Multi-Scale Flow Matching for Continuous-time Generative Modeling of Spatiotemporal Tissue Dynamics from Spatial Transcriptomics

Pinar Demetci [1] **Tim Guan** [2] **Bo Xia** [3 4 5] **Lazar Atanackovic** [6 7]

Abstract

Understanding how tissue organization changes over time in response to signals from the local cellular environment is a fundamental challenge in developmental biology, cancer biology, and regenerative medicine. Spatial transcriptomics enables the characterization of cell states within their native microenvironment. Yet tissue remodeling unfolds continuously, while spatial transcriptomics only captures fragments, yielding unmatched cross-sectional snapshots from a few time points due to its destructive nature. This necessitates computational approaches to reconstruct dynamics from sparse observations. Existing methods either do not model the temporal dynamics or ignore cell signaling effects. Existing methods either recover discrete cell-cell correspondences without modeling dynamics, treat cells as independent particles ignoring cell-cell interactions and microenvironmental context, or consider fixed-size neighborhood effects without supporting temporal modeling beyond two time points. We introduce ChronoTILE , a multi-scale, multi-marginal flow matching framework that jointly models continuous-time spatiotemporal tissue dynamics. ChronoTILE accounts for cell-cell interactions, and models niche-mediated signaling effects, across the full course of a biological process.

[1]Eric and Wendy Schmidt Center, Broad Institute of MIT and Harvard, Cambridge, MA, USA [2]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA [3]Gene Regulation Observatory, Broad Institute of MIT and Harvard, Cambridge, MA, USA [4]Harvard Medical School, Boston, MA, USA [5]Massachusetts General Hospital, Boston, MA, USA [6]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada [7]Department of Biochemistry, University of Alberta, Edmonton, AB, Canada. Correspondence to: Pinar Demetci <demetci@broadinstitute.edu>, Lazar Atanackovic <atanacko@ualberta.ca>.

Accepted at the Workshop on Generative and Agentic AI for Biology at the 43^{rd} International Conference on Machine Learning, Seoul, South Korea, 2026.

1. Introduction

Throughout development, disease progression, tissue regeneration, and homeostasis, cells transit between functional states shaped by continuous crosstalk with their local tissue environment. These local tissue environments (or "spatial niches"[1]) actively regulate cell state transitions through signaling via secreted ligands or mechanical cues, such that the same progenitor cell type, for example, may follow distinct differentiation trajectories depending on which niche it inhabits (Lee et al., 2025a; Heitz et al., 2025).

Spatial transcriptomics technologies have enabled us to study these processes by jointly measuring cells' gene expression profiles and spatial positions at single-cell resolution, placing each cell's transcriptional state directly within its native tissue context (Ståhl et al., 2016; Marx, 2021). When collected from multiple time points, such snapshots are a powerful tool for studying tissue remodeling dynamics (Heitz et al., 2025) They also raise fundamental questions about how remodeling is coordinated: which cell populations interact during, and what molecular signals mediate their communication? How do these interactions drive transcriptional state changes at different stages of the process? How do niches emerge, reorganize, and instruct neighboring populations over time?

Yet, the destructive nature of sequencing technologies imposes a fundamental limitation: Each slide captures a snapshot of a distinct tissue specimen, yielding no direct correspondence between cells across time points. Temporal observations must therefore be inferred from independent cross-sectional populations. Compounding this, spatial transcriptomics experiments are expensive and time-intensive, resulting in datasets with limited sample sizes from a few and often irregularly spaced time points, and considerable biological variability across tissue slices (Heitz et al., 2025). These constraints necessitate computational models capable of inferring continuous tissue remodeling dynamics from sparse, unmatched snapshots while accounting for the spatial structure and cell-cell interactions that govern these processes.

[1]A niche refers to a spatially localized tissue microenvironment characterized by a specific composition of cell types and molecular signals that collectively define its functional identity.

Existing approaches address parts of this problem but fall short of a comprehensive solution, as discussed in Section 2. Some recover cell-cell correspondences between discrete time points without modeling the dynamic process (Klein et al., 2025), and others model cells as independent particles without accounting for interactions or microenvironmental effects (Peng et al., 2026). Although the most recent approach makes progress towards modeling the microenvironment (Sakalyan et al., 2025), it defines "niches" by a fixed spatial radius rather than in a data-driven and size-adaptive manner. It additionally does not model the full temporal evolution across all available time points, instead generating tissue profiles only between two fixed time points (Sakalyan et al., 2025). Importantly, except for stVCR, none of the others methods can generate a tissue profile at arbitrary, intermediate (e.g. held-out) time points; however, stVCR's formulation limits its computational scalability. Together, these leave a gap in generative modeling of multi-marginal temporal dynamics of tissue architecture changes while taking into account microenvironment- or niche-level signaling effects.

To address this, we introduce ChronoTILE , a multi-scale, multi-marginal flow matching framework for modeling continuous tissue remodeling dynamics from spatial transcriptomics snapshots. ChronoTILE learns a generative model of the joint transcriptomic and spatial changes in tissue remodeling by parameterizing the velocity field underlying these changes with a hierarchical Transformer architecture that captures cell-cell interactions and niche-level effects. Rather than imposing fixed neighborhood sizes, ChronoTILE discovers spatial niches in a data-driven manner, accommodating the variable compositions and sizes of real tissue microenvironments that form functional units. By operating across all available time points together, rather than solely between pairs, ChronoTILE models the full temporal evolution of tissue architecture, enabling inference of how niche-mediated interactions drive cell state transitions across the course of a biological process.

2. Related Works

2.1. Discrete correspondence methods

Early works sought to establish probabilistic cell-cell correspondences across consecutive time points. **moscot** (Klein et al., 2025) computes such correspondences via fused unbalanced Gromov-Wasserstein (FUGW) OT, jointly accounting for transcriptomic similarity and spatial structure through intra-slice pairwise distances, with cell masses initialized from proliferation scores. **SpaTrack** (Shen et al., 2025) uses a similar FUGW formulation with uniform marginals, while additionally deriving spatial velocity vectors from the transport plan to organize cells into differentiation trajectories. While principled and interpretable, these methods produce discrete transport plans between adjacent time point pairs and do not learn a generative model of dynamics, precluding prediction at unseen time points or continuous interpolation between observations.

2.2. Continuous generative models

More recently, efforts have shifted to go beyond discrete correspondences to leverage generative models, typically to learn a continuous model of tissue dynamics. **STORIES** (Huizing et al., 2025) uses an FGW gradient flow approach to learn a Waddington-style potential landscape J_θ over gene expression, with dynamics governed by $-\nabla J_\theta$ and trained using an FGW objective that incorporates spatial structure. While it yields interpretable pseudotime and velocity outputs, the potential is expression-only, so STORIES cannot predict future spatial positions and does not model cell-cell interactions or niche-level effects. Furthermore, because the dynamics are constrained to descend a scalar potential, the framework cannot represent cyclical or oscillatory cellular processes such as the cell cycle, where cells return to previously visited transcriptomic states.

stVCR (Peng et al., 2026) extends dynamical OT to jointly reconstruct continuous differentiation, proliferation, and physical migration via three coupled neural networks in a neural ODE framework, and simultaneously learns rigid-body alignment parameters across time points. However, it treats each cell's dynamics independently without modeling local microenvironment structure, and training via dynamical OT requires solving forward ODE integrations at every gradient step, which is a computationally intensive procedure that can become numerically unstable when dynamics are stiff, in contrast to the simulation-free flow matching objective used in ChronoTILE .

NicheFlow (Sakalyan et al., 2025) is the most closely related work, modeling generative dynamics of cellular microenvironments as point clouds via variational flow matching with entropic OT. However, it operates only between two biological time points: it learns to generate the spatial transcriptomic profile of a target tissue slice noise distribution by transporting it to the target, conditioned on the source slice at the previous time point. The continuous-time axis in NicheFlow is the flow matching interpolation parameter, not biological time, so the model has no notion of the rate of change of tissue state over real time and cannot generate profiles at arbitrary biological time points. Furthermore, niches are defined by fixed-radius neighborhoods rather than discovered data-adaptively, and each microenvironment is processed independently with no mechanism for inter-niche communication.

2.3. Our contributions

ChronoTILE addresses these complementary limitations by learning a continuous-time multi-marginal velocity field jointly over all time points, discovering variable-size niches in a data-driven manner, and explicitly modeling both cell–cell interactions through a multi-scale Transformer architecture. Table 1 summarizes the key differences with existing methods. Notably, ChronoTILE is the first generative model to produce spatial transcriptomic tissue profiles at arbitrary, held-out biological time points while explicitly accounting for microenvironmental effects.ChronoTILE is also trained via an efficient, simulation-free flow matching objective, in contrast to methods like stVCR (Peng et al., 2026) that require computationally expensive forward ODE integration at every training step.

3. Methods

Given spatial transcriptomic snapshots $\mathcal{D}^t = (\mathbf{X}^t, \mathbf{S}^t)$ from time points $t \in \{t_0, t_1, \ldots, t_T\}$, where $\mathbf{X}^t \in \mathbb{R}^{n_t \times d}$ denotes gene expression (i.e. *"transcriptomic"*) measurements for n_t cells and d genes, and $\mathbf{S}^t \in \mathbb{R}^{n_t \times 2}$ specifies the corresponding spatial coordinates in 2-dimensional tissue slice, ChronoTILE aims to learn a generative model of the continuous-time dynamics underlying cells' transcriptomic states and spatial organization. Critically, cells are destroyed upon measurement, so there is no direct correspondence of cells across time points in the dataset.

ChronoTILE uses a multi-scale Transformer to parameterize a time-dependent velocity field $v_\theta \colon \mathbb{R}^{n\times(d+2)} \times [0,1] \to \mathbb{R}^{n\times(d+2)}$. This describes the instantaneous rate of change of joint cell states $(\mathbf{X}, \mathbf{S})$ over continuous time. This multi-scale Transformer models the hierarchical organization of tissues, accounting for cell-cell interactions *within* niches, then niche-level interactions that may capture longer-range tissue-level coordination. The overall methodology is presented in the schematic in Figure 1. Spatial niche assignments are learned from data in a pre-computed stage (Figure 1A), which involves an encoder-decoder Transformer, inspired by CellTransformer (Lee et al., 2025a), that encodes "neighborhood representations" for each cell. These "neighborhood representations" are learned through a self-supervised training procedure that decodes cells' expression profiles and then are clustered to assign niches. We describe this procedure in Section B. We further detail the Transformer architecture and our design rationale in Section 3.2. To train the velocity field in a computationally efficient manner, we use a multi-marginal flow matching framework using the temporally-sparse spatial transcriptomic measurements, as detailed in Section 3.1. At inference, ChronoTILE generates predicted tissue states at arbitrary time points by integrating the learned velocity field via an ordinary differential equation (ODE) solver.

3.1. Multi-marginal flow matching for spatiotemporal tissue dynamics

3.1.1. Background on conditional flow matching (CFM)

Modeling the spatiotemporal evolution of tissues naturally defines a neural ODE problem: we seek a time-dependent velocity field v_θ such that integrating

$$\frac{d(\mathbf{X}, \mathbf{S})}{dt} = v_\theta(\mathbf{X}, \mathbf{S}, \mathbf{c}, t), \qquad (\mathbf{X}, \mathbf{S}) \in \mathbb{R}^{n\times(d+2)}, \quad (1)$$

forward in time reproduces the observed tissue dynamics, where $\mathbf{c} \in \{1, \ldots, K\}^n$ denotes the niche assignments for

Table 1. Comparison of spatiotemporal modeling methods for spatial transcriptomics. Methods considered in our benchmarking experiments so far are denoted with *. stVCR is not included due to the recency of their code release. NicheFlow has "∼" in the "Generative modeling" row because while it is technically a generative model, it generates spatial transcriptomic profiles from noise conditioned on the slices from a previous time point, and therefore is not a generative model of temporal dynamics and cannot be used to generate tissue profiles at arbitrary time points.

	moscot*	SpaTrack	STORIES*	stVCR	NicheFlow*	ChronoTILE*
Generative model	×	×	✓	✓	∼	✓
Generates full spatial transcriptomics	×	×	×	✓	×	✓
Multi-marginal (>2 timepoints)	×	×	✓	✓	×	✓
Continuous-time dynamics	×	×	✓	✓	×	✓
Models cell–cell interactions	×	×	×	×	✓	✓
Data-driven niche discovery	×	×	×	×	×	✓
Works with variable-size niches	×	×	×	×	×	✓
Models inter-niche communication	×	×	×	×	×	✓

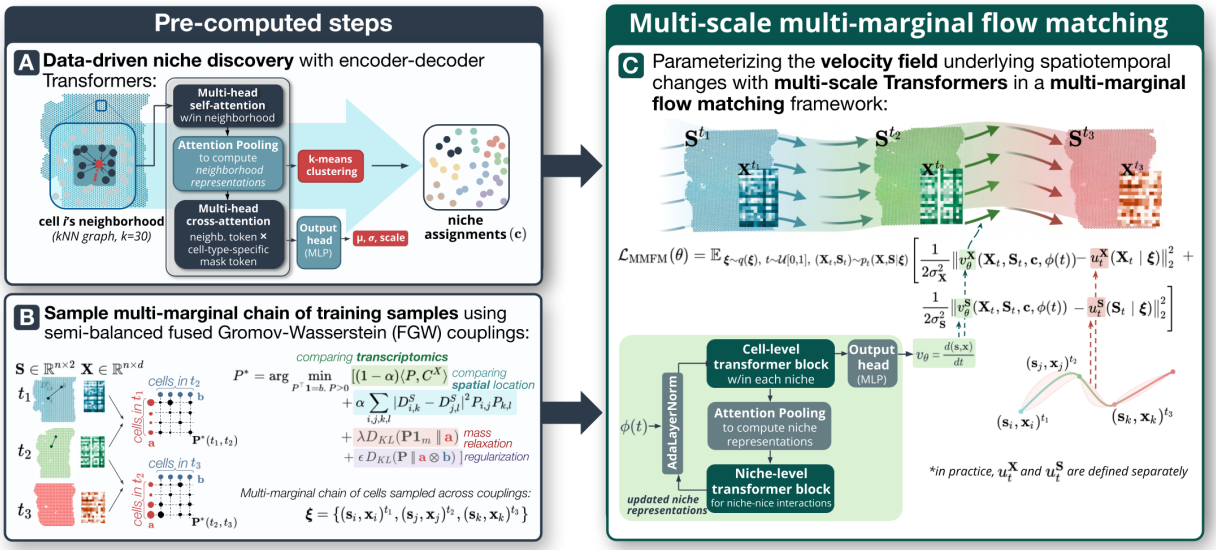

Figure 1. **Overview of ChronoTILE .** ChronoTILE learns continuous-time generative models of tissue remodeling dynamics from spatial transcriptomic snapshots at discrete time points, in three steps: **(A) Data-driven niche discovery.** An encoder-decoder Transformer learns a local neighborhood representation for each cell by masking its expression profile and reconstructing it via cross-attention over neighboring cells in a kNN graph (k=30). Neighborhood representations are clustered via k-means. Based on these clusters, cells are assigned discrete labels, $\mathbf{c}$, after ensuring that unconnected regions of the tissue with the same cluster label are assigned unique labels. These niche labels $\mathbf{c}$ are then passed as input to the velocity field Transformer in (C). More detailed information is in Appendix Section B **(B) Multi-marginal chain sampling.** Semi-balanced fused Gromov–Wasserstein (FGW) couplings between consecutive time points yield multi-marginal chains $\boldsymbol{\xi} = \{(\mathbf{s}_i, \mathbf{x}_i)^{t_1}, (\mathbf{s}_j, \mathbf{x}_j)^{t_2}, (\mathbf{s}_k, \mathbf{x}_k)^{t_3}\}$ used as training trajectories. **(C) Multi-scale velocity field and MMFM training.** A multi-scale Transformer parameterizes v_θ via cell-level self-attention within niches, niche-level self-attention for tissue-scale coordination, and AdaLayerNorm conditioning on time embedding $\phi(t)$. It is trained under $\mathcal{L}_{\text{MMFM}}(\theta)$, regressing v_θ onto reference velocities $u_t^{\mathbf{X}}$ and $u_t^{\mathbf{S}}$ derived from spline interpolation through $\boldsymbol{\xi}$.

K niches (Section B). Neural ODEs are typically trained by repeatedly solving forward ODE integrations at every gradient step, which is computationally expensive and can become numerically unstable when the dynamics are stiff. The conditional flow matching (CFM) framework (Lipman et al., 2022) sidesteps this by instead regressing v_θ onto a closed-form conditional velocity field induced by a prescribed interpolation between sampled endpoint pairs $(\mathbf{X}_0, \mathbf{S}_0)$, $(\mathbf{X}_1, \mathbf{S}_1)$:

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{\mathcal{P}} \left\| v_\theta(\mathbf{X}_t, \mathbf{S}_t, \mathbf{c}, t) - u_t(\mathbf{X}_t, \mathbf{S}_t \mid \mathbf{X}_0, \mathbf{S}_0, \mathbf{X}_1, \mathbf{S}_1) \right\|_2^2 \tag{2}$$

where $\mathcal{P}$ denotes the joint distribution over $t \sim \mathcal{U}(0,1)$, $(\mathbf{X}_0, \mathbf{S}_0, \mathbf{X}_1, \mathbf{S}_1) \sim \pi$, and $(\mathbf{X}_t, \mathbf{S}_t) \sim p_t(\cdot \mid \mathbf{X}_0, \mathbf{S}_0, \mathbf{X}_1, \mathbf{S}_1)$; and π is a coupling over pairs from consecutive time points, either from independent marginal sampling (Lipman et al., 2022) or a structured coupling, such as from optimal transport (Tong et al., 2023).

3.1.2. Background on Multi-Marginal Flow Matching (MMFM)

The CFM objective in Eq. (2), however, is defined only between two time points. In many studies investigating spatiotemporal changes of tissues across biological processes (Chen et al., 2022; Wei et al., 2022; Cadinu et al., 2024), including the axolotl telencephalon regeneration process we study, datasets span $T+1$ time points. Therefore, we instead adopt the multi-marginal flow matching (MMFM) framework (Lee et al., 2025b) and generalize CFM by conditioning on a full chain of cells $\boldsymbol{\xi} = \{(\mathbf{x}_i^{t_0}, \mathbf{s}_i^{t_0}), \ldots, (\mathbf{x}_k^{t_T}, \mathbf{s}_k^{t_T})\}$ sampled across all time points:

$$\mathcal{L}_{\text{MMFM}}(\theta) = \mathbb{E}_{\mathcal{Q}} \left\| v_\theta(\mathbf{X}_t, \mathbf{S}_t, \mathbf{c}, \phi(t)) - u_t(\mathbf{X}_t, \mathbf{S}_t \mid \boldsymbol{\xi}) \right\|_2^2, \tag{3}$$

where $\mathcal{Q}$ denotes the joint distribution over $\boldsymbol{\xi} \sim q(\boldsymbol{\xi})$, $t \sim \mathcal{U}[0,1]$, and $(\mathbf{X}_t, \mathbf{S}_t) \sim p_t(\mathbf{X}, \mathbf{S} \mid \boldsymbol{\xi})$. Here, $q(\boldsymbol{\xi})$ is the distribution over multi-marginal chains obtained via semi-balanced fused Gromov-Wasserstein (FGW) optimal transport couplings (Section 3.1.4); $p_t(\mathbf{X}, \mathbf{S} \mid \boldsymbol{\xi})$ is a conditional path through the chain; $u_t(\mathbf{X}_t, \mathbf{S}_t \mid \boldsymbol{\xi})$ is the corresponding analytic reference velocity; and $\phi(t)$ is a sinusoidal time embedding.

3.1.3. Our MMFM formulation for spatial transcriptomics

Since $\mathbf{X}_t \in \mathbb{R}^{n \times d}$ and $\mathbf{S}_t \in \mathbb{R}^{n \times 2}$ live in spaces of vastly different dimensionality and dynamic range, the squared norm in Eq. (3) is dominated by the transcriptomic term. To balance the two modalities without introducing additional tunable hyperparameters, we normalize each term by its

expected squared magnitude under the reference velocity, estimated once from the data prior to training:

$$\mathcal{L}_{\text{MMFM}}(\theta) = \mathbb{E}_{\mathcal{Q}}\left[\frac{1}{2\sigma_{\mathbf{X}}^2}\left\|v_\theta^{\mathbf{X}}(\mathbf{X}_t, \mathbf{S}_t, \mathbf{c}, \phi(t)) - u_t^{\mathbf{X}}(\mathbf{X}_t \mid \boldsymbol{\xi})\right\|_2^2 + \frac{1}{2\sigma_{\mathbf{S}}^2}\left\|v_\theta^{\mathbf{S}}(\mathbf{X}_t, \mathbf{S}_t, \mathbf{c}, \phi(t)) - u_t^{\mathbf{S}}(\mathbf{S}_t \mid \boldsymbol{\xi})\right\|_2^2\right] \tag{4}$$

where $v_\theta^{\mathbf{X}}$ and $v_\theta^{\mathbf{S}}$ are the transcriptomic and spatial output heads of v_θ respectively, $u_t^{\mathbf{X}}$ and $u_t^{\mathbf{S}}$ are the corresponding modality-specific reference velocities derived from coordinate-wise spline interpolation (Section 3.1.4), and

$$\sigma_{\mathbf{X}}^2 = \mathbb{E}_{\mathcal{Q}}\left[\|u_t^{\mathbf{X}}\|_2^2\right], \qquad \sigma_{\mathbf{S}}^2 = \mathbb{E}_{\mathcal{Q}}\left[\|u_t^{\mathbf{S}}\|_2^2\right] \tag{5}$$

are fixed normalizing constants computed over a large sample of chains and uniformly drawn $t \sim \mathcal{U}[0,1]$ before training begins. By construction, each term in Eq. (4) has expected value $\frac{1}{2}$ at initialization, ensuring that transcriptomic and spatial supervision contribute equally to the total loss throughout training. We next describe the construction of conditional paths and reference velocities (Section 3.1.4).

3.1.4. Defining conditional probability paths via spline interpolation

We construct conditional probability paths $p_t(\mathbf{X}, \mathbf{S} \mid \boldsymbol{\xi})$ by first sampling multi-marginal chains $\boldsymbol{\xi}$ of cells across all time points via sequential composition of semi-balanced FGW couplings, then fitting a monotone cubic Hermite spline (PCHIP) through the chain knots to define smooth mean paths $\boldsymbol{\mu}_{\mathbf{X}}^{\boldsymbol{\xi}}(t)$ and $\boldsymbol{\mu}_{\mathbf{S}}^{\boldsymbol{\xi}}(t)$ for the transcriptomic and spatial modalities respectively. To encourage learning of smooth interpolating velocities rather than memorization of exact trajectories, we place a modality-specific time-varying variance $\tau_m^2(t, \boldsymbol{\xi})$ around each spline mean, yielding block-factorized Gaussian conditional paths with separate noise models for transcriptomic and spatial coordinates. The analytic reference velocities $u_t^{\mathbf{X}}$ and $u_t^{\mathbf{S}}$ used in Eq. (4) are then derived in closed form from the spline derivatives and the variance schedule, without any ODE integration at training time. We describe each of these components in detail below:

Fused Gromov-Wasserstein OT couplings. For a given pair of consecutive time points t_1 and t_2, we compute the semi-balanced FGW coupling between cells as:

$$\mathbf{P}^*(t_1, t_2) = \arg\min_{\mathbf{P} \geq 0} \alpha \sum_{i=1}^{n_{t_1}} \sum_{j=1}^{n_{t_2}} \tilde{C}_{ij}^{\mathbf{x}} P_{ij} + (1-\alpha) \sum_{i,k=1}^{n_{t_1}} \sum_{j,l=1}^{n_{t_2}} \tilde{C}_{ikjl}^{\mathbf{s}} P_{ij} P_{kl} + \epsilon\, \mathrm{KL}(\mathbf{P} \,\|\, \mathbf{a} \otimes \mathbf{b}) + \lambda\, \mathrm{KL}(\mathbf{P}\mathbf{1} \,\|\, \mathbf{a}) \tag{6}$$

where the transcriptomic and spatial cost matrices are each normalized by their respective medians prior to optimization:

$$\tilde{C}_{ij}^{\mathbf{x}} = \frac{\|\mathbf{x}_i^{t_1} - \mathbf{x}_j^{t_2}\|^2}{\text{median}_{i,j}\|\mathbf{x}_i^{t_1} - \mathbf{x}_j^{t_2}\|^2}, \qquad \tilde{C}_{ikjl}^{\mathbf{s}} = \frac{\|D_{ik}^{\mathbf{s}} - D_{jl}^{\mathbf{s}}\|^2}{\text{median}_{i,k,j,l}\|D_{ik}^{\mathbf{s}} - D_{jl}^{\mathbf{s}}\|^2} \tag{7}$$

with $\mathbf{x}_i^{t_1} \in \mathbb{R}^d$ denoting the gene expression profile of cell i at time t_1, and $D_{ik}^{\mathbf{s}} = \|\mathbf{s}_i^{t_1} - \mathbf{s}_k^{t_1}\|^2$ the squared Euclidean distance between the spatial coordinates of cells i and k at time t_1. Without this normalization, the transcriptomic cost, being a sum over d gene dimensions, exceeds the spatial cost in magnitude, rendering α ineffective as an interpolation weight. Median normalization ensures $\alpha \in [0,1]$ meaningfully controls the trade-off between the two terms. The hyperparameter α therefore interpolates the transcriptomic similarity term (first line) against the spatial structural alignment term (second line), which compares intra-time-point pairwise distances $D^{\mathbf{s}}$ rather than absolute coordinates, and therefore does not require spatial alignment across time points.

The vectors $\mathbf{a} \in \mathbb{R}^{n_{t_1}}$ and $\mathbf{b} \in \mathbb{R}^{n_{t_2}}$ are the prescribed mass distributions over cells at t_1 and t_2, respectively; $\mathbf{b}$ is uniform, while $\mathbf{a}$ is set proportionally to gene-expression-derived cell proliferation scores computed via `scanpy`'s `score_genes` function on a curated set of cell-cycle marker genes, so that cells with higher estimated proliferation rates contribute more outgoing mass. The entropic regularization term $\epsilon\,\mathrm{KL}(\mathbf{P}\|\mathbf{a}\otimes\mathbf{b})$ enables efficient computation and encourages soft many-to-many matchings, while the semi-balanced term $\lambda\,\mathrm{KL}(\mathbf{P}\mathbf{1}\|\mathbf{a})$ allows deviation from the growth prior when it is uncertain.

Sampling multi-marginal chain of cells for training. Based on these couplings, we sample multi-marginal training chains following the rolling window strategy of (Lee et al., 2025c). For consecutive time points t_1, t_2, t_3, we first sample a source cell i at t_1 proportionally to its outgoing mass $a_i = \sum_j P_{ij}^*(t_1, t_2)$, then sample its matched cell j at t_2 from the row-normalized conditional:

$$(\mathbf{x}_i^{t_1}, \mathbf{s}_i^{t_1},\ \mathbf{x}_j^{t_2}, \mathbf{s}_j^{t_2}) \sim \pi_{12}, \qquad \pi_{12}(j \mid i) = \frac{P_{ij}^*(t_1, t_2)}{\sum_{j'} P_{ij'}^*(t_1, t_2)}.$$

Given j, we sample its successor cell k at t_3 from the next coupling:

$$(\mathbf{x}_k^{t_3}, \mathbf{s}_k^{t_3}) \sim \pi_{23}(\cdot \mid j), \qquad \pi_{23}(k \mid j) = \frac{P_{jk}^*(t_2, t_3)}{\sum_{k'} P_{jk'}^*(t_2, t_3)}.$$

Extending this procedure sequentially across all $T+1$ time points yields the full multi-marginal chain $\boldsymbol{\xi} = \{(\mathbf{x}_i^{t_0}, \mathbf{s}_i^{t_0}), (\mathbf{x}_j^{t_1}, \mathbf{s}_j^{t_1}), \ldots, (\mathbf{x}_k^{t_T}, \mathbf{s}_k^{t_T})\}$, consistent with the definition of $\boldsymbol{\xi}$ in Eq. (3).

Defining conditional probability paths with splines. Given a sampled multi-marginal chain

$\boldsymbol{\xi} = \{(\mathbf{x}_i^{t_0}, \mathbf{s}_i^{t_0}), \ldots, (\mathbf{x}_k^{t_T}, \mathbf{s}_k^{t_T})\}$, we define smooth conditional probability paths $p_t(\mathbf{X}, \mathbf{S} \mid \boldsymbol{\xi})$ as Gaussian paths centered on a spline mean and perturbed by a time-varying variance. Since $\mathbf{X}_t$ and $\mathbf{S}_t$ live in spaces of incompatible geometry and scale, we define a block-factorized conditional path:

$$p_t(\mathbf{X}, \mathbf{S} \mid \boldsymbol{\xi}) = p_t^{\mathbf{X}}(\mathbf{X} \mid \boldsymbol{\xi})\, p_t^{\mathbf{S}}(\mathbf{S} \mid \boldsymbol{\xi}), \tag{8}$$

where each block is an isotropic Gaussian:

$$p_t^{\mathbf{X}}(\mathbf{X} \mid \boldsymbol{\xi}) = \mathcal{N}\Big(\mathbf{X};\ \boldsymbol{\mu}_{\mathbf{X}}^{\boldsymbol{\xi}}(t),\ \tau_{\mathbf{X}}^2(t, \boldsymbol{\xi})\, \mathbf{I}\Big), \tag{9}$$

$$p_t^{\mathbf{S}}(\mathbf{S} \mid \boldsymbol{\xi}) = \mathcal{N}\Big(\mathbf{S};\ \boldsymbol{\mu}_{\mathbf{S}}^{\boldsymbol{\xi}}(t),\ \tau_{\mathbf{S}}^2(t, \boldsymbol{\xi})\, \mathbf{I}\Big). \tag{10}$$

This factorization defines separate noise models per modality and does not impose biological independence: both blocks are conditioned on the same multi-marginal chain $\boldsymbol{\xi}$ and share the same biological time t, and the neural velocity field v_θ remains joint so that transcriptomic velocity can depend on spatial context and vice versa.

Spline mean paths. The mean paths $\boldsymbol{\mu}_{\mathbf{X}}^{\boldsymbol{\xi}}(t)$ and $\boldsymbol{\mu}_{\mathbf{S}}^{\boldsymbol{\xi}}(t)$ are obtained by fitting a monotone cubic Hermite spline (PCHIP; Fritsch & Carlson 1980) through the chain knots $\{(\mathbf{x}^{t_k}, t_k)\}_{k=0}^T$ and $\{(\mathbf{s}^{t_k}, t_k)\}_{k=0}^T$, respectively, using all $T+1$ time points simultaneously. Because the spline is fitted coordinate-wise, fitting the two modalities jointly on the concatenated state or separately is equivalent; we use a single PCHIP fit over the full chain and read off the transcriptomic and spatial components of the derivative. The PCHIP construction assigns tangents at interior knots via the Fritsch-Carlson weighted harmonic mean of the neighboring secant slopes, and applies a one-sided three-point formula at the endpoints. This guarantees C^1 continuity and shape preservation: interpolated states remain within the convex hull of neighboring observations and the spline never overshoots, which is important for biologically plausible reference paths. Unlike the natural cubic splines used in (Lee et al., 2025b), PCHIP does not minimize global bending energy and therefore allows the acceleration to jump at knot points, which is appropriate for biological systems that can undergo transcriptional bursts or rapid spatial reorganization.

Modality-specific adaptive variance. The path variance $\tau_m^2(t, \boldsymbol{\xi})$ for each modality $m \in \{\mathbf{X}, \mathbf{S}\}$ is chosen to vanish at the observed knot times and peak between them, following the piece-wise schedule:

$$\tau_m^2(t, \boldsymbol{\xi}) = M_m^2 \cdot \frac{(t - t_k)^2\,(t_{k+1} - t)^2}{(t_{k+1} - t_k)^2}, \qquad t \in [t_k, t_{k+1}], \tag{11}$$

where t_k and t_{k+1} are the knot times bracketing t, and $M_m > 0$ is a modality-specific scale hyperparameter. This schedule is structurally equivalent to the time-dependent variance used in (Lee et al., 2025b) and vanishes exactly at every observed knot, so that sampling $(\mathbf{X}_t, \mathbf{S}_t) \sim p_t(\cdot \mid \boldsymbol{\xi})$ at a knot time recovers the observed cell state. The noise added between knots encourages the model to learn a smooth interpolating velocity rather than memorizing exact trajectories, and provides the signal-sharing benefit across time points identified in (Lee et al., 2025b).

Analytic reference velocities. For an isotropic Gaussian path with time-varying variance, the analytic conditional reference velocity for modality m is (Lipman et al., 2022):

$$u_t^m(\mathbf{Y}_t \mid \boldsymbol{\xi}) = \dot{\boldsymbol{\mu}}_m^{\boldsymbol{\xi}}(t) + \frac{\dot{\tau}_m(t, \boldsymbol{\xi})}{\tau_m(t, \boldsymbol{\xi})}\Big(\mathbf{Y}_t - \boldsymbol{\mu}_m^{\boldsymbol{\xi}}(t)\Big), \tag{12}$$

where $\dot{\boldsymbol{\mu}}_m^{\boldsymbol{\xi}}(t)$ is the PCHIP spline derivative (available in closed form as $p'(s) = 3as^2 + 2bs + c$ for the local polynomial coefficients) and $\dot{\tau}_m(t, \boldsymbol{\xi})$ is the time derivative of Eq. (11):

$$\dot{\tau}_m(t, \boldsymbol{\xi}) = M_m \cdot \frac{(t - t_k)(t_{k+1} - t)(t_k + t_{k+1} - 2t)}{(t_{k+1} - t_k)^2 \sqrt{\frac{(t-t_k)^2(t_{k+1}-t)^2}{(t_{k+1}-t_k)^2}}}, \quad \text{with } t \in (t_k, t_{k+1}). \tag{13}$$

The second term in Eq. (12) is a score-like correction that pulls the sampled noisy state $\mathbf{Y}_t$ back toward the spline mean as t approaches a knot. This reference velocity is evaluated at training time without any ODE integration.

3.2. Multi-scale transformers

We parameterize the velocity field v_θ with a multi-scale Transformer that mirrors the multi-scale organization of tissues: cells are embedded within niches, and niches are spatially arranged across the tissue. Rather than applying global self-attention over all cells, which would scale quadratically with n and ignore biological structure, we decompose the computation into three successive stages that process information at the cell ("Stage A"), niche ("Stage B"), and tissue ("Stage C") levels respectively, as detailed below and visualized in Figure 7. The network is conditioned on flow time t via Adaptive Layer Normalization (AdaLN) in Stage A.

Cell-level transformer block: self-attention within niches (Stage A). Each cell's input token is formed by concatenating two modality-specific projections: its h highly variable gene expression features are projected into a $d_{\mathbf{X}}$-dimensional subspace, and its 2D spatial coordinates are projected into a $d_{\mathbf{S}}$-dimensional subspace via separate learned linear layers, with $d_{\mathbf{X}} + d_{\mathbf{S}} = d$. The two projections are concatenated to form a d-dimensional input token that preserves modality identity throughout the network, which is necessary for decoding modality-specific velocity predictions at the output. Spatial coordinates additionally en-

ter the attention computation via Rotary Position Embeddings (RoPE) (Su et al., 2024), which encode relative spatial relationships by rotating query and key vectors, injecting translation-equivariant spatial context into the attention geometry without modifying the token features directly. Cells are grouped by their niche assignment $\mathbf{c}$, and multi-head self-attention is applied independently within each niche group using a block-diagonal attention mask, so cells in different niches do not interact at this stage. This captures cell–cell interactions and co-expression patterns within the local microenvironment. The restriction to within-niche attention is motivated by computational efficiency and is biologically justified, as most cell–cell interactions occur locally through secreted ligands or mechanical cues. Longer-range effects, such as morphogen gradients or boundary signals between adjacent niches, are instead captured by Stage C.

Each Stage A block follows the adaLN-Zero design of Li et al. (2022). The flow time t is encoded as a d-dimensional sinusoidal embedding $\phi(t)$, concatenated with the niche representation $\mathbf{e}_{\text{niche}} \in \mathbb{R}^d$ of the cell's assigned niche, and passed through a fusion MLP to produce a conditioning vector $\mathbf{c} = \text{MLP}(\phi(t) \oplus \mathbf{e}_{\text{niche}}) \in \mathbb{R}^d$. A zero-initialized linear projection of $\mathbf{c}$ predicts six parameter vectors $(\gamma_1, \beta_1, \alpha_1, \gamma_2, \beta_2, \alpha_2)$ that modulate the pre-attention LayerNorm, self-attention residual gate, pre-MLP LayerNorm, and MLP residual gate via scale, shift, and gating respectively. The zero initialization ensures each Stage A block acts as an identity function at the start of training, providing stability while the niche representations are still uninformative. The conditioning is per-cell: each cell looks up the representation of its own assigned niche, so cells within the same niche share niche-level context while retaining individual cell-level representations. The per-cell velocity outputs $v_\theta^{\mathbf{X}}$ and $v_\theta^{\mathbf{S}}$ are read from the final Stage A block via the output head described below, retaining cell-level resolution while having been informed by niche- and tissue-level context injected through adaLN conditioning in earlier rounds.

Attention pooling from cells to niche tokens (Stage B). After Stage A, a single learnable query vector performs cross-attention over the cell tokens within each niche, with a padding mask to handle variable niche sizes, producing a d-dimensional niche embedding $\mathbf{e}_{\text{niche}}$ for each niche. This compresses the distributed cell-level representations into a compact summary of each niche's current state. The niche embeddings are normalized via plain LayerNorm and passed forward into Stage C. The updated niche representations returned from Stage C are subsequently fed back into Stage A's adaLN conditioning in the next round, enabling iterative refinement of cell representations with progressively more informed niche context.

Niche-level transformer block: self-attention across all niches (Stage C). The niche tokens attend to one another via multi-head self-attention with 2D RoPE applied to niche centroids, encoding the relative spatial arrangement of niches across a tissue slide. Stage C blocks use plain LayerNorm rather than AdaLN, as temporal information enters niche representations implicitly through the cell representations from which they were pooled in Stage B. A subsequent MLP produces updated niche representations that are designed to capture longer-range tissue-level coordination, such as signaling gradients or boundary effects between adjacent niche types. Each cell then performs a per-cell lookup of its assigned niche's updated representation, which is fused with the sinusoidal time embedding and used to condition the next round of Stage A via the AdaLN transformation layer.

Output head. Per-cell velocity vectors are produced from the final Stage A block's cell token representations by a two-layer MLP, yielding the joint velocity $v_\theta(\mathbf{X}_t, \mathbf{S}_t, \mathbf{c}, \phi(t)) \in \mathbb{R}^{n \times (d_{\mathbf{X}} + d_{\mathbf{S}})}$, whose transcriptomic ($v_\theta^{\mathbf{X}} \in \mathbb{R}^{n \times d_{\mathbf{X}}}$) and spatial ($v_\theta^{\mathbf{S}} \in \mathbb{R}^{n \times d_{\mathbf{S}}}$) components are supervised separately by the modality-balanced MMFM loss (Eq. (4)). No additional normalization is applied in the output head, as the cell representations entering it are already normalized and rescaled by the AdaLN inside the final Stage A block. At inference, an ODE solver integrates the predicted velocity field to obtain spatiotemporal profile predictions at unseen intermediate time points.

4. Results

We model the continuous-time spatiotemporal dynamics of axolotl telencephalon regeneration using the spatial transcriptomics dataset generated by Wei et al. (2022). This dataset was collected using Stereo-seq (spatial enhanced resolution omics sequencing) at single-cell resolution, spanning seven time points (2, 5, 10, 15, 20, 30, and 60 DPI; days post-injury) following surgical injury to the dorsal pallium of the left telencephalic hemisphere of adult axolotls (*Ambystoma mexicanum*). With this dataset, we first evaluate the fidelity of learned tissue dynamics by **ChronoTILE** and then interpret the interactions captured by the learned velocity field to investigate what they reveal about the telencephalon regeneration process.

4.1. Benchmarking Dynamics Modeling Performance

We benchmark ChronoTILE against three existing methods, namely moscot (Klein et al., 2025), STORIES (Huizing et al., 2025), and NicheFlow (Sakalyan et al., 2025), on how well they generate tissue profiles at different time points as a proxy for the quality of their learned dynamics, using two categories of evaluation metrics: gene expression fidelity,

and spatial architecture fidelity (detailed in Appendix D). We exclude stVCR (Peng et al., 2026) due to the recency of their functional code release but plan to include it in future experiments. Given that none of the baselines are designed to generate full spatial transcriptomic profiles at arbitrary time points, we evaluate each method in its most natural operational mode under two scenarios:

1. **Interpolation of observed time points:** All methods are trained on the full dataset and evaluated on their ability to reconstruct tissue profiles at two of the time points seen during training (10 and 30 DPI).
2. **Generalization to held-out time points:** Each method is assessed for their ability to generate tissue profiles at a time point withheld entirely from training (10 and 30 DPI, held out separately).

ChronoTILE integrates its trained velocity field forward from the first observed time point (2 DPI) via an ODE solver to generate tissue profiles at the target time point. The only difference between the two scenarios is whether the target time point's ground-truth profile was seen during training.

STORIES generates gene expression profiles at the target time point by applying its learned potential function sequentially from 2 DPI via forward Euler steps. As STORIES generates gene expression data only, we exclude it from spatial fidelity benchmarking.

NicheFlow trains a single generative model conditioned on source and target slide labels across all pairs of adjacent time points. In both evaluation scenarios, we generate profiles at the target time point by autoregressively chaining pairwise flows from 2 DPI, feeding each step's generated output as the source for the next. In the interpolation scenario, the model is trained on all adjacent pairs. In the held-out scenario, the model is retrained excluding pairs involving the held-out time point (e.g., for 10 DPI held out, this leaves the training pairs of 2→5, 15→20, 20→30, and 30→60 DPI), and the autoregressive chain proceeds from 2 DPI to the target using the dynamics learned from the remaining pairs. As NicheFlow's flow matching interpolation parameter is an internal denoising variable rather than biological time, it cannot represent the rate of biological change and is not expected to account for differences in temporal gap length between steps, which is an inherent limitation of the method here.

Lastly, we adapt **moscot** differently across the two scenarios as it is not a generative model. It, however, does provide functions to leverage its coupling matrices to project spatial transcriptomic data from one time point to another seen time point. In the interpolation scenario, we compose the FGW coupling matrices moscot produces for adjacent time points via its `cell_transition()` function. With this, we obtain a push-forward distribution over observed cells at the target time point, starting from 2 DPI. In the held-out scenario, `cell_transition()` is inapplicable as it requires a solved OT coupling involving the target time point's cells. In this case, we instead use moscot to compute an FGW coupling between the two observed slices bracketing the held-out target (5 and 15 DPI for the 10 DPI target; 20 and 60 DPI for the 30 DPI target) and interpolate between them at weight $\lambda = \Delta t_{\text{pre}}/(\Delta t_{\text{pre}} + \Delta t_{\text{post}})$, where Δt_{pre} and Δt_{post} are the temporal gaps from the held-out time point to its preceding and following observed slices, respectively.

Figure 3 displays the results, where each neural model

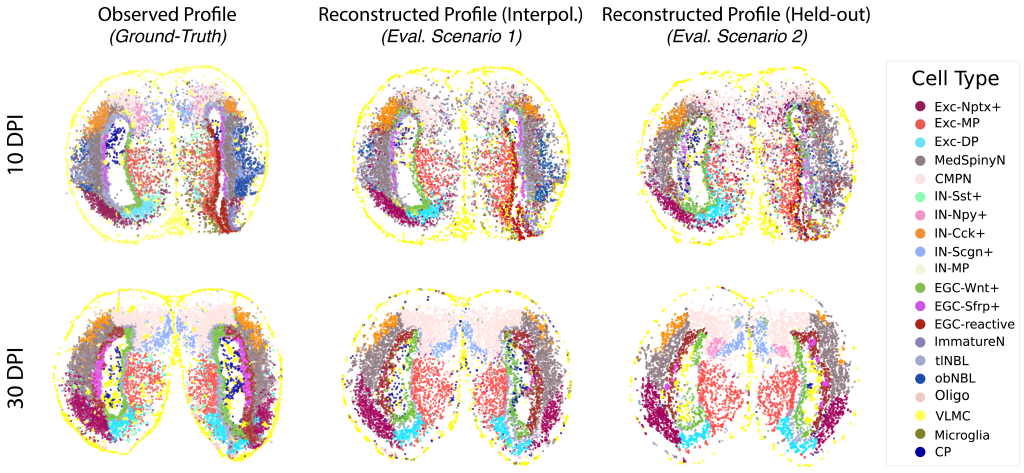

Figure 2. **Visualization of tissue profiles generated by ChronoTILE** under both evaluation scenarios, compared to the observed (ground-truth) profiles at 10 DPI and 30 DPI. In both scenarios, the trained model is integrated forward in time, starting from 2 DPI to generate profiles at 10 DPI and 30 DPI. In the first scenario, training data contained observed profiles from 10 DPI and 30 DPI, while in the second scenario, these were held out. Cell type abbreviations are described in the caption of Figure 5.

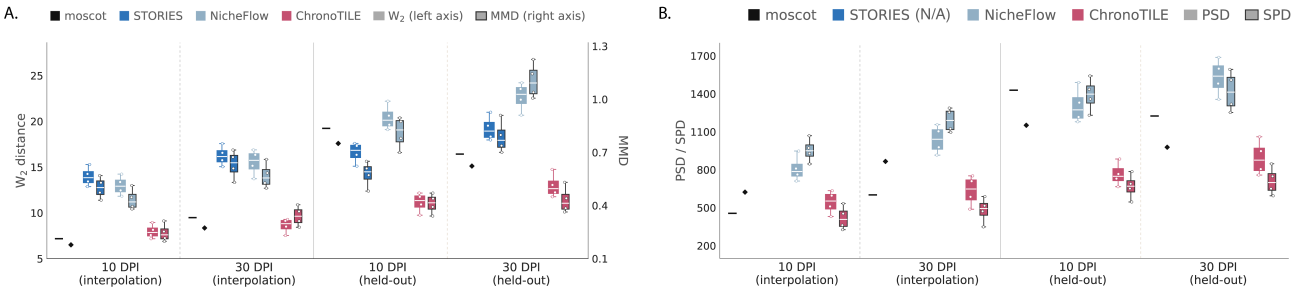

Figure 3. **Benchmarking ChronoTILE against existing methods using gene expression and spatial coordinate reconstruction error metrics. A. Evaluation on gene expression reconstruction,** as measured by Wasserstein (W_2) distance and maximum mean discrepacy (MMD), both of which measure reconstruction error, and therefore, lower values are better. The left axis gives the W_2 distance, while the right axis gives the MMD values. **B. Evaluation on spatial coordinate reconstruction,** as measured by point-to-shape (PSD) and shape-to-point distances (SPD), as described in Appendix Section D. PSD asks whether each generated point lies near the true tissue (akin to a measure of precision); SPD asks whether each ground-truth point is covered by a generated point (akin to measure of recall). Each neural method is benchmarked across four runs with different random seeds. Moscot is run only once and the W_2 values are shown by the horizontal line while the MMD values are shown with the diamond marker.

(ChronoTILE , STORIES, and NicheFlow) is run with four different random seeds upon hyperparameter tuning. As moscot's entropic FGW solver is deterministic given fixed inputs,[2] we report benchmarking results based only on a single run.

Among the neural methods, ChronoTILE achieves the strongest overall performance across the evaluation scenarios. While we observe that reconstruction errors generally increase at later target time points, consistent with the greater difficulty of predicting over longer temporal horizons, the increase is substantially milder for ChronoTILE than for STORIES or NicheFlow. This highlights the advantage of ChronoTILE's multi-marginal formulation, which leverages cells from all observed training time points jointly to learn a single continuous-time model of the dynamics. NicheFlow performs particularly poorly in the held-out setting, where the task requires generating profiles at biological time points that were not observed during training. This is consistent with a limitation of NicheFlow for learning temporal dynamics: its flow-matching time corresponds to an internal denoising process in interpolating between pairs of profiles rather than biological time, making it less suitable for modeling temporal dynamics beyond interpolating between two time points.

An exception to these trends is moscot, reflecting the distinct evaluation setting we use due to its non-generative mode of operation. In the interpolation setting, moscot performs strongly because unlike the neural methods which learn parameterized dynamics across all observed profiles, it instead composes couplings to obtain a push-forward distribution at the target time point, over cells that it already uses to solve the final coupling specific to this time point. Thus, its reconstruction is directly supported by couplings involving the ground-truth profile it is asked to recover. Although this makes the benchmarking less directly comparable, it remains an informative correspondence-based baseline for the neural models. In the held-out setting, however, this advantage disappears because no coupling involving the target slice can be computed. Moscot instead reconstructs the missing profile by interpolating between the two observed slices that bracket the held-out time point. Interestingly, although the 30 DPI target is separated from its bracketing slices by a larger temporal gap than the 10 DPI target, the 20 and 60 DPI profiles appear more similar to the 30 DPI profile than the 5 and 15 DPI profiles are to the 10 DPI profile (Figure 5), likely explaining moscot's comparatively better performance at 30 DPI than 10 DPI in the held-out scenario. Nevertheless, its overall weaker held-out performance indicates that interpolating between observed profiles is insufficient for reconstructing unseen intermediate tissue states, highlighting the need for generative models of continuous biological dynamics.

Together, these results suggest that ChronoTILE's continuous-time, multi-marginal formulation provides a valuable framework for reconstructing both observed and unseen stages of tissue remodeling than existing generative models or discrete OT-based interpolation.

4.2. Interpreting Cell-Cell Interaction Drivers

After benchmarking ChronoTILE's generative performance, we then interpret the Transformer model that parameterizes the velocity field using integrated gradients (Sundararajan et al., 2017) for the cell interaction drivers it captures. As a reference baseline, we use the "control" profile from the original dataset, which reflects the state of the telencephalon before any injury-inducing intervention (Figure 5C).

[2]The Sinkhorn scaling vectors are initialized uniformly by default in the `ott-jax` package used by moscot, yielding a unique solution for fixed regularization parameters. We do not modify this default initialization.

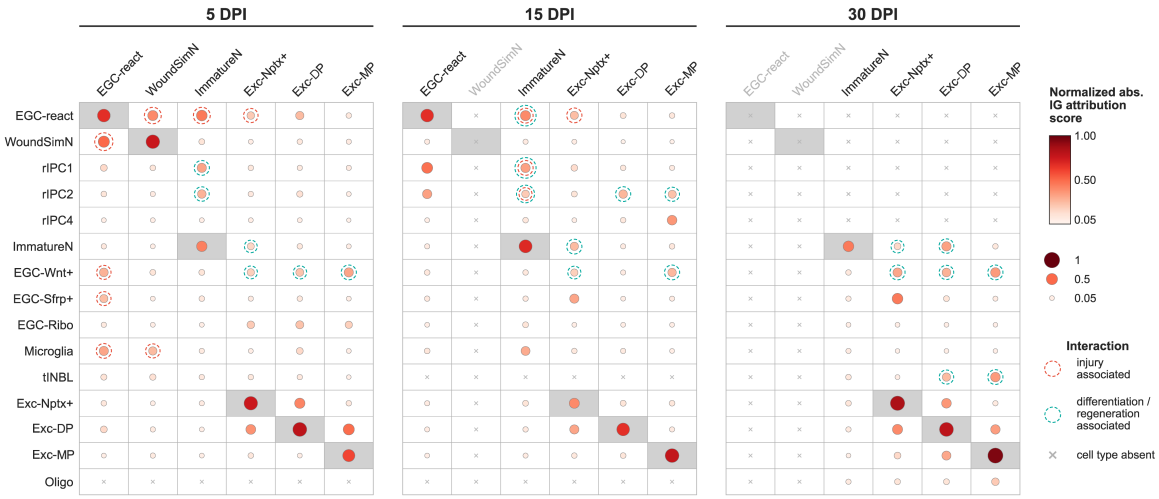

Figure 4. **Learned interactions between cell types within a shared niche across three time points**. Dot plot showing normalized absolute integrated gradient attribution scores for interactions between select "source" cell types (rows) and "target" cell types (columns) at 5, 15, and 30 days post-injury (DPI). Dot size and color reflect the magnitude of the attribution score, indicating the degree to which the presence of a given source cell type in the local niche influences the predicted velocity of the target cell type. Dot size and color reflect the normalized absolute integrated gradient attribution score, quantifying the contribution of the source cell type's input token to the velocity predictions of target cell type within the same niche. These are computed on a cell-pair-level, and then averaged across cell types for visualization. Red dashed border indicates interactions known to be injury-associated (involving wound-stimulated neurons, WSN; microglia; and reactive ependymoglial, EGC-react), while the green dashed borders indicate interactions known to be differentiation/regeneration-associated (involving regeneration intermediate progenitor cells, rIPCs; immature neurons, ImmatureN; telencephalon neuroblasts, tINBL; their interactions with mature neurons such as dorsal pallium excitatory neuron, Exc-DP and medial pallium excitatory neuron, Exc-MP), excluding cell-type-level self-interactions.

Figure 4 summarizes normalized absolute integrated-gradient attributions, averaged at the cell-type level, to identify which cell types most influence velocity predictions within shared niches. Across regeneration, the learned attributions shift from wound-associated interactions at 5 DPI toward cell differentiation-associated interactions at 15 and 30 DPI, involving progenitor, immature neuronal, neuroblast, and mature excitatory neuronal populations. These results suggest that ChronoTILE likely captures biologically meaningful temporal changes in local cell-cell dependencies during tissue remodeling, while providing model-based attributions rather than causal interaction estimates.

5. Discussion

ChronoTILE introduces a multi-scale, multi-marginal flow matching framework for jointly modeling continuous-time transcriptomic and spatial dynamics from unmatched spatial transcriptomic snapshots, while accounting for cell–cell interactions and niche-mediated signaling across all observed time points. Applied to axolotl telencephalon regeneration, ChronoTILE outperforms existing generative models in both interpolation and held-out generalization settings. Integrated gradients analysis of the learned velocity field further recovers cell-cell interactions driving tissue remodeling dynamics.

Several directions remain open. The current interpretability analysis focuses on cell-type-level interactions within niches; extending attribution methods to tissue-wide effects would enable the study of longer-range tissue coordination, which could possibly enable studying signaling gradients between adjacent niche types. A complementary direction is to disentangle the drivers behind gene expression and spatial migration components of the learned velocity field. Adding uncertainty quantification to integrated-gradient scores, for example through bootstrap resampling or permutation tests, would help distinguish statistically significant signals. In addition, the current framework assumes a roughly stable cell population size. Our ongoing work is extending it to accommodate large-scale tissue growth for applications to developmental settings substantial temporal change in number of cells.

ChronoTILE is currently designed and evaluated for interpolation between observed time points. Extrapolation beyond the last observation remains an important but harder problem, requiring generalization outside the training distribution. Another promising direction is perturbation prediction: by conditioning the velocity field on ligand knockdown or cell-type ablation signals, ChronoTILE could help identify signaling pathways whose disruption may steer pathological tissues toward regenerative outcomes, connecting generative tissue modeling to rational therapeutic design.

Impact Statement

This paper presents work that advances generative modeling methodology for spatial transcriptomics, with application to understanding tissue remodeling dynamics in a non-human model organism. We anticipate that such tools that model continuous-time tissue dynamics could accelerate basic biological research into development, regeneration, and disease, while reducing experimental burden by enabling richer inference from existing limited data. We do not foresee immediate societal risks from the use or distribution of this model. However, should the framework be applied to human tissue data in the future, standard data governance practices, including patient consent, data anonymization, and institutional review, would apply to both the use of the model and the distribution of any trained checkpoints.

References

Cadinu, P., Sivanathan, K. N., Misra, A., Xu, R. J., Mangani, D., Yang, E., Rone, J. M., Tooley, K., Kye, Y.-C., Bod, L., Geistlinger, L., Lee, T., Mertens, R. T., Ono, N., Wang, G., Sanmarco, L., Quintana, F. J., Anderson, A. C., Kuchroo, V. K., Moffitt, J. R., and Nowarski, R. Charting the cellular biogeography in colitis reveals fibroblast trajectories and coordinated spatial remodeling. *Cell*, 187 (8):2010–2028.e30, April 2024.

Chen, A., Liao, S., Cheng, M., Ma, K., Wu, L., Lai, Y., Qiu, X., Yang, J., Xu, J., Hao, S., Wang, X., Lu, H., Chen, X., Liu, X., Huang, X., Li, Z., Hong, Y., Jiang, Y., Peng, J., Liu, S., Shen, M., Liu, C., Li, Q., Yuan, Y., Wei, X., Zheng, H., Feng, W., Wang, Z., Liu, Y., Wang, Z., Yang, Y., Xiang, H., Han, L., Qin, B., Guo, P., Lai, G., Muñoz-Cánoves, P., Maxwell, P. H., Thiery, J. P., Wu, Q.-F., Zhao, F., Chen, B., Li, M., Dai, X., Wang, S., Kuang, H., Hui, J., Wang, L., Fei, J.-F., Wang, O., Wei, X., Lu, H., Wang, B., Liu, S., Gu, Y., Ni, M., Zhang, W., Mu, F., Yin, Y., Yang, H., Lisby, M., Cornall, R. J., Mulder, J., Uhlén, M., Esteban, M. A., Li, Y., Liu, L., Xu, X., and Wang, J. Spatiotemporal transcriptomic atlas of mouse organogenesis using dna nanoball-patterned arrays. *Cell*, 185(10):1777–1792.e21, May 2022. ISSN 0092-8674. doi: 10.1016/j.cell.2022.04.003. URL https://doi.org/10.1016/j.cell.2022.04.003.

Fritsch, F. N. and Carlson, R. E. Monotone piecewise cubic interpolation. *SIAM Journal on Numerical Analysis*, 17 (2):238–246, 1980. doi: 10.1137/0717021. URL https://doi.org/10.1137/0717021.

Heitz, M., Ma, Y., Kubal, S., and Schiebinger, G. Spatial transcriptomics brings new challenges and opportunities for trajectory inference. *Annual Review of Biomedical Data Science*, 8(1):1–19, 2025.

Huizing, G.-J., Samaran, J., Capocefalo, D., Audit, A., Peyré, G., and Cantini, L. Stories: learning cell fate landscapes from spatial transcriptomics using optimal transport. *Nature Methods*, pp. 1–10, 2025.

Klein, D., Palla, G., Lange, M., Klein, M., Piran, Z., Gander, M., Meng-Papaxanthos, L., Sterr, M., Saber, L., Jing, C., et al. Mapping cells through time and space with moscot. *Nature*, 638(8052):1065–1075, 2025.

Lee, A. J., Dubuc, A., Kunst, M., Yao, S., Lusk, N., Ng, L., Zeng, H., Tasic, B., and Abbasi-Asl, R. Data-driven fine-grained region discovery in the mouse brain with transformers. *Nature Communications*, 16(1):8536, 2025a.

Lee, J., Moradijamei, B., and Shakeri, H. Multi-marginal stochastic flow matching for high-dimensional snapshot data at irregular time points. *arXiv preprint arXiv:2508.04351*, 2025b.

Lee, J., Moradijamei, B., and Shakeri, H. Multi-marginal stochastic flow matching for high-dimensional snapshot data at irregular time points. *arXiv preprint arXiv:2508.04351*, 2025c.

Li, J., Xu, Y., Lv, T., Cui, L., Zhang, C., and Wei, F. Dit: Self-supervised pre-training for document image transformer. In *Proceedings of the 30th ACM international conference on multimedia*, pp. 3530–3539, 2022.

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Marx, V. Method of the year: spatially resolved transcriptomics. *Nature Methods*, 18(1):9–14, Jan 2021. ISSN 1548-7105. doi: 10.1038/s41592-020-01033-y. URL https://doi.org/10.1038/s41592-020-01033-y.

Peng, Q., Zhou, P., and Li, T. stvcr: spatiotemporal dynamics of single cells. *Nature Methods*, pp. 1–12, 2026.

Sakalyan, K., Palma, A., Guerranti, F., Theis, F., and Günnemann, S. Modeling microenvironment trajectories on spatial transcriptomics with nicheflow. In Belgrave, D., Zhang, C., Lin, H., Pascanu, R., Koniusz, P., Ghassemi, M., and Chen, N. (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 103328–103371. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/954d303839bf7fc458acea7384322255-Paper-Conference.pdf.

Shen, X., Zuo, L., Ye, Z., Yuan, Z., Huang, K., Li, Z., Yu, Q., Zou, X., Wei, X., Xu, P., et al. Inferring cell trajectories of spatial transcriptomics via optimal transport analysis. *Cell systems*, 16(2), 2025.

Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J. O., Huss, M., Mollbrink, A., Linnarsson, S., Codeluppi, S., Åke Borg, Pontén, F., Costea, P. I., Sahlén, P., Mulder, J., Bergmann, O., Lundeberg, J., and Frisén, J. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, 2016. doi: 10.1126/science.aaf2403. URL https://www.science.org/doi/abs/10.1126/science.aaf2403.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568(127063):127063, February 2024.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks, 2017. URL https://arxiv.org/abs/1703.01365.

Tong, A., Fatras, K., Malkin, N., Huguet, G., Zhang, Y., Rector-Brooks, J., Wolf, G., and Bengio, Y. Improving and generalizing flow-based generative models with minibatch optimal transport. *arXiv preprint arXiv:2302.00482*, 2023.

Wei, X., Fu, S., Li, H., Liu, Y., Wang, S., Feng, W., Yang, Y., Liu, X., Zeng, Y.-Y., Cheng, M., Lai, Y., Qiu, X., Wu, L., Zhang, N., Jiang, Y., Xu, J., Su, X., Peng, C., Han, L., Lou, W. P.-K., Liu, C., Yuan, Y., Ma, K., Yang, T., Pan, X., Gao, S., Chen, A., Esteban, M. A., Yang, H., Wang, J., Fan, G., Liu, L., Chen, L., Xu, X., Fei, J.-F., and Gu, Y. Single-cell stereo-seq reveals induced progenitor cells involved in axolotl brain regeneration. *Science*, 377(6610):eabp9444, 2022. doi: 10.1126/science.abp9444. URL https://www.science.org/doi/abs/10.1126/science.abp9444.

APPENDIX

A. Extended Results

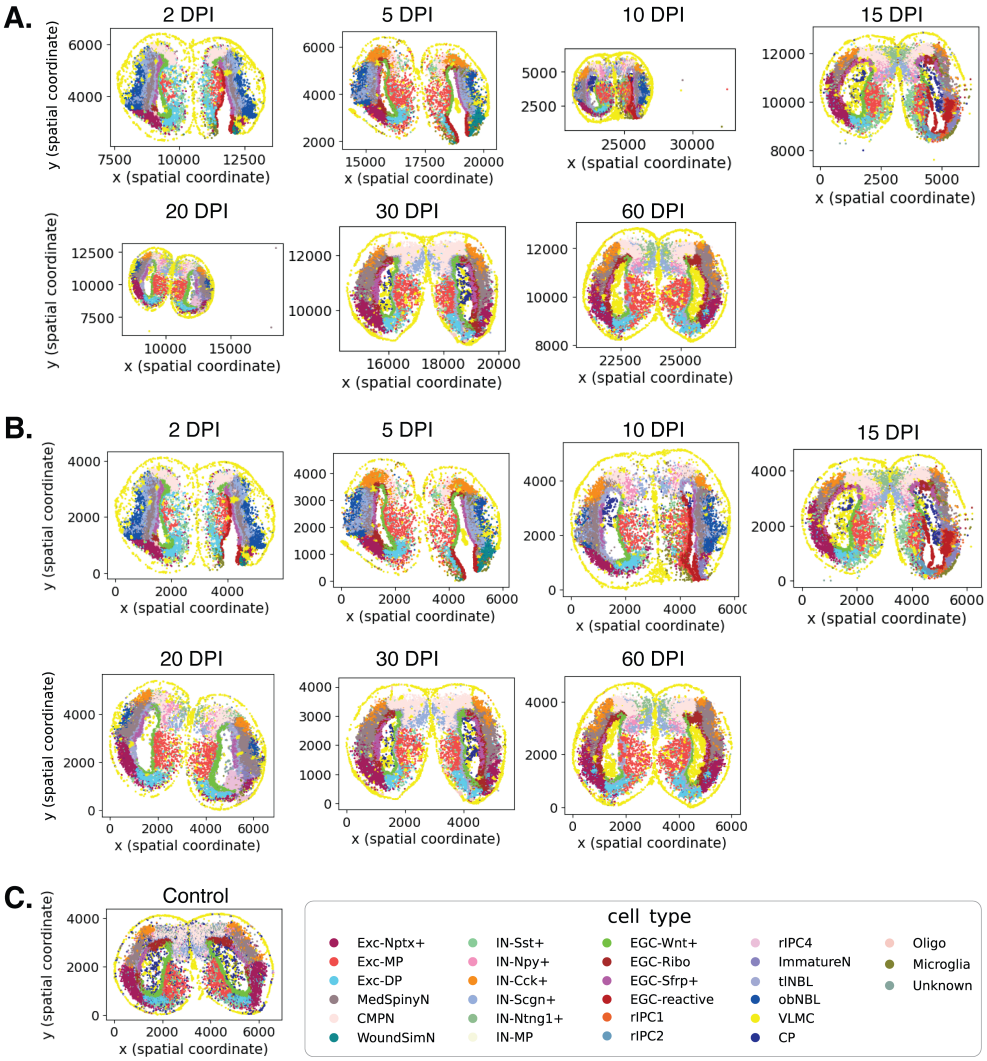

Figure 5. **Visualization of data before and after preprocessing**. **A.** We visualize the original data by plotting the spatial coordinates of cells from each slice and coloring cells based on cell-type identity. The original dataset contains tissue slices with cells that are dissociated from the slice and are located in "outlier" spatial locations, especially clearly visible in the 10DPI, 15DPI and 20DPI slices (DPI: Days Post-Injury). **B.** We manually remove cells with "outlier" spatial coordinates based on visual judgment and work with the displayed pre-processed data. **C.** Visualizing the "control" slice, used to define a reference for the integrated gradients analysis. This control slice is profiled by using the same region in a healthy axolotl telencephalon without any intervention or injury.**Cell type legend:** "Exc": excitatory neurons, "MP": medial pallium region, "DP": dorsal pallium region, "MedSpinyN": medium spiny neurons, "CMPN": cholinergic, monoaminergic, and peptidergic neuron, "WoundSimN": wound-stimulated neuron, "IN": inhibitory neurons, "EGC": ependymoglial cells, "rIPC": regeneration intermediate progenitor cells, "ImmatureN": immature neurons, "tlNBL": telencephalon neuroblasts, "obNBL": olfactory bulb neuroblasts, "VLMC": vascular leptomeningeal cells, "CP": choroid plexus, "Oligo": oligodendrocytes.

B. Data-Driven Niche Discovery with an Encoder-Decoder Transformer

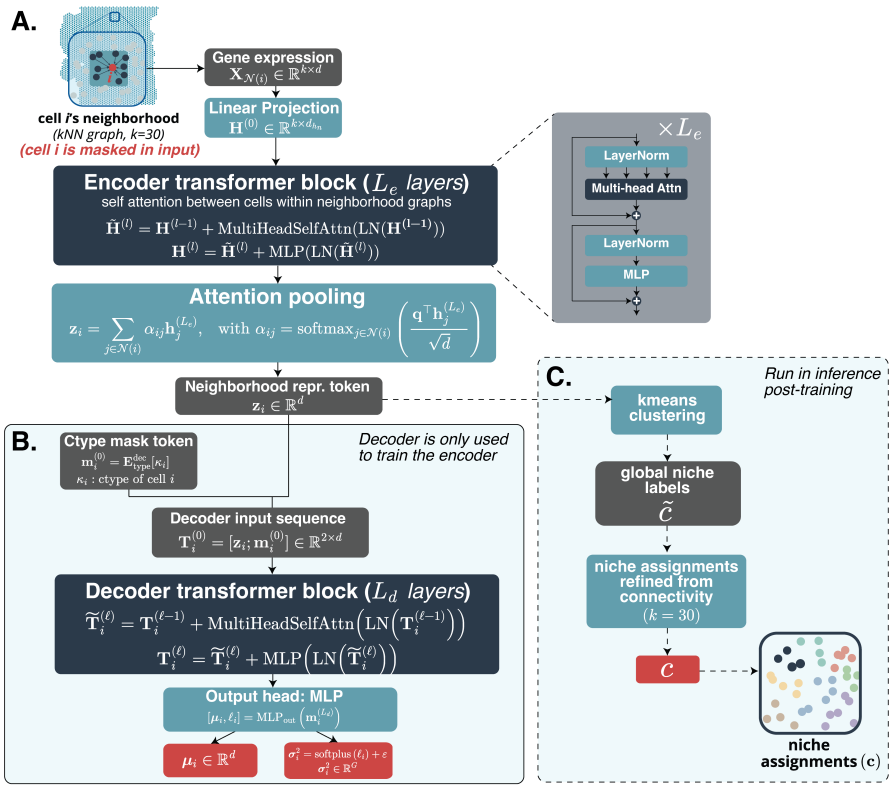

Figure 6. **Architecture diagram for the encoder–decoder Transformer used for computing niche assignments in ChronoTILE . A. Encoder and neighborhood representation learning.** For each reference cell i, we extract its spatial k-nearest-neighbor neighborhood $\mathcal{N}(i)$ with a default of $k = 30$ used in this study, mask the reference cell in the input, and encode only the log-normalized gene expression profiles of the observed neighboring cells. The neighborhood expression matrix with d genes is linearly projected into hidden cell tokens $\mathbf{H}^{(0)}$. The architecture is based on CellTransformer (Lee et al., 2025a); however, we do not use register tokens and do not include cell-type embeddings in the encoder. The encoder applies L_e-many Transformer layers (default of 4 is used here) with self-attention among cells in the neighborhood graph. A learned attention-pooling operation then aggregates the encoded neighboring-cell tokens into a single neighborhood representation token $\mathbf{z}_i$.**B. Decoder used for self-supervised encoder training.** The decoder is only used for training the encoder. It receives the neighborhood representation token $\mathbf{z}_i$ and a cell-type-specific mask token $\mathbf{m}_i^{(0)} = \mathbf{E}_{\text{type}}^{\text{dec}}[\kappa_i]$, where κ_i denotes the cell type of the masked reference cell. These two tokens are concatenated into the decoder input sequence. A shallow decoder Transformer performs self-attention between the neighborhood token and the cell-type mask token, and an output MLP maps the final mask-token representation to Gaussian reconstruction parameters $[\boldsymbol{\mu}_i, \boldsymbol{\ell}_i]$, with $\boldsymbol{\sigma}_i^2 = \text{softplus}(\boldsymbol{\ell}_i) + \varepsilon$. The model is trained to reconstruct the masked log-normalized expression profile using a Gaussian negative log-likelihood loss, $\mathcal{L}_{\text{G-NLL}}(\theta) = -\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \log \mathcal{N}\left(\mathbf{x}_i; \boldsymbol{\mu}_i, \text{diag}(\boldsymbol{\sigma}_i^2)\right)$, where $\boldsymbol{\sigma}_i^2 = \text{softplus}(\boldsymbol{\ell}_i) + \varepsilon$, and $\mathbf{x}_i \in \mathbb{R}^d$ is the masked log-normalized expression profile of the reference cell. **C. Post-training niche assignment.** After training, the decoder is discarded and the encoder plus attention-pooling module is run in inference mode to compute neighborhood representation tokens $\mathbf{z}_i$ for all cells. These representations are clustered with k-means to obtain global initial niche labels $\tilde{\mathbf{c}}$, where the number of clusters k is determined through the elbow method. We then refine these labels using spatial connectivity based on the cell neighborhood graph, assigning disconnected tissue regions with the same initial cluster label to distinct final niche labels $\mathbf{c}$. The resulting niche assignments are passed to the multi-scale velocity-field Transformer in Figure 7.

C. Multi-Scale Transformer Architecture for the Velocity Field

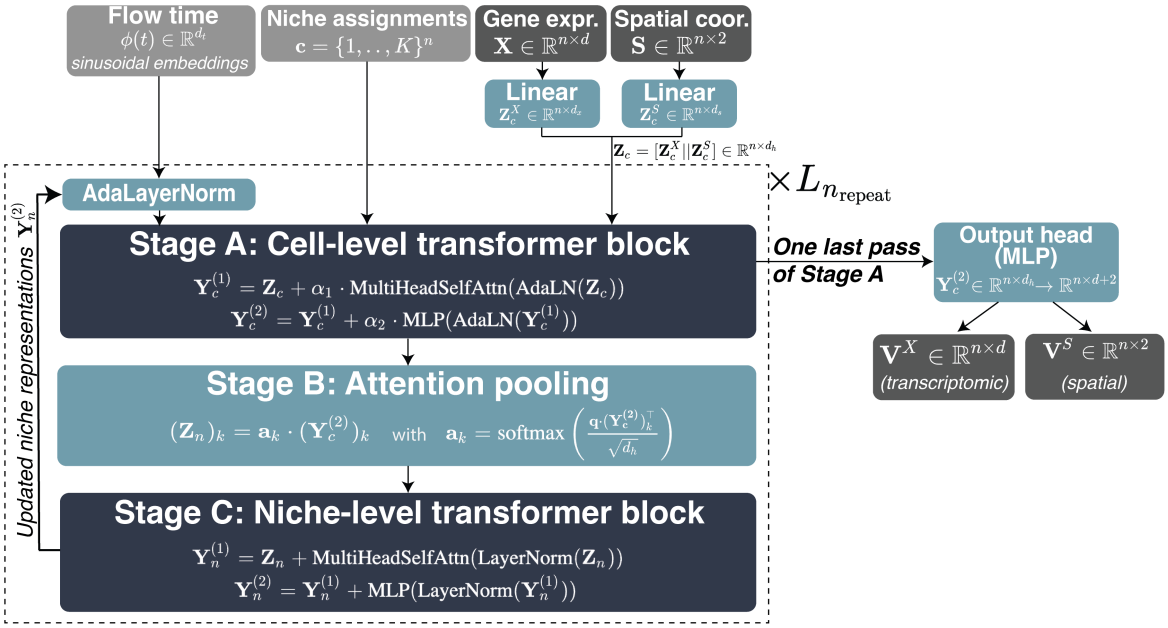

Figure 7. **Architecture diagram for the multi-scale Transformer parameterizing the velocity field in ChronoTILE .** The network parameterizes the time-dependent velocity field $v_\theta(\mathbf{X}_t, \mathbf{S}_t, \mathbf{C}, \phi(t))$, which predicts the instantaneous transcriptomic and spatial velocities of all cells at flow time t. The inputs are the gene-expression matrix $\mathbf{X}_t \in \mathbb{R}^{n \times d}$, spatial coordinate matrix $\mathbf{S}_t \in \mathbb{R}^{n \times 2}$, refined niche assignments $\mathbf{C} \in \{1, \ldots, K\}^n$, and a sinusoidal embedding $\phi(t)$ of biological flow time, with experimental time normalized to $[0, 1]$. Gene expression and spatial coordinates are separately projected into hidden representations $\mathbf{Z}_c^X \in \mathbb{R}^{n \times d_x}$ and $\mathbf{Z}_c^S \in \mathbb{R}^{n \times d_s}$, which are concatenated to form cell tokens $\mathbf{Z}_c = [\mathbf{Z}_c^X \| \mathbf{Z}_c^S] \in \mathbb{R}^{n \times d_h}$, where $d_h = d_x + d_s$. **Stage A: cell-level Transformer block.** Stage A performs multi-head self-attention among cells within each niche, using the niche assignments $\mathbf{C}$ to define a block-diagonal attention structure. Thus, cells exchange information with other cells assigned to the same niche, capturing local cell–cell interactions and co-expression structure while avoiding global all-to-all attention. The Stage A block is conditioned on flow time and niche context through AdaLayerNorm: the sinusoidal time embedding $\phi(t)$ and the updated representation of each cell's assigned niche are used to modulate the cell-level Transformer computation. This allows cell-state updates to depend jointly on biological time and the current niche-level context. **Stage B: attention pooling from cells to niches.** Stage B aggregates cell-level representations into niche-level tokens. For each niche k, a learned query vector computes attention weights over the Stage A cell tokens assigned to that niche, producing a niche representation $(\mathbf{Z}_n)_k$. This operation compresses the variable number of cells in each niche into a fixed-dimensional token summarizing that niche's current transcriptomic and spatial state. **Stage C: niche-level Transformer block.** Stage C performs multi-head self-attention across all niche tokens, allowing communication between niches and capturing tissue-scale coordination, such as longer-range signaling gradients or boundary effects between neighboring microenvironments. The updated niche representations $\mathbf{Y}_n^{(2)}$ are fed back into the AdaLayerNorm conditioning of Stage A in the next repeat, so that subsequent cell-level updates are informed by progressively refined niche-level context. The Stage A–B–C module is repeated n_{repeat} times, followed by one final Stage A pass to produce cell-level output representations. A final output MLP maps these representations to the joint velocity field in $\mathbb{R}^{n \times (d+2)}$, which is split into transcriptomic velocity $\mathbf{v}_\theta^X \in \mathbb{R}^{n \times d}$ and spatial velocity $\mathbf{v}_\theta^S \in \mathbb{R}^{n \times 2}$. These velocity components are trained with the modality-balanced multi-marginal flow matching objective, which regresses the predicted velocities onto analytic reference velocities derived from spline-based interpolation through multi-marginal chains of cells.

D. Evaluation Metrics

Our goal requires capturing dynamics of both gene expression changes and the spatial architecture. As such, our evaluation metrics aim to cover two main areas: (1) fidelity of gene expression profile reconstruction, and (2) fidelity of spatial architecture prediction. We evaluate our model by holding out slices from 10 DPI (DPI: days post-injury) and 30 DPI, one at a time, training on the remaining time points, and generating tissue at t_k by integrating the learned velocity field from the first observed time point: 2DPI. We then compare the generated tissue against the held-out real slice using the metrics described below. For each metric, we indicate in parentheses whether lower (↓) or higher (↑) values are better, along with the range of possible values.

Notation Let $\mathcal{D}_{t_k} = (\mathbf{X}_{t_k}, \mathbf{S}_{t_k})$ denote the real (held-out) slide with n cells, and $\hat{\mathcal{D}}_{t_k} = (\hat{\mathbf{X}}_{t_k}, \hat{\mathbf{S}}_{t_k})$ denote the generated slide with m cells, where $\mathbf{X} \in \mathbb{R}^{n\times g}$ and $\hat{\mathbf{X}} \in \mathbb{R}^{m\times g}$ are gene expression matrices and $\mathbf{S} \in \mathbb{R}^{n\times d}$ and $\hat{\mathbf{S}} \in \mathbb{R}^{m\times d}$ are spatial coordinate matrices. We denote the i-th row of $\mathbf{X}$ and $\mathbf{S}$ by $\mathbf{x}_i \in \mathbb{R}^g$ and $\mathbf{s}_i \in \mathbb{R}^d$, respectively, and similarly $\hat{\mathbf{x}}_j$ and $\hat{\mathbf{s}}_j$ for the generated slide. Since the measurement process is destructive, there is no cell-level correspondence between $\mathcal{D}_{t_k}$ and $\hat{\mathcal{D}}_{t_k}$; all metrics below compare distributions rather than paired cells.

D.1. Gene Expression Reconstruction Fidelity

These metrics evaluate whether the generated tissue contains the correct distribution of gene expression profiles, irrespective of where those profiles are located in space.

- **Wasserstein-2 (W_2) Distance (↓, $[0,\infty)$).** Measures the minimum "cost" of transporting the generated expression distribution onto the real expression distribution, providing a geometrically meaningful comparison that accounts for the metric structure of gene expression space.

  Given empirical gene expression distributions over n real cells and m generated cells, we first compute a joint PCA on the concatenated expression matrices $[\mathbf{X}_{t_k}; \hat{\mathbf{X}}_{t_k}] \in \mathbb{R}^{(n+m)\times g}$ and retain the top P principal components ($P = 50$), yielding projected representations $\mathbf{z}_i, \hat{\mathbf{z}}_j \in \mathbb{R}^P$ for real and generated cells, respectively. This avoids the curse of dimensionality in the full g-dimensional gene space, where Euclidean distances lose discriminative power. The discrete 2-Wasserstein distance is then:

  $$W_2^2(\mathcal{D}_{t_k}, \hat{\mathcal{D}}_{t_k}) = \min_{\mathbf{T}\in\Pi(\mathbf{a},\mathbf{b})} \sum_{i=1}^{n}\sum_{j=1}^{m} T_{ij}\,\|\mathbf{z}_i - \hat{\mathbf{z}}_j\|_2^2, \tag{14}$$

  where $\mathbf{T} \in \mathbb{R}^{n\times m}$ is a transport plan with row sums $\mathbf{a} = (\frac{1}{n}, \ldots, \frac{1}{n})^T$ and column sums $\mathbf{b} = (\frac{1}{m}, \ldots, \frac{1}{m})^T$.

- **Maximum Mean Discrepancy (MMD) (↓, $[0,\infty)$).** Measures whether the overall distribution of gene expression profiles in the generated tissue matches that of the real tissue, using a kernel-based comparison that captures differences in all moments of the distributions.

  Given two sets of PCA-projected gene expression profiles $\{\mathbf{z}_i\}_{i=1}^n$ from the real slide and $\{\hat{\mathbf{z}}_j\}_{j=1}^m$ from the generated slide (using the same top-50 PCA projection as W_2 above), MMD is defined as:

  $$\mathrm{MMD}^2(\mathcal{D}_{t_k}, \hat{\mathcal{D}}_{t_k}) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{i'=1}^{n} k(\mathbf{z}_i, \mathbf{z}_{i'}) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} k(\mathbf{z}_i, \hat{\mathbf{z}}_j) + \frac{1}{m^2}\sum_{j=1}^{m}\sum_{j'=1}^{m} k(\hat{\mathbf{z}}_j, \hat{\mathbf{z}}_{j'}), \tag{15}$$

  where $k(\cdot,\cdot)$ is a Gaussian RBF kernel $k(\mathbf{z}, \mathbf{z}') = \exp\left(-\|\mathbf{z} - \mathbf{z}'\|^2/2\sigma^2\right)$. The bandwidth σ is set via the *median heuristic*: we compute all pairwise Euclidean distances between points in the pooled set $\{\mathbf{z}_1, \ldots, \mathbf{z}_n\} \cup \{\hat{\mathbf{z}}_1, \ldots, \hat{\mathbf{z}}_m\}$ and set σ to the median of these distances, ensuring the kernel operates at a scale commensurate with the data. An MMD^2 of zero indicates that the two distributions are identical in the reproducing kernel Hilbert space induced by k.

D.2. Spatial Architecture Fidelity

These metrics evaluate whether the model populates the correct spatial locations with cells, regardless of what those cells express or what cell types they are.

- **Point-to-Shape Distance (PSD) (↓, $[0,\infty)$).** Measures whether generated cells are placed in spatially plausible locations by computing how far each generated cell is from the nearest real cell (a precision-like measure over spatial coordinates). This metric was first introduced in *NicheFlow* (Sakalyan et al., 2025).

  $$\mathrm{PSD}(\hat{\mathcal{D}}_{t_k}, \mathcal{D}_{t_k}) = \frac{1}{m}\sum_{j=1}^{m} \min_{i\in\{1,\ldots,n\}} \|\hat{\mathbf{s}}_j - \mathbf{s}_i\|_2^2. \tag{16}$$

  A low PSD indicates that every generated cell has a nearby counterpart in the real tissue; a high PSD indicates the model is hallucinating cells in regions where no cells actually exist.

- **Shape-to-Point Distance (SPD)** ($\downarrow$, $[0, \infty)$)**.** Measures whether the model covers the full spatial extent of the real tissue by computing how far each real cell is from the nearest generated cell (a recall-like measure over spatial coordinates). This metric was first introduced in *NicheFlow* (Sakalyan et al., 2025).

$$\mathrm{SPD}(\mathcal{D}_{t_k}, \hat{\mathcal{D}}_{t_k}) = \frac{1}{n} \sum_{i=1}^{n} \min_{j \in \{1,\dots,m\}} \|\mathbf{s}_i - \hat{\mathbf{s}}_j\|_2^2. \tag{17}$$

  A low SPD indicates that every region of the real tissue has generated cells nearby; a high SPD indicates the model fails to populate certain tissue regions. PSD and SPD are complementary: a model that generates cells in only a small correct region achieves low PSD but high SPD, while a model that scatters cells everywhere achieves low SPD but high PSD.