66.45	67.03	67.53	70.20	64.14	67.31	50.73	93.33
Ours	**70.94**	**69.12**	**69.59**	**75.75**	**70.67**	**68.54**	**75.74**	**60.96**	92.53
Mouse Brain									
COSMOS	18.81	29.44	29.71	27.17	30.84	28.67	26.94	15.57	96.04
SpatialGlue	30.26	38.15	38.34	40.33	35.58	41.56	39.70	24.77	91.99
ProtoPSR	34.05	39.22	39.47	41.42	40.14	38.81	41.40	26.10	93.33
SMART	38.14	43.93	44.15	45.31	44.47	43.84	45.31	29.29	93.62
Ours	**43.50**	**46.80**	**47.02**	**50.54**	**46.29**	**47.77**	**50.43**	**33.72**	90.12

Table 3: Clustering results on the simulated spatial multi-omics dataset.

Methods	ARI	AMI	NMI	FMI	HOMO	COMPL	F1	Jaccard	sLISI
Simulation									
COSMOS	65.43	74.44	74.54	72.93	75.86	73.27	72.81	57.25	**94.51**
SpatialGlue	97.62	96.79	96.80	98.16	96.90	96.70	98.16	96.38	90.38
ProtoPSR	99.15	98.80	98.81	99.34	98.82	98.80	99.34	98.69	90.38
SMART	97.85	96.96	96.97	98.34	97.00	96.94	98.34	96.73	90.38
Ours	**99.78**	**99.66**	**99.66**	**99.83**	**99.65**	**99.68**	**99.83**	**99.66**	90.38

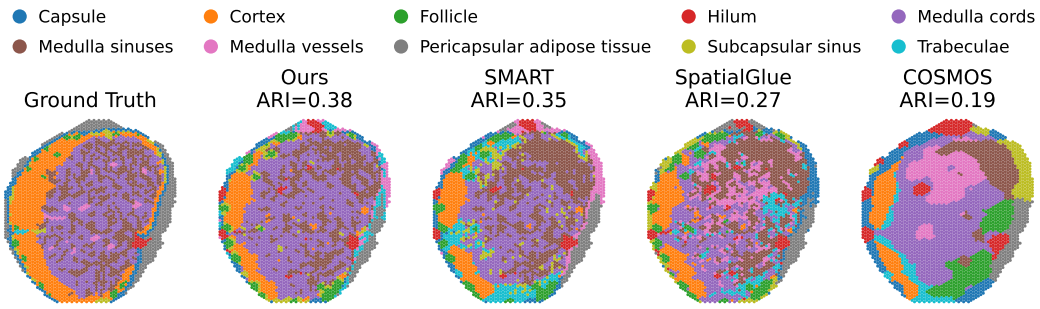

Figure 2: Spatial domain detection results by different methods on Human Lymph Node A1 data.

preserves both the dominant central domain and its rightward extension, while more accurately capturing the curved peripheral structures and the distributions of Domains 5 and 6 than the competing methods. This advantage is further reflected in the UMAP projection (Fig. 3(c)), where CSR yields more clearly separated domain representations with reduced inter-domain overlap.

Effect of Location-Specific Context on Synchronization Dynamics. To examine the effect of contextual regulation on synchronization dynamics, we track the evolution of S_{intra}, S_{inter}, and ΔS (Fig. 4).Without contextual regulation, both $S_{\text{intra}}$and S_{inter} increase, causing ΔS to rise initially but decline with further synchronization. In contrast, CSR progressively increases S_{intra} while reducing S_{inter}, yielding a sustained increase in ΔS.These results indicate that contextual regulation promotes within-domain synchronization while preventing excessive synchronization across spatial domains.

Ablation Studies. Across Human Lymph Node A1, Mouse Brain, and Human Tonsil Section 1, the full CSR model shows the strongest overall clustering performance. Removing contextual guidance, neighbor coupling, GST-based multiscale context, the synchronization module, or iterative updates generally reduces ARI, NMI, and FMI, with the largest degradations often observed after disrupting synchronization or multiscale context modeling. Despite a slight NMI increase on Mouse Brain without context, the overall results demonstrate that contextual guidance, relational coupling, and iterative synchronization contribute complementarily to spatial domain identification.

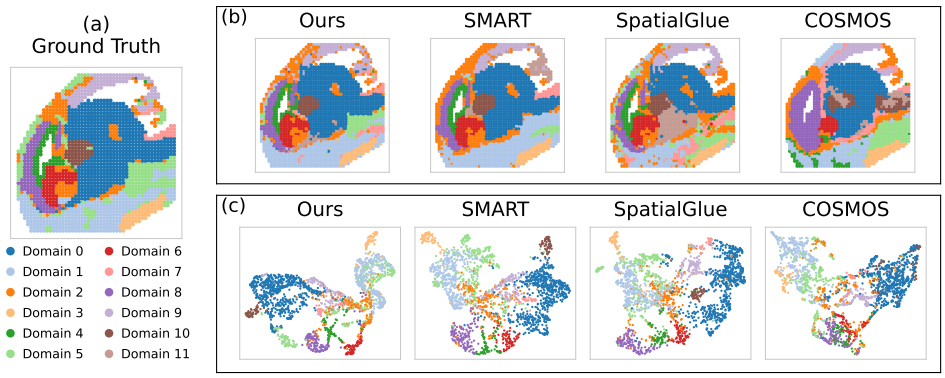

Figure 3: Qualitative results on Mouse Embryonic Brain E15.5. (a) Ground-truth spatial domain annotations. (b) Spatial clustering results and (c) UMAP visualizations of the learned representations obtained by CSR and competing methods.

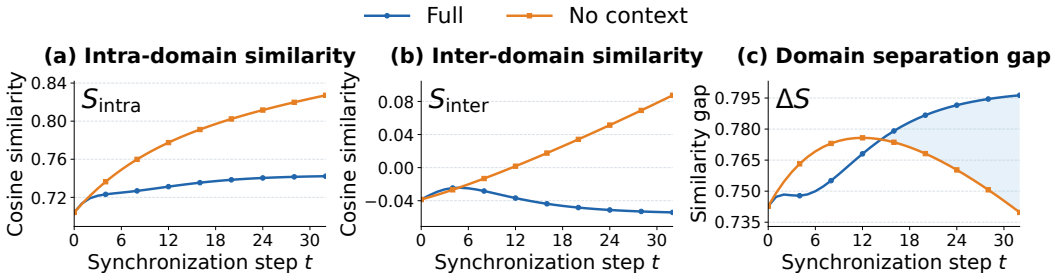

Figure 4: Evolution of intra-domain similarity (S_{intra}), inter-domain similarity (S_{inter}), and the domain separation gap ($\Delta S := S_{\text{intra}} - S_{\text{inter}}$) during synchronization.

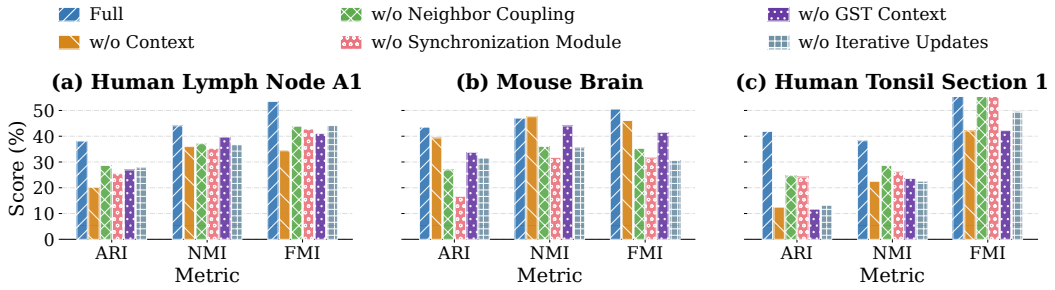

Figure 5: Ablation study of key CSR components on three datasets.

Sensitivity Analysis. Sensitivity to the context strength γ and the number of synchronization steps T is evaluated on three datasets. CSR remains stable across a broad range of settings, with moderate context strengths generally yielding better performance and further synchronization providing only marginal gains beyond a certain point. Additional sensitivity analysis and computational efficiency results are provided in Appendix E.

8 Conclusion

This paper introduced Context-aware Synchronization for Spatial Multi-omics Representation Learning (CSR), which formulates latent geometry optimization as context-regulated synchronization under graph-based relational constraints. CSR constructs location-specific contextual references from multimodal representations and multiscale graph contexts to guide hyperspherical synchronization. Theoretical analysis establishes monotonic descent of the objective along the continuous dynamics and characterizes when fully synchronized states remain stationary under a location-specific context. Empirical evaluations on spatial multi-omics benchmarks support the effectiveness of CSR for spatial domain identification.

AI USE STATEMENT

Generative AI tools were used to assist with methodological and theoretical discussions, experimental design, implementation and code debugging, translation, result interpretation, figure preparation, literature review, and manuscript drafting and editing. They were not used to generate synthetic datasets or conduct qualitative or thematic data analysis; other non-applicable disclosure categories are omitted. All AI-assisted mathematical arguments, code, experimental interpretations, literature-related statements, figures, and manuscript text were independently reviewed and verified by the authors. Code was inspected and tested before use, theoretical claims were checked for consistency, and literature-related statements were verified against the original sources. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

CSR is developed for research on spatial multi-omics representation learning and spatial domain identification. All experiments are conducted on publicly available spatial multi-omics datasets, and no new human-subject data are collected in this work. The method is intended to support computational analysis of tissue organization rather than clinical diagnosis or medical decision-making. Because conclusions obtained from spatial omics analyses may depend on dataset composition, preprocessing, and reference annotations, downstream biological interpretation should be made with appropriate domain expertise and validation. We encourage users to follow the applicable data-use terms, privacy requirements, and ethical guidelines when applying CSR to additional datasets or biomedical studies.

REPRODUCIBILITY STATEMENT

To ensure reproducibility, we provide complete implementation details of CSR, including data preprocessing, relational graph construction, model architecture, training settings, and evaluation protocols. The main hyperparameters for representation learning, multiscale graph context extraction, and synchronization dynamics are specified in the paper and appendix. Complete experimental results, sensitivity analyses, computational efficiency measurements, and theoretical proofs are also included in the appendix. Source code and instructions for reproducing all experiments are publicly available through an anonymous repository:`https://anonymous.4open.science/r/CSR-demo-C17D/`.

REFERENCES

Nicolas Boumal. An introduction to optimization on smooth manifolds. *To appear with Cambridge University Press,(Apr 2022)*, 2020.

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.

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

Ruben Dries, Qian Zhu, Rui Dong, Chee-Huat Linus Eng, Huipeng Li, Kan Liu, Yuntian Fu, Tianxiao Zhao, Arpan Sarkar, Feng Bao, et al. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome biology*, 22(1):78, 2021.

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.

Feng Gao, Guy Wolf, and Matthew Hirn. Geometric scattering for graph data analysis. In *International Conference on Machine Learning*, pp. 2122–2131. PMLR, 2019.

Aoyun Geng, Chunyan Cui, Zhenjie Luo, Junlin Xu, Yajie Meng, Feifei Cui, Leyi Wei, Quan Zou, and Zilong Zhang. Computational methods for spatial multi-omics integration. *Biotechnology Advances*, pp. 108807, 2026.

John Healy and Leland McInnes. Uniform manifold approximation and projection. *Nature Reviews Methods Primers*, 4(1):82, 2024.

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

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.

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.

Haoyang Li, Juexiao Zhou, Zhongxiao Li, Siyuan Chen, Xingyu Liao, Bin Zhang, Ruochi Zhang, Yu Wang, Shiwei Sun, and Xin Gao. A comprehensive benchmarking with practical guidelines for cellular deconvolution of spatial transcriptomics. *Nature communications*, 14(1):1548, 2023.

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

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

Sophia K Longo, Margaret G Guo, Andrew L Ji, and Paul A Khavari. Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. *Nature Reviews Genetics*, 22 (10):627–644, 2021.

Takeru Miyato, Sindy Löwe, Andreas Geiger, and Max Welling. Artificial kuramoto oscillatory neurons. In *International Conference on Learning Representations*, volume 2025, pp. 44278–44322, 2025.

David I Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE signal processing magazine*, 30(3):83–98, 2013.

Marco Varrone, Daniele Tavernari, Albert Santamaria-Martínez, Logan A Walsh, and Giovanni Ciriello. Cellcharter reveals spatial cell niches associated with tissue remodeling and cell plasticity. *Nature genetics*, 56(1):74–84, 2024.

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.

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.

Yu Zhang, Mengzhu Wang, Luoxi Jing, Shanshan Wang, Nan Yin, Kaixiang Yang, and Hongxing Zhang. Protopsr: Prototype-guided pairwise similarity regularization for robust spatial multi-omics integration. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 12940–12951, 2026.

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

A Representative Relation-Guided Geometry-Shaping Terms

This section provides representative examples of the relation-guided geometry-shaping terms formalized in Eq. 3. The comparison is restricted to explicit objectives that directly constrain latent geometry. Modality-specific encoders, reconstruction objectives, graph construction, and multimodal fusion are therefore outside the scope of this mapping.

Table 4: Representative relation-guided geometry-shaping objectives.

Method	Geometry-shaping mechanism	Relation instance
MISO	Spectral graph regularization	Graph pair (i, j)
SMART	Triplet metric regularization	Triplet (a, p, n)
SpaMosaic	Cross-view contrastive regularization	$(i^{(m)}, i^{(n)}, \mathcal{N}_i^-)$
PRAGA	Prototype contrastive regularization	$(i, p_{q_i}, \{p_k\})$
ProtoPSR	Prototype contrastive and pairwise regularization	$(i, p_{q_i}), (i, j)$

Two representative cases are detailed below to illustrate the mapping.

MISO. After adapting the notation to this paper, the spectral regularization term can be written as

$$\mathcal{L}_{\mathrm{SC}} = \frac{1}{2} \sum_{i,j} A_{ij} \left\| \mathbf{u}_i - \mathbf{u}_j \right\|_2^2 . \tag{19}$$

Each graph-connected pair (i, j) defines a relation instance, with the corresponding pairwise distance directly regularizing latent geometry. The term therefore follows the formulation in Eq. 3.

SMART. SMART constructs anchor–positive–negative triplets and applies a margin-based metric-learning objective. Under the notation used in this paper,

$$\mathcal{L}_{\mathrm{tri}} = \sum_{(a,p,n)\in\mathcal{T}} \left[d(\mathbf{u}_a, \mathbf{u}_p) - d(\mathbf{u}_a, \mathbf{u}_n) + m \right]_+ . \tag{20}$$

Each triplet specifies relative geometric relations between the anchor–positive and anchor–negative pairs, and therefore constitutes an instance of Eq. 3.

A.1 Relation-Conditioned Representation Learning

Some spatial multi-omics methods exploit relational information without introducing an explicit geometry-shaping objective of the form in Eq. 3. SpatialGlue and COSMOS, for example, incorporate relational structure through mechanisms such as graph propagation, attention-based aggregation, or spatial regularization. These methods are therefore treated as relation-conditioned representation learning rather than explicit instances of relation-guided geometry regularization.

Other methods in the empirical comparison are not included in Table 4 when their objectives do not explicitly instantiate the geometry-shaping formulation in Eq. 3.

B Proof of Proposition 1

Proof. Consider the relation-guided regularizer in Eq. 3. For a latent entity $\mathbf{u}_i$, only relation instances involving i contribute to $\nabla_{\mathbf{u}_i} \mathcal{R}_{\mathrm{rel}}$. By the chain rule, the gradient is a linear combination of terms of the form

$$\nabla_{\mathbf{u}_i} \kappa(\mathbf{u}_i, \mathbf{u}_j), \qquad j \in \mathcal{N}_{\mathcal{R}}(i). \tag{21}$$

By the assumption in Proposition 1,

$$\nabla_{\mathbf{u}_i} \kappa(\mathbf{u}_i, \mathbf{u}_j) \in \mathrm{span}\{\mathbf{u}_i, \mathbf{u}_j\}. \tag{22}$$

Since $\nabla_{\mathbf{u}_i} \mathcal{R}_{\mathrm{rel}}$ is a linear combination of these derivatives,

$$\nabla_{\mathbf{u}_i} \mathcal{R}_{\mathrm{rel}} \in \mathrm{span}\left(\{\mathbf{u}_i\} \cup \{\mathbf{u}_j : j \in \mathcal{N}_{\mathcal{R}}(i)\} \right). \tag{23}$$

For hinge-based objectives, the same conclusion holds for any valid subgradient at nondifferentiable points. □

C Proofs of Theoretical Results

C.1 Proof of Theorem 1

Proof. Consider the product manifold

$$\mathcal{M} = (\mathbb{S}^{d-1})^N.$$

For the objective in Eq. 5, symmetry of $\mathbf{A}$ gives

$$\nabla_{\mathbf{z}_i} \mathcal{E} = \beta \sum_j A_{ij}(\mathbf{z}_i - \mathbf{z}_j) - \alpha \mathbf{c}_i. \tag{24}$$

Since $\Pi_{\mathbf{z}_i}(\mathbf{z}_i) = \mathbf{0}$, projecting Eq. 24 onto the tangent space yields

$$\operatorname{grad}_{\mathbf{z}_i} \mathcal{E} = -\Pi_{\mathbf{z}_i} \left(\beta \sum_j A_{ij} \mathbf{z}_j + \alpha \mathbf{c}_i \right). \tag{25}$$

Hence, the continuous dynamics in Eq. 10 satisfy

$$\dot{\mathbf{z}}_i = -\operatorname{grad}_{\mathbf{z}_i} \mathcal{E}. \tag{26}$$

Applying the chain rule on $\mathcal{M}$ gives

$$\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}t} = \sum_{i=1}^{N} \langle \operatorname{grad}_{\mathbf{z}_i} \mathcal{E}, \dot{\mathbf{z}}_i \rangle \tag{27}$$

$$= -\sum_{i=1}^{N} \left\| \operatorname{grad}_{\mathbf{z}_i} \mathcal{E} \right\|_2^2 \leq 0. \tag{28}$$

Therefore, the objective is non-increasing along the continuous synchronization trajectory.

It remains to show that the objective is bounded from below. Since $A_{ij} \geq 0$ and $\beta > 0$, the graph-dependent term satisfies

$$\frac{\beta}{4} \sum_{i,j} A_{ij} \left\| \mathbf{z}_i - \mathbf{z}_j \right\|_2^2 \geq 0. \tag{29}$$

Moreover, $\|\mathbf{z}_i\|_2 = 1$ and the Cauchy–Schwarz inequality imply

$$-\alpha \mathbf{c}_i^\top \mathbf{z}_i \geq -\alpha \|\mathbf{c}_i\|_2. \tag{30}$$

Consequently,

$$\mathcal{E} \geq -\alpha \sum_{i=1}^{N} \|\mathbf{c}_i\|_2. \tag{31}$$

Thus, the objective value is non-increasing and bounded from below, and therefore converges to a finite limit along the continuous trajectory. □

C.2 Proof of Proposition 2

Proof. Consider a fully synchronized state $\mathbf{z}_i = \mathbf{z}$ for all $i = 1, \ldots, N$, where $\mathbf{z} \in \mathbb{S}^{d-1}$. The graph-coupling component in Eq. 10 becomes

$$\beta \sum_j A_{ij} \mathbf{z}_j = \beta \left(\sum_j A_{ij} \right) \mathbf{z}, \tag{32}$$

which is parallel to $\mathbf{z}$. Therefore,

$$\Pi_{\mathbf{z}} \left(\beta \sum_j A_{ij} \mathbf{z}_j \right) = \mathbf{0}. \tag{33}$$

Table 5: Summary of the spatial multi-omics datasets used in our experiments.

Dataset	Spots	Omics	Technology	Domains
Mouse Embryonic Brain E15.5	1,949	RNA + ATAC	MISAR-seq	12
Human Lymph Node A1	3,484	RNA + ADT	Visium CytAssist	10
Human Tonsil Section 1	4,326	RNA + ADT	Visium CytAssist	4
Mouse Brain	9,196	RNA + ATAC	Spatial ATAC–RNA-seq	14
Simulation	1,296	RNA + ATAC + ADT	Simulation	5

The dynamics at a fully synchronized state consequently reduce to

$$\dot{\mathbf{z}}_i = \alpha \Pi_{\mathbf{z}}(\mathbf{c}_i). \tag{34}$$

Since $\alpha > 0$, the synchronized state is an equilibrium if and only if

$$\Pi_{\mathbf{z}}(\mathbf{c}_i) = \mathbf{0}, \qquad i = 1, \ldots, N. \tag{35}$$

Using the tangent projection in Eq. 2,

$$\Pi_{\mathbf{z}}(\mathbf{c}_i) = \mathbf{c}_i - (\mathbf{z}^\top \mathbf{c}_i)\mathbf{z}. \tag{36}$$

Hence, $\Pi_{\mathbf{z}}(\mathbf{c}_i) = \mathbf{0}$ if and only if $\mathbf{c}_i \in \mathrm{span}\{\mathbf{z}\}$, which proves the proposition. □

C.3 Implement details

Datasets. Evaluation is performed on five publicly available, annotated spatial multi-omics datasets spanning three modality combinations: (1) **RNA+protein/ADT:** Human Tonsil Section 1 (Yan et al., 2026) and Human Lymph Node A1 (Long et al., 2024); (2) **RNA+ATAC:** Mouse Embryonic Brain E15.5 (Jiang et al., 2023) and Mouse Brain (Zhang et al., 2023); and (3) **RNA+ADT+ATAC:** a simulated tri-omics dataset (Simulation) (Long et al., 2024). Dataset statistics are provided in Table 5.

Data preprocessing. Each modality is preprocessed separately following established practice in spatial multi-omics analysis (Long et al., 2024). RNA, ADT, and ATAC modalities were preprocessed separately according to their data characteristics. For RNA, genes detected in fewer than 10 spots were removed for HT1, HLN, MBE15, and MB, while MB additionally excluded spots expressing fewer than 200 genes. The top 3,000 highly variable genes were selected using the Seurat v3 method. RNA counts were normalized to a total of 10^4 per spot and log-transformed; the real datasets were further standardized before dimensionality reduction. PCA was then applied to obtain 30-dimensional representations for HT1, HLN, and MBE15, a 50-dimensional representation for MB, and a 100-dimensional representation for the simulated dataset. For ADT, protein measurements were normalized using centered log-ratio (CLR) normalization and reduced by PCA, yielding 30-dimensional representations for HT1 and HLN and a 100-dimensional representation for the simulated dataset; ADT features in HT1 and HLN were additionally standardized before PCA. For ATAC, preprocessing depended on the dataset. In MBE15, ATAC observations were aligned with the RNA spot identifiers, followed by TF–IDF normalization, per-spot normalization to 10^4, log transformation, and PCA to obtain a 60-dimensional representation. For MB, the provided latent semantic indexing (LSI) representation was used when available; otherwise, LSI was computed from the ATAC matrix, resulting in a 50-dimensional representation. The simulated ATAC modality was processed using LSI to obtain a 100-dimensional representation. Spatial coordinates were obtained from the corresponding spatial AnnData objects, primarily from the RNA modality, and the provided reference annotations were aligned with the retained spots and encoded as integer labels for evaluation.

Baselines. The comparison includes eight representative methods for spatial multi-omics analysis: CellCharter, SpatialGlue, PRAGA, COSMOS, SMART, MISO, ProtoPSR, and SpaMosaic. Together, these methods cover a range of strategies for incorporating neighborhood relations and structuring multimodal latent representations for spatial domain discovery.For clarity, the main text presents five representative baselines, while retaining all nine evaluation metrics. Complete comparisons across all eight baselines are provided in Appendix D.

Metrics. Nine metrics are adopted to evaluate clustering quality and spatial coherence. Eight reference-based metrics (ARI, AMI, NMI, FMI, F1, Jaccard, Homogeneity (Homo), and Completeness (Compl)) assess agreement between inferred domains and reference annotations from different

perspectives. The remaining metric, sLISI evaluates local spatial coherence without relying on reference annotations.

Implementation Details. The framework is implemented in PyTorch and trained in full-graph mode, with modality-specific pretraining followed by joint fine-tuning. We use a hidden dimension of 512 and a unified representation dimension of 128.The graph scattering transform uses two scales and a maximum scattering order of two.We construct a spatial k-nearest-neighbor (kNN) graph from the spatial coordinates and a modality-specific feature graph from the molecular measurements of each modality. The spatial and molecular relations are combined through a union operation to obtain a relational graph. The resulting adjacency matrix is symmetrized and normalized before being used for representation learning. The spatial and feature neighborhood sizes are set to $k_s = [8]$ and $k_f = [4]$, respectively. The same relational graph is used for multiscale graph context extraction and neighbor coupling during synchronization.Multiscale graph context is extracted using a graph scattering transform (GST) with two scales and a maximum scattering order of two. The scattering responses are concatenated with the fused multimodal representation and passed through a context network to obtain a location-specific contextual reference. During each synchronization trajectory, the contextual reference remains fixed, while the representation states evolve on the unit hypersphere. We perform $K = [24]$ synchronization steps using a projected Euler update, with a step size of $\eta = [0.1]$, relational coupling strength $\beta = [1]$, and contextual strength $\alpha = [0.01]$. The representation states are normalized after each update to maintain the unit-norm constraint. Experiments are conducted on an NVIDIA GeForce RTX 3090 GPU.

D Complete Quantitative Results

Table 6: Clustering results on the RNA and ADT datasets.

Methods	ARI	AMI	NMI	FMI	HOMO	COMPL	F1	Jaccard	sLISI
Human Tonsil Section 1									
CellCharter	10.83	18.37	18.45	45.80	19.07	17.86	45.67	29.60	**87.18**
COSMOS	10.16	18.06	18.14	43.61	19.76	16.76	43.25	27.59	84.31
MISO	20.20	25.36	25.43	49.65	27.62	23.57	49.15	32.58	71.09
SpaMosaic	15.21	25.59	25.65	43.77	29.34	22.79	42.52	27.00	73.33
SpatialGlue	17.69	28.50	28.57	48.33	30.72	26.70	48.51	32.02	67.83
PRAGA	31.17	31.92	31.99	58.35	33.66	30.47	58.35	41.04	66.96
ProtoPSR	24.58	29.17	29.24	55.08	30.32	28.23	55.03	37.96	69.23
SMART	36.46	35.86	35.92	61.87	**37.55**	34.42	61.77	44.69	71.09
Ours	**41.86**	**38.18**	**38.25**	**67.61**	37.47	**39.05**	**67.54**	**50.99**	73.33
Human Lymph Node A1									
CellCharter	17.12	33.21	33.63	31.88	38.46	29.88	30.49	17.99	**94.77**
COSMOS	18.65	33.15	33.58	34.50	37.48	30.42	33.97	20.46	93.06
MISO	26.72	35.42	35.85	47.11	33.63	39.78	46.37	30.18	88.46
SpaMosaic	17.98	34.78	35.18	32.23	40.94	30.85	30.29	17.85	91.95
SpatialGlue	27.25	39.03	39.43	41.61	43.50	36.05	40.99	25.78	83.94
PRAGA	27.86	38.11	38.51	42.09	42.54	35.19	41.44	26.14	82.38
ProtoPSR	30.08	40.86	41.26	44.80	43.97	38.87	44.63	28.72	88.49
SMART	35.23	42.79	43.18	48.78	**46.51**	40.29	48.54	32.05	87.75
Ours	**38.13**	**43.63**	**44.03**	**53.49**	44.28	**43.78**	**53.33**	**36.36**	88.99

E Sensitivity and Efficiency Analysis

Sensitivity Analysis. Sensitivity to the context strength γ and the number of synchronization steps T is evaluated on three datasets. As shown in Fig. 6, the model remains stable across a broad range of settings, with moderate context strengths generally yielding better clustering performance, while further synchronization beyond a certain point brings only marginal gains. These results demonstrate the robustness of the framework and highlight the importance of balancing contextual guidance with neighbor coupling during iterative synchronization.

Runtime and memory usage. We evaluate computational efficiency in terms of runtime and peak GPU memory usage. All experiments are conducted on an NVIDIA GeForce RTX 3090 GPU. Run-

Table 7: Clustering results on the RNA and ATAC datasets.

Methods	ARI	AMI	NMI	FMI	HOMO	COMPL	F1	Jaccard	sLISI
Mouse Embryonic Brain E15.5									
CellCharter	45.50	60.78	61.34	53.88	65.54	57.66	52.70	35.78	94.64
COSMOS	33.84	49.32	50.05	43.55	53.37	47.11	42.65	27.10	89.88
MISO	18.30	35.65	36.61	34.73	34.25	39.32	34.23	20.65	84.72
SpaMosaic	54.87	61.51	62.06	61.98	65.60	58.89	61.26	44.16	**94.94**
SpatialGlue	44.96	58.81	59.41	53.36	62.99	56.21	52.56	35.65	90.58
PRAGA	53.22	64.55	65.07	60.53	67.85	62.51	60.14	43.00	92.26
ProtoPSR	50.76	63.46	63.99	58.42	67.72	60.65	57.60	40.45	92.12
SMART	61.39	66.45	67.03	67.53	70.20	64.14	67.31	50.73	93.33
Ours	**70.94**	**69.12**	**69.59**	**75.75**	**70.67**	**68.54**	**75.74**	**60.96**	92.53
Mouse Brain									
CellCharter	28.82	35.14	35.40	37.51	35.31	35.48	37.45	23.04	**98.26**
COSMOS	18.81	29.44	29.71	27.17	30.84	28.67	26.94	15.57	96.04
MISO	23.95	37.19	37.44	34.55	34.95	40.31	34.11	20.56	85.33
SpaMosaic	24.90	33.37	33.63	32.91	34.55	32.76	32.78	19.61	96.38
SpatialGlue	30.26	38.15	38.34	40.33	35.58	41.56	39.70	24.77	91.99
PRAGA	32.97	41.22	41.46	41.05	41.13	41.79	41.01	25.80	95.91
ProtoPSR	34.05	39.22	39.47	41.42	40.14	38.81	41.40	26.10	93.33
SMART	38.14	43.93	44.15	45.31	44.47	43.84	45.31	29.29	93.62
Ours	**43.50**	**46.80**	**47.02**	**50.54**	**46.29**	**47.77**	**50.43**	**33.72**	90.12

Table 8: Clustering results on the simulated spatial multi-omics dataset.

Methods	ARI	AMI	NMI	FMI	HOMO	COMPL	F1	Jaccard	sLISI
Simulation									
CellCharter	97.71	97.43	97.44	98.23	97.57	97.30	98.23	96.52	90.38
COSMOS	65.43	74.44	74.54	72.93	75.86	73.27	72.81	57.25	**94.51**
MISO	17.31	46.60	46.84	42.03	42.34	52.41	40.70	25.50	75.22
SpaMosaic	96.10	95.18	95.20	96.98	95.44	94.97	96.98	94.13	90.38
SpatialGlue	97.62	96.79	96.80	98.16	96.90	96.70	98.16	96.38	90.38
PRAGA	99.36	99.07	99.07	99.51	99.10	99.05	99.51	99.02	90.38
ProtoPSR	99.15	98.80	98.81	99.34	98.82	98.80	99.34	98.69	90.38
SMART	97.85	96.96	96.97	98.34	97.00	96.94	98.34	96.73	90.38
Ours	**99.78**	**99.66**	**99.66**	**99.83**	**99.65**	**99.68**	**99.83**	**99.66**	90.38

time includes both training and inference, excluding data preprocessing, downstream clustering, and file I/O.Peak GPU memory usage is measured as the maximum allocated GPU memory during execution.As shown in Table 9, CSR completes training and inference in 6.10 seconds, with a peak GPU memory usage of 4.72 GB. It requires substantially less runtime than ProtoPSR, PRAGA, CellCharter, COSMOS, MISO, and SMART, although SpatialGlue and SpaMosaic exhibit shorter runtimes. In terms of memory usage, our method requires less GPU memory than ProtoPSR, PRAGA, COSMOS, and MISO, while consuming more memory than CellCharter, SpatialGlue, SpaMosaic, and SMART. These results indicate that our method achieves relatively low computational cost while incorporating multiscale graph context extraction and synchronization dynamics.

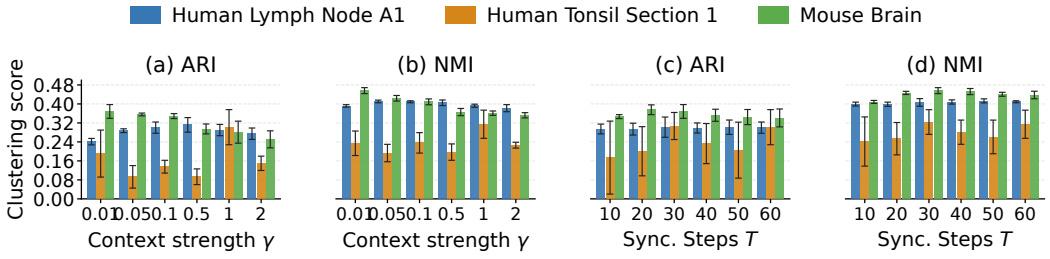

Figure 6: Analysis of context strength and synchronization steps.

Table 9: Runtime and GPU memory usage of different methods.

Method	Runtime (s)	GPU memory (GB)
ProtoPSR	2545.48	14.45
PRAGA	2557.93	11.28
CellCharter	3143.51	2.17
COSMOS	108.72	14.34
MISO	32.14	6.64
SpatialGlue	3.35	0.17
SpaMosaic	2.74	0.34
SMART	23.64	1.30
CSR	**6.10**	**4.72**