TEMPUS: TEMPORAL MODELING PLATFORM FOR UNIFIED SPATIAL OMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Tissues evolve over time through development, homeostasis, and disease, and understanding these dynamics requires measurements that are spatially resolved and temporally informative. Spatial transcriptomic studies increasingly profile tissues at multiple time points alongside hematoxylin and eosin (H&E) histology, but the cost of the two modalities is disproportionate: H&E is acquired routinely while spatial transcriptomics remains expensive and is profiled at only a few time points. We formulate *tissue dynamics prediction* as a multi-modal cross-time task: predict spatial gene expression at any time point given dense H&E images across all time points and sparse spatial transcriptomics at few. We introduce **TEMPUS** (Temporal Modeling Platform for Unified Spatial Omics), the first model for this task. TEMPUS combines three design choices: (i) a learned per-query top-K selection over a two-dimensional grid of past observations indexed by time point and tissue coordinate, compressed into one per-query, per-time-point token consumed by a cross-attention decoder; (ii) a chained cross-modal and temporal objective with cycle-consistency, where an auxiliary decoder reconstructs each pre-fusion history token from the prediction-time latent, preventing collapse onto the most recent observation; and (iii) operation in frozen foundation-model embedding space, decoded to counts only at evaluation, which keeps the model gene-panel and platform agnostic and scales to whole-tissue in Xenium resolution. On two synthetic benchmarks where the target provably cannot be predicted from any single time point, and four experimental datasets, the mean Pearson correlation improves over the strongest baselines from non-temporal histology-to-expression methods and temporal models. Ablations validate that learned top-K outperforms k-nearest-neighbor retrieval, and that per-query attention is interpretable against ground-truth niche identity.

1 INTRODUCTION

Cells and their interactions with surrounding tissue change over time. Spatial transcriptomics (ST) (Ståhl et al., 2016; Moses & Pachter, 2022) measures not only which genes a cell expresses but also where it is located in the tissue and which neighbors it is exposed to, capturing the microenvironmental context that defines cellular state (Zhou et al., 2024; Sakalyan et al., 2026). When such measurements are combined across multiple time points, developmental stages, disease progression and treatment conditions (Ortiz et al., 2020; Chelebian et al., 2025), they yield a far richer picture: how individual cells, their molecular programs, and their spatial neighbourhoods co-evolve through development and disease (Marx, 2021; Longo et al., 2021; Zhang et al., 2025). These changes are shaped by the internal state of a cell and its microenvironment, and, in turn, influence its underlying morphology and molecular profile. Therefore, this necessitates the study of tissue dynamics jointly across spatio-temporal axes.

In practice, however, we cannot collect such measurements densely along the time axis. Spatial transcriptomics is expensive and laborious to acquire, requiring specialized reagents, instrumentation, and protocols that limit how many tissues a study can profile (Jaume et al., 2024). In contrast, Hematoxylin and Eosin (H&E) histology, the standard stain in clinical pathology, sits at the opposite extreme: it captures detailed tissue morphology cheaply, at scale, and is already collected routinely on essentially every tissue of interest. The result is a structural asymmetry across the temporal axis of any developmental or disease study: paired ST is available at a few sparse time points, while

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.

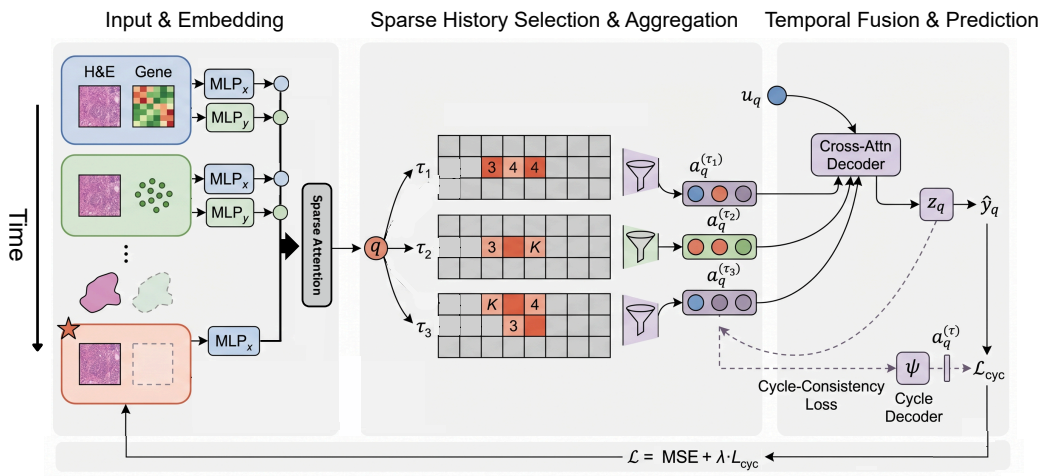

Figure 1: **Overview of TEMPUS. Left:** Each history time point provides paired H&E and gene expression embeddings via modality-specific projections with spatial and relative temporal positional encodings. **Centre:** For each target query q, the top-K most relevant entries (orange) are selected independently at every history time step based on learned relevance scores, then compressed by a shared aggregator into one token. **Right:** A cross-attention decoder fuses the aggregated history tokens with the target query to produce the prediction. A cycle-consistency decoder regularising the model to retain information from all time steps.

H&E is available at all of them. This has motivated an active line of research on predicting spatial gene expression from histology alone (He et al., 2020; Xie et al., 2023; Zeng et al., 2022; Li et al., 2024; Chung et al., 2024; Ganguly et al., 2025; Fu et al., 2025; Huang et al., 2025; Pan et al., 2025; Chen et al., 2024), aiming to extract molecular information from images that are already routinely collected. Yet these approaches cannot model tissue dynamics; they map a single histology image to its paired expression, one tissue at a time, with no mechanism to propagate information across time points (Wang et al., 2025; Jaume et al., 2024).

We argue that the right framing for this problem is not single-image expression prediction but predicting tissue dynamics: given dense H&E across many time points and ST at only a sparse subset, predict the spatial gene-expression state of the tissue at any time point, including those where only histology is available. This requires a model that can ingest a cheap imaging modality (H&E) at the target time and infer molecular information from the few profiled time points to the unprofiled ones, accounting for cellular dynamics and how they are shaping their neighborhoods and their morphology (Sakalyan et al., 2026; Zhou et al., 2024; Zhu et al., 2025; Sanian et al., 2025).

Doing so requires addressing five intertwined properties of the data that no current method models jointly. First, the data is **multi-modal**: histology and expression carry complementary but structurally distinct information, and the mapping between them is complex and many-to-many (Xie et al., 2023). Second, the data has **temporal structure**: observations at adjacent time points are not independent but reflect gradually changing tissue states, and each profiled time point is highly predictive of its neighbors (Monjo et al., 2022). Third, the data is **temporally sparse**: paired ST is only available for a subset of time points due to cost constraints, and within each profiled tissue transcript dropout may further reduce coverage (Ruiz et al., 2025; Wang et al., 2025). Fourth, the data **scales simultaneously in two directions at once**: gene panels routinely cover thousands of genes, and a single high-resolution Xenium tissue can contain hundreds of thousands of cells, so methods that train end-to-end against the raw count tensor quickly become memory-bound and cannot scale to whole-tissue inputs (Jaume et al., 2024). Fifth, the data is **unpaired across time**: tissue sectioning is destructive, so the same cell cannot be tracked across time points; at each time point we observe a different population sampled from a related but distinct tissue, and any cross-time correspondence must be inferred rather than measured.

We introduce TEMPUS, a framework that addresses all five properties jointly. TEMPUS has two core components: (i) a **chained training objective** whose forward arm jointly performs cross-

modal mapping and temporal prediction from histology at the target time conditioned on observed expression–histology pairs, while a backward arm enforces cycle consistency by reconstructing each history token from the fused prediction, coupling the multi-modal and temporal structure in a single loss; and (ii) a **top-K history selection**, inspired by the latent bottleneck of Perceiver IO (Jaegle et al., 2021) but operating per query rather than through a fixed set of latent vectors, which scores all entries in a time-by-position buffer, retains the K most relevant per query and time step, and compresses them into a single aggregated token for downstream cross-attention, preserving cell-specific spatial structure.

Our contributions are as follows:

- We introduce TEMPUS, a scalable and interpretable multi-modal architecture for predicting spatially resolved gene expression from H&E histology at unseen time points.
- We design a learned top-K history selection mechanism over a two-dimensional grid of past observations that allows each target cell to independently attend to the most informative cells and time points in its spatiotemporal neighbourhood. This design addresses both the temporal sparsity of the available observations and the spatial scalability challenge as tissue size and resolution grow.
- We release *Phase Memory* and *Trajectory Memory*, two synthetic benchmarks constructed so that the held-out target provably cannot be inferred from any single time point, providing a clean diagnostic for whether a model genuinely uses cross-time information.
- We evaluate on six datasets, two synthetic and four real, with the latter spanning fetal heart and pancreas tissue on both Visium and Xenium platforms. TEMPUS improves mean Pearson correlation over the strongest single-tissue and temporal baselines on every dataset, and ablations isolate the contribution of each architectural component.

2 METHOD

We introduce TEMPUS, an architecture for predicting tissue dynamics: at any time point and any spatial location, predict spatial gene expression from histology together with a temporally sparse set of past expression observations. The method operates on a 2D point cloud that evolves across ordered time points, where each point carries two co-localized modalities: a histology embedding $\mathbf{x}$ available at every time point, and an expression embedding $\mathbf{y}$ observed only at a subset of time points and predicted at the rest.[1] We address two prediction regimes within this setting: **cross-time completion**, where expression at a held-out time point is predicted from its histology together with earlier observed pairs, and **within-slide imputation**, where a spatially held-out region of the target time point is predicted from the rest of the same slide together with earlier time points. The architecture has four stages: (i) modality-specific projections map each point into a shared space with spatial and temporal positional encodings (§2.2); (ii) a sparse history selection module identifies the most relevant history points at each past time point (§2.3); (iii) selected entries are compressed into one latent token per query and per time point, then fused across time by a cross-attention decoder (§2.4); and (iv) a cycle-consistency loss prevents temporal collapse by requiring the fused representation to reconstruct each pre-fusion history token (§2.5). An overview is given in Figure 1.

2.1 PROBLEM FORMULATION

Let $t \in \{1, \dots, T\}$ index ordered time points and let $\mathcal{D}^{(t)} = \{(\mathbf{p}_i, \mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N_t}$ denote the data at time t, where $\mathbf{p}_i \in \mathbb{R}^2$ is a spatial coordinate, $\mathbf{x}_i \in \mathbb{R}^{d_x}$ is the input embedding, and $\mathbf{y}_i \in \mathbb{R}^{d_y}$ is the content embedding. Training and prediction operate entirely in embedding space; count-space evaluation, where reported, uses a frozen decoder defined in §2.7.

At training time we sample a target time t and construct a history set from the remaining observed times $\mathcal{T}_{\text{hist}} \subset \{1, \dots, t-1\}$. For each $\tau \in \mathcal{T}_{\text{hist}}$, a spatial crop of radius r_h centred on the target location yields:

$$\mathcal{H}_t = \left\{ \mathcal{S}^{(\tau)} = \left(X^{(\tau)}, Y^{(\tau)}, P^{(\tau)}, \tau\right) \;:\; \tau \in \mathcal{T}_{\text{hist}} \right\}, \tag{1}$$

[1] The formalism is general: $\mathbf{x}$ can be any cheaply-acquired modality available at every time point, and $\mathbf{y}$ any expensive, harder to obtain modality observed sparsely. We use histology and gene expression throughout for concreteness.

where $X^{(\tau)} \in \mathbb{R}^{N_\tau \times d_x}$, $Y^{(\tau)} \in \mathbb{R}^{N_\tau \times d_y}$, and $P^{(\tau)} \in \mathbb{R}^{N_\tau \times 2}$ stack the cropped points. Restricting each history step to a local spatial neighbourhood bounds the cost of the subsequent top-K selection and biases context toward spatial proximity. For in-slide imputation, a same-time context step $\mathcal{S}^{(t,0)}$ containing the non-held-out points at time t is prepended to $\mathcal{H}_t$ with relative offset $\Delta\tau = 0$.

The output is a predicted content embedding $\hat{\mathbf{y}}_q$ for every query point q in the held-out set.

2.2 Modality Embeddings

Each point is mapped into a shared d-dimensional space through two modality-specific projections:

$$\mathbf{u}_i = \mathrm{MLP}_x(\mathbf{x}_i) + \mathrm{PE}_{\mathrm{sp}}(\mathbf{p}_i), \qquad \mathbf{v}_i = \mathrm{MLP}_y(\mathbf{y}_i), \tag{2}$$

where MLP_x and MLP_y are two-layer networks. The token $\mathbf{u}_i$ combines input with spatial identity and supplies the query and key in attention; values come only from history points and are the concatenation $[\mathbf{u}_i;\, \mathbf{v}_i]$. At the target time only $\mathbf{u}_q$ is built, since $\mathbf{y}$ is the prediction target.

Encoding. Each point receives a spatial and a temporal encoding alongside its modality projections. Spatial coordinates $\mathbf{p}_i$ are normalised and embedded by Fourier features at geometric frequencies, $\mathrm{PE}_{\mathrm{sp}}$ (Appendix C). The offset between each history step and the target is encoded by a sinusoidal embedding $\boldsymbol{\gamma}(\Delta\tau)$ of the relative time $\Delta\tau = t - \tau$, followed by a learned linear projection. Using relative rather than absolute time keeps the feature distribution stationary across train/test splits.

2.3 Sparse History Selection

Each history step $\mathcal{S}^{(\tau)}$ may contain thousands of points after spatial cropping. Attending to all of them at every query is both unnecessary and expensive: only a small subset carries information relevant to a given target location. We address this through a per-query top-K selection that identifies the K most informative points within each time step independently.

Beyond efficiency, for a query (target cell) at the held-out time we can directly interpret the model's prediction: the top-K set $\mathcal{I}_q^{(\tau)}$ is a small, discrete list of past cells the model *actually* used, in contrast to dense softmax attention, where retrieval evidence is diffused across all N_τ history entries. On the Trajectory Memory dataset, where the ground-truth informative history is provably defined, this discrete set tracks the correct same-niche cells through time on the queries the model predicts well, and locks onto the wrong niche precisely on the queries it predicts poorly (§3.5, Appendix J).

For target query q and history step τ, relevance scores are computed via scaled dot-product attention with normalised queries and keys:

$$\mathbf{Q}_q^{(\tau)} = \mathrm{LN}\big(\mathbf{W}^Q(\mathbf{u}_q + \boldsymbol{\gamma}(\Delta\tau))\big), \qquad \mathbf{K}_i^{(\tau)} = \mathrm{LN}\big(\mathbf{W}^K\mathbf{u}_i\big), \tag{3}$$

$$s_{q,i}^{(\tau)} = \frac{{\mathbf{Q}_q^{(\tau)}}^\top \mathbf{K}_i^{(\tau)}}{\sqrt{d_h}\,/\,\sigma_h}, \tag{4}$$

where σ_h is a per-head learnable inverse temperature and LayerNorm on both queries and keys prevents entropy collapse. The top-K indices are selected as $\mathcal{I}_q^{(\tau)} = \mathrm{argtopK}\big(\{s_{q,i}^{(\tau)}\}_{i=1}^{N_\tau}\big)$, and the restricted softmax aggregates the corresponding values into a single token:

$$\boldsymbol{\alpha}_q^{(\tau)} = \mathrm{softmax}\big(\{s_{q,i}^{(\tau)}\}_{i\in\mathcal{I}_q^{(\tau)}}\big), \qquad \mathbf{a}_q^{(\tau)} = \mathrm{Agg}\Big(\sum_{i\in\mathcal{I}_q^{(\tau)}} \alpha_{q,i}^{(\tau)}\,[\mathbf{u}_i;\, \mathbf{v}_i]\Big) + \boldsymbol{\gamma}(\Delta\tau), \tag{5}$$

where Agg is a shared MLP applied identically at every time step. The relative time embedding is added after aggregation so that all per-step tokens occupy a comparable basis before temporal fusion. This design compresses each history time step into a single token per query, reducing the downstream fusion from attention over $|\mathcal{T}_{\mathrm{hist}}| \cdot K$ entries to attention over $|\mathcal{T}_{\mathrm{hist}}|$ tokens. Since argtopK is non-differentiable through the index, we use a straight-through estimator: the forward pass selects hard top-K indices, while the backward pass routes gradients through a softmax over all N_τ keys weighted by the same scores.

Table 1: Toy-benchmark results on Phase Memory (PM) and Trajectory Memory (TM). Mean ± std over 3 seeds, averaged across held-out timepoints.

	Phase Memory (PM)			**Trajectory Memory** (TM)		
Method	PCC↑	MSE↓	MAE↓	PCC↑	MSE↓	MAE↓
Single-section H&E to expression (no temporal memory)						
Hist2ST (Zeng et al., 2022)	0.515 ± 0.078	0.366 ± 0.063	0.472 ± 0.041	0.236 ± 0.186	0.156 ± 0.045	0.279 ± 0.034
BLEEP (Xie et al., 2023)	0.467 ± 0.077	0.442 ± 0.084	0.492 ± 0.044	0.452 ± 0.230	0.104 ± 0.040	0.225 ± 0.045
MERGE (Ganguly et al., 2025)	0.503 ± 0.052	0.385 ± 0.052	0.484 ± 0.032	0.231 ± 0.437	0.118 ± 0.046	0.253 ± 0.058
STFlow (Huang et al., 2025)	0.520 ± 0.070	0.364 ± 0.065	0.468 ± 0.040	0.522 ± 0.169	0.099 ± 0.024	0.224 ± 0.027
TEMPUS (no memory)	0.527 ± 0.080	0.367 ± 0.058	0.468 ± 0.037	0.310 ± 0.357	0.106 ± 0.061	0.235 ± 0.081
Temporal models (use cross-time information)						
NicheFlow (Sakalyan et al., 2026)	0.633 ± 0.066	0.341 ± 0.055	0.395 ± 0.019	0.488 ± 0.068	0.102 ± 0.083	0.241 ± 0.062
NicheFlow+X	0.730 ± 0.016	0.327 ± 0.041	0.311 ± 0.016	0.534 ± 0.166	0.081 ± 0.038	0.206 ± 0.023
TEMPUS (Ours)	$\mathbf{0.871 \pm 0.011}$	$\mathbf{0.141 \pm 0.039}$	$\mathbf{0.215 \pm 0.025}$	$\mathbf{0.883 \pm 0.025}$	$\mathbf{0.023 \pm 0.008}$	$\mathbf{0.117 \pm 0.021}$

2.4 Temporal Fusion

The sparse selection stage produces one aggregated token $\mathbf{a}_q^{(\tau)}$ per query per history time step a short sequence that must be fused into a single representation capturing the most relevant temporal context.

Each target query cross-attends to its own per-step token sequence through an L_c-layer transformer decoder:

$$\mathbf{z}_q = \mathrm{CrossAttnDecoder}\big(\mathbf{u}_q,\ \{\mathbf{a}_q^{(\tau)}\}_{\tau \in \mathcal{H}_t}\big). \tag{6}$$

Since the same-time context step (when present) appears in this sequence with $\Delta\tau = 0$, within-section spatial information and cross-time temporal information are fused through the same mechanism. After temporal fusion, an optional self-attention stack allows target queries within the same spatial crop to exchange information. A prediction head maps the fused representation to the content embedding: $\hat{\mathbf{y}}_q = h(\mathbf{z}_q)$.

2.5 Cycle-Consistency Regularisation

The top-K aggregation followed by temporal fusion is lossy by design: each history step is compressed to a single token, and the model may collapse to relying on only the most recent step. To prevent this and to encourage the fused representation to retain information from all relevant time points, we introduce an auxiliary cycle-consistency objective.

A decoder ψ receives the post-fusion representation $\mathbf{z}_q$ together with a relative time tag $\boldsymbol{\gamma}(\Delta\tau)$ and must reconstruct the corresponding pre-fusion aggregated token:

$$\mathcal{L}_{\mathrm{cyc}} = \frac{1}{|\mathcal{H}_t|} \sum_{\tau} \Big\| \psi\big([\mathbf{z}_q;\, \boldsymbol{\gamma}(\Delta\tau)]\big) - \mathrm{sg}\big(\mathbf{a}_q^{(\tau)}\big) \Big\|_2^2, \tag{7}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient on the target. This asymmetry is deliberate: the prediction loss drives the aggregator and projections, while the cycle loss drives ψ and enriches $\mathbf{z}_q$ by requiring it to preserve per-step information through the fusion bottleneck.

2.6 Training and Inference

Training. The objective combines embedding-space prediction with a cycle-consistency regulariser,

$$\mathcal{L} = \mathrm{MSE}(\hat{\mathbf{y}}, \mathbf{y}^\star) + \lambda_{\mathrm{cyc}}\, \mathcal{L}_{\mathrm{cyc}}, \tag{8}$$

with the MSE masked to the held-out region under within-slide imputation. Both terms act in the embedding space; the model never predicts raw counts during training (Algorithm 1, Appendix A).

Inference. The target section is partitioned into overlapping spatial chunks, each processed independently with its own cropped history, and predictions are averaged over overlapping tiles. For multi-step autoregressive evaluation, content predictions at earlier test time-points replace missing ground truth in the history buffer; under within-slide imputation, the propagated representation mixes predictions for held-out points with ground-truth content elsewhere. The cycle decoder is disabled at inference (Appendix B).

Table 2: Real-data results in count space on three datasets: Human heart, Human pancreas, and a kidney ischemia–reperfusion injury (IRI) time course. Mean ± std over 3 seeds, averaged across held-out timepoints. The complete table including Hist2ST and STFlow is given in Appendix H.4 (Table 11).

	Human heart			**Human pancreas**			**Kidney IRI**		
Method	PCC↑	MSE↓	MAE↓	PCC↑	MSE↓	MAE↓	PCC↑	MSE↓	MAE↓
Single-section H&E to expression (no temporal memory)									
BLEEP	0.671 ± 0.011	4.828 ± 0.101	0.369 ± 0.002	0.805 ± 0.005	17.72 ± 0.20	**0.448 ± 0.004**	0.474 ± 0.010	0.142 ± 0.005	0.180 ± 0.004
MERGE	0.698 ± 0.026	4.681 ± 0.253	0.371 ± 0.008	0.805 ± 0.001	18.92 ± 1.84	0.561 ± 0.024	0.461 ± 0.011	0.147 ± 0.004	0.184 ± 0.005
TEMPUS (no memory)	0.738 ± 0.008	4.218 ± 0.081	0.331 ± 0.001	0.788 ± 0.004	17.05 ± 0.33	0.490 ± 0.022	0.480 ± 0.003	0.132 ± 0.003	0.171 ± 0.003
Temporal models (use cross-time information)									
NicheFlow	0.613 ± 0.037	5.930 ± 0.360	0.365 ± 0.004	0.760 ± 0.008	27.19 ± 1.20	0.638 ± 0.012	0.431 ± 0.013	0.149 ± 0.006	0.193 ± 0.006
NicheFlow+X	0.705 ± 0.003	4.529 ± 0.137	0.353 ± 0.002	0.781 ± 0.003	22.96 ± 1.30	0.594 ± 0.018	0.463 ± 0.014	0.143 ± 0.005	0.185 ± 0.005
TEMPUS (Ours)	**0.794 ± 0.006**	**3.680 ± 0.082**	**0.318 ± 0.002**	**0.812 ± 0.006**	**16.32 ± 0.07**	0.477 ± 0.018	**0.512 ± 0.008**	**0.129 ± 0.003**	**0.165 ± 0.003**

Table 3: In-slide imputation results on Phase Memory (PM), Trajectory Memory (TM) and Human heart. Mean ± std over 3 seeds on the held-out region (90% of cells held out, 10% prompt), averaged across test timepoints. Human heart in count space.

	Phase Memory (PM)			**Trajectory Memory** (TM)			**Human heart**		
Method	PCC↑	MSE↓	MAE↓	PCC↑	MSE↓	MAE↓	PCC↑	MSE↓	MAE↓
PH2ST (Niu et al., 2025)	0.582 ± 0.032	0.301 ± 0.035	0.434 ± 0.027	0.064 ± 0.123	0.175 ± 0.051	0.298 ± 0.038	0.235 ± 0.012	5.292 ± 0.341	0.348 ± 0.009
TEMPUS (Ours)	**0.850 ± 0.013**	**0.154 ± 0.034**	**0.235 ± 0.017**	**0.891 ± 0.189**	**0.030 ± 0.012**	**0.130 ± 0.029**	**0.810 ± 0.001**	**3.210 ± 0.062**	**0.307 ± 0.003**

2.7 COUNT DECODER

TEMPUS predicts in the embedding space. To compare against baselines that emit raw counts and to report tissue-dataset metrics on a common axis, the predicted embedding $\hat{\mathbf{y}}$ is passed through a frozen Negative-Binomial `CountDecoder` parameterised by a softplus mean and a per-gene dispersion. The decoder is fitted once per dataset on (TERRA (Birk et al., 2026) embedding, raw count) pairs from the training split by NB negative log-likelihood, and is never updated jointly with TEMPUS. On synthetic datasets the decoder is not used and metrics are computed in embedding space directly. The full architecture and pmf are in Appendix C.

3 EXPERIMENTS

We evaluate TEMPUS on two synthetic and on four experimental datasets, two public Visium datasets with serial-section structure, and two Xenium datasets spanning multiple time points. The synthetic benchmarks are designed to make the role of cross-time attention unambiguous; the real datasets stress the model along distinct biological axes.

3.1 DATASETS

Synthetic benchmarks. We design two synthetic datasets as a controlled testbed for our two central claims: that the model genuinely uses memory, and that its per-query attention is interpretable. By construction, the data require cross-time information to predict the target, and provide a ground-truth reference against which the selected history can be audited. *Phase Memory (PM)*: a rotating mode-to-content mapping where $\mathbf{x}_t$ reveals the current mode but not the rotation phase, giving a structural ceiling for any memory-less predictor. *Trajectory Memory (TM)*: orthogonal-niche subspaces with a per-niche cumulative target, requiring recency-weighted integration and providing a falsifiable same-niche ground truth for the attention audit (§3.5). Generator equations, biological analogues, and parameter defaults are in Appendix E.

Real datasets. We benchmark on four real-world spatial-transcriptomics datasets, each chosen to stress the model on a distinct biological axis. Human heart (Visium, Lázár et al., 2025; 38 sections, $\sim$64k spots, 9 timepoints PCW 6.0–12.0) is our primary long-range temporal completion benchmark. Human fetal pancreas (Visium, Olaniru et al., 2023; 7 sections, 4 timepoints PCW 12, 15, 18, 20) is the sparse-supervision benchmark: fewer sections and a coarser temporal grid stress robustness where the training set is small. Developing human pancreas (Xenium; 12 sections, $\sim$1.78M cells, 6

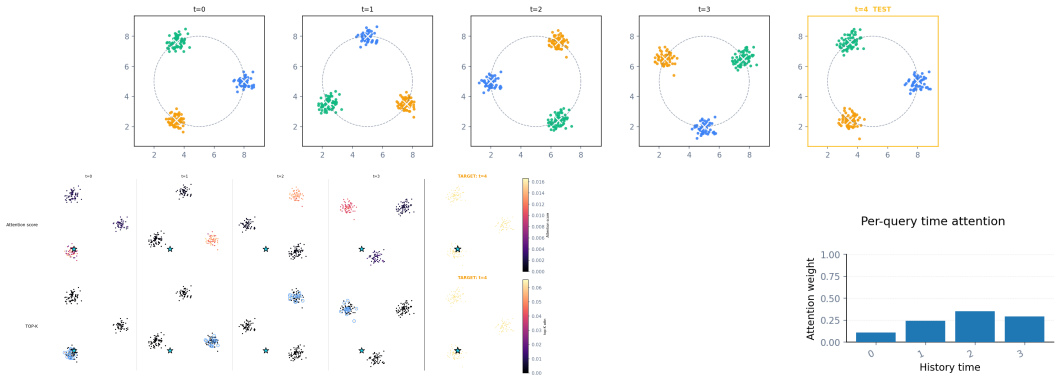

Figure 2: **Trajectory Memory: dataset and best-cosine attention.** *Top*: three niches (green/blue/orange) rotate around the unit circle; cells keep niche identity through time, the right panel (t=4) is held out. *Bottom*: spatial (left) and time (right) top-K attention for the best-cosine held-out query (cyan star).

timepoints PCW 8–17) is the scale-and-resolution benchmark: sub-cellular resolution and an order-of-magnitude larger cell count per section that exposes scaling behaviour not captured at Visium resolution. Kidney IRI (Xenium, Xuanyuan et al., 2025; 6 timepoints: sham, 4 h, 12 h, 2 d, 14 d, 6 wk) is the non-monotonic benchmark: unlike the smooth fetal series, acute injury and later fibrosis make early and late tissue differ sharply. Section/cell counts, image specs, and split details for each dataset are in Appendix F.

3.2 Evaluation Metrics

We report Pearson correlation (PCC), mean squared error (MSE), and mean absolute error (MAE) on held-out points and held-out time; for in-slide imputation the same metrics are restricted to the held-out region. On synthetic datasets metrics are computed in embedding space (details in Appendix F).

3.3 Baselines

We compare against four single-section H&E to expression baselines: Hist2ST (Zeng et al., 2022) (ConvMixer + Transformer + GNN over spot patches), BLEEP (Xie et al., 2023) (k-NN retrieval over a bimodal contrastive embedding), MERGE (Ganguly et al., 2025) (CNN + graph backbone with a count head), and STFlow (Huang et al., 2025) (flow-matching backbone with no temporal memory; our no-memory reference). For in-slide imputation we additionally compare against PH2ST (Niu et al., 2025), the strongest recent in-slide method. For temporal flow-matching baselines we use NicheFlow (Sakalyan et al., 2026), a single-modality flow on $(\mathbf{y}, \mathbf{p})$ with no conditioning on $\mathbf{x}$, and our extension NicheFlow+X that adds $\mathbf{x}$ conditioning to the same flow; the gap between them isolates the second-modality contribution from the history-attention contribution.

3.4 Main Results

Tables 1–2 report cross-time completion and Table 3 reports in-slide imputation. On the synthetic benchmarks, TEMPUS lifts test PCC by $+0.14$ on Phase Memory and $+0.35$ on Trajectory Memory over the strongest temporal baseline (NicheFlow+X: 0.871 vs. 0.730, and 0.883 vs. 0.534), and by $+0.34$ and $+0.36$ respectively over the strongest single-section baseline; the targets cannot be recovered from a single time-point by construction. On real data (count space, Table 2), TEMPUS reaches 0.794 PCC on Human heart against 0.705 for the next-best temporal model (NicheFlow+X, $+0.089$); on the smaller Human pancreas dataset the gap is tighter (0.812 vs. 0.805, $+0.007$). The same comparison on the Developing human pancreas (Xenium) dataset is reported in Appendix G.1 (Table 7), where TEMPUS reaches PCC 0.535 vs. 0.438 for NicheFlow+X ($+0.097$). For in-slide imputation, where a contiguous region of the test time-point is held out and predicted from the rest of the same scene plus earlier time-points, the gap to PH2ST is large on every dataset: $+0.27$ on PM (0.850 vs. 0.582), $+0.83$ on TM (0.891 vs. 0.064), and $+0.57$ on Human heart (0.810 vs. 0.235); the

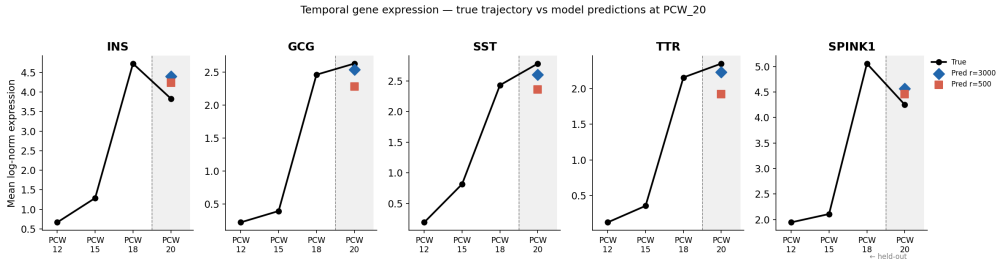

Figure 3: **Temporal gene expression trajectories and model predictions at the held-out timepoint.** Mean log-normalised expression of five marker genes across four developmental timepoints (PCW 12, 15, 18, 20) in the developing human pancreas (Olaniru et al., 2023). Solid lines show ground truth expression in the training slides (PCW 12–18); the shaded region marks the held-out test timepoint (PCW 20), never seen during training. Coloured markers show model predictions at PCW 20 for two history crop radii (r=3000 μm, diamonds; r=500 μm, squares).

TM gap is the largest because the target is a per-niche cumulative trajectory that cannot be recovered from a single time-point's morphology even when same-time spatial prompts are available. Further analyses and ablation studies are provided in the Appendices H.3–H.9.

3.5 Attention Analysis

We audit which past cells the model attends to, and at which past time-points it places weight, for a single target cell on the held-out evaluation pool of Trajectory Memory. The dataset provides a ground-truth answer to compare against: three niches rotate around the unit circle while keeping niche identity through time (Fig. 2, top), so for a target at t=4 the correct retrieval is the same-colour cells at every past step, regardless of where those cells sit in space (§E). We pick the target cell whose prediction is the most accurate in the held-out pool, shown as a cyan star in Fig. 2 (bottom); on this cell the model achieves a test PCC of 1.00 and an MSE of 0.0288. Two observations follow. First, in every past panel the top-K set lands on the correct-niche cluster even though that cluster moves spatially with t, so the model is tracking niche identity rather than geometric proximity. Second, the time-attention is recency-weighted, with roughly 0.65 of the mass on t=2, 3 and only 0.10 on t=0, matching the cumulative structure of the target. The worst-case query is reported in Appendix J.

3.6 Spatial-Temporal Gene Expression Prediction in the Developing Human Pancreas

We evaluated our method on spatial transcriptomics data from the developing human pancreas (Olaniru et al., 2023), spanning four timepoints across mid-gestation (PCW 12, 15, 18, and 20). The model was trained on PCW 12–18 and asked to predict cellular gene expression profiles at the held-out timepoint PCW 20, using only histology image embeddings and temporal memory from the training history.

Predicting the maturation of endocrine lineages PCW 20 corresponds to a critical window of pancreatic endocrine maturation, during which the islets of Langerhans become progressively organized and the major hormone-producing cell populations, beta (insulin, INS), alpha (glucagon, GCG), and delta (somatostatin, SST) cells, undergo coordinated transcriptional activation. The model correctly anticipates this developmental trajectory without ever observing a PCW 20 slide during training. As shown in the temporal curves of Fig. 3, the ground truth expression of INS, GCG, and SST rises monotonically across PCW 12–18, and continues increasing at PCW 20. The model's predictions at PCW 20 track this upward trend closely for all three genes, with INS showing the strongest signal. Transthyretin (TTR), another gene expressed by alpha cells, follows the same pattern: near-absent at PCW 12, strongly elevated by PCW 20, and correctly predicted to be high by the model. Notably, this result is not trivially explained by simple extrapolation of a mean trend, the model must identify, from the histology image of the unseen PCW 20 slide, which spatial locations are likely to harbour high hormone expression, and assign those predictions accordingly. The spatial expression maps (Fig. 14) confirm that the model does not predict uniform high expression across the

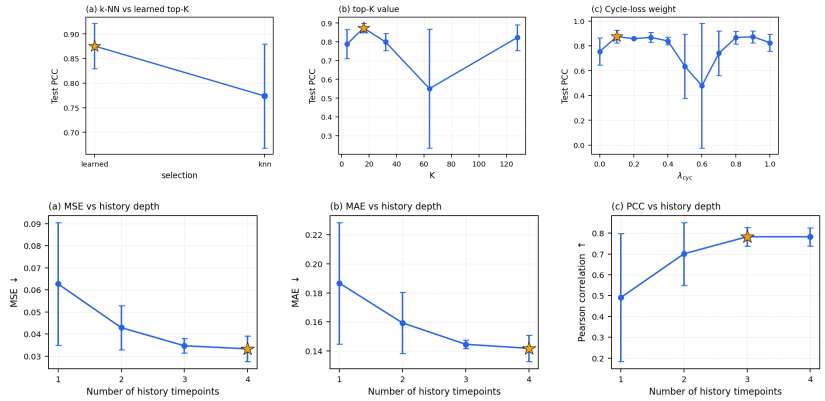

Figure 4: **Ablations on Trajectory Memory.** Mean across 3 seeds; bars are seed std; orange star marks the best setting in each panel. *Top*: (a) learned top-K vs. deterministic k-NN in the same d_{model} feature space; (b) $K \in \{4, 16, 32, 64, 128\}$; (c) cycle-loss weight $\lambda_{\text{cyc}} \in \{0.0, \ldots, 1.0\}$. *Bottom*: history-depth at inference (MSE / MAE / PCC vs. number of past time-points).

tissue. Instead, predicted INS, GCG, SST, and TTR are elevated in specific spatial regions consistent with the expected localisation of islet clusters, while remaining low elsewhere. This spatial specificity, achieved by conditioning on the H&E image embedding of the target slide, demonstrates that the model jointly leverages temporal context (what endocrine maturation looks like at PCW 15–18) and spatial context (where in the tissue such maturation is occurring at PCW 20).

Exocrine dynamics on SPINK1 and the per-spot spatial expression maps for the pancreas (Fig. 14) are reported in Appendix K. The same spatial-fidelity check on the human heart Visium dataset, covering cardiomyocyte and stromal markers across the held-out test timepoints (PCW 9.0–12.0), is reported in Appendix K (Figure 15).

4 Ablations

We isolate four architectural choices on Trajectory Memory; mean across 3 seeds with seed-std error bars is shown in Fig. 4. Replacing the learned top-K scorer with deterministic k-NN in the same feature space test PCC drops noticeably (top-a), so the learning-to-retrieve signal adds lift beyond raw geometric proximity. Sweeping $K \in \{4, 16, 32, 64, 128\}$ shows mid-range $K \in [16, 32]$ is sufficient and $K{=}4$ is unstable (top-b); the cycle-consistency loss is helpful at small weights ($\lambda_{\text{cyc}}{=}0.1$ wins) and recovers near $\lambda_{\text{cyc}}{=}1$ with a non-monotonic trough in between (top-c). Varying the number of past time-points fed to a fixed trained model at inference (bottom row, MSE/MAE/PCC) gives a consistent trend: the second past time-point provides the largest single jump, with diminishing returns thereafter. Full sweeps and the heart-data variant are in Appendix H.

5 Conclusion

We presented TEMPUS, a single architecture that unifies cross-time completion and within-slide imputation on serial spatial transcriptomics. The model combines per-query top-K differentiable attention over history cells, relative-time conditioning, and a cycle-consistency auxiliary loss; same-time observed context in the imputation regime is treated as just another history time-step, so both tasks share one architecture and one training procedure. On two synthetic benchmarks designed to make memory necessary, and on four real developmental datasets, human heart (Visium), human fetal pancreas (Visium), kidney, and developing human pancreas (Xenium), TEMPUS improves test PCC over the strongest single-section H&E to expression baseline and over the strongest temporal flow-matching baseline (NicheFlow and NicheFlow+X) on every dataset.

Limitations. The frozen pre-trained embedding-to-count decoder gives a fair count-space metric, but conflates errors from the embedding model and the decoder; jointly training a calibrated decoder is a natural extension we deliberately leave out of scope to keep the temporal model dataset-agnostic.

AI USAGE CLARIFICATION

In this work, we used generative AI tools for:

1. Design or provide feedback on research methodology or experiments. The core ideas of all experiments are human generated. We use AI to improve the rigor of experiments.
2. Implement methods. We use coding agents for implementation. We check all code and take full responsibility for its correctness.
3. Support qualitative and thematic data analysis. We use agents to help with making plots and visualizations.

We have not used generative AI tools for:

1. Clean and reformat datasets.
2. Interpret results.
3. we did not use generative AI to generate the synthetic datasets.

These required disclosure tasks are not applicable to this work:

1. Help develop theoretical models or conceptual frameworks.
2. Formulate mathematical claims.
3. Provide critical ingredients for proving mathematical claims.
4. Assist in the writing of proofs.
5. Propose or refine hypotheses.
6. Assist with translation.

Additionally, we used generative AI tools for:

1. Editing the research paper to improve readability. We use agents to find and correct grammar mistakes and point out sections which require more human editing.
2. Identifying relevant literature. We use agents to perform an additional literature search on top of our human search.

We reviewed all AI-assisted work and take responsibility for the final content of this paper, including its text, claims, code, and figures.

ETHICS STATEMENT

This work uses only previously published, publicly available datasets (Lázár et al., 2025; Olaniru et al., 2023; Xuanyuan et al., 2025) and collects no new human or animal data; all tissue data are de-identified and were obtained under the ethical approvals reported in the original studies. Our model predicts gene expression rather than measuring it, so its outputs are intended for research use and should not be treated as measured data or used for clinical decisions.

REPRODUCIBILITY STATEMENT

All datasets used are publicly available (Lázár et al., 2025; Olaniru et al., 2023; Xuanyuan et al., 2025). Architecture, hyperparameters, and training details are given in Appendices C and D, and dataset splits in Appendix F. The supplementary material ships the full implementation.

REFERENCES

Sebastian Birk, Mohammad Vali Sanian, Amirhossein Vahidi, Samuel Ogden, Daniyal J Jafree, Adib Miraki Feriz, Carlo Leonardi, Arpit Merchant, Zichen He, Lloyd Steele, Adam Boxall, Marta Rosa Sallese, Jonas Maaskola, Ciro Ramirez-Suastegui, Sevda Ogut, Koen Rademaker, M S Vijayabaskar, Batuhan Cakir, Sergio Marco Salas, Valentina Lorenzi, Josephine M Bryant, Cecilia Kyany'a, James O Jones, Raheleh Rahbari, Arti Mala Raghubar, Grant D Stewart, Benjamin Rumney, Catherine Tudor, Minal Patel, Jasmine Halliwell, Hon Man Chan, Tong Li, Heather Stanley, April Rose Foster, Fani Memi, Kenny Roberts, Andrew L Trinh, Elizabeth Tuck, Tannia Gracia, Shreya Rai, David J Adams, Simone Webb, Martin Prete, Roser Vento-Tormo, Mats Nilsson, Heiko Lickert, Fabian J Theis, Amin Ardestani, James J Sun, Foad J Rouhani, Lassi Paavolainen, Menna R Clatworthy, Omer Ali Bayraktar, Muzlifah Haniffa, Thomas Mitchell, Mostafa Bakhti, and Mo Lotfollahi. Multi-scale modeling of human tissues from spatial transcriptomics with TERRA. *bioRxiv*, pp. 2026.07.29.741565, August 2026.

Eduard Chelebian, Christophe Avenel, and Carolina Wählby. Combining spatial transcriptomics with tissue morphology. *Nature Communications*, 16(1):4452, 2025.

Richard J Chen, Tong Ding, Ming Y Lu, Drew F K Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L Weishaupt, Judy J Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nat. Med.*, 30(3):850–862, March 2024.

Youngmin Chung, Ji Hun Ha, Kyeong Chan Im, and Joo Sang Lee. Accurate spatial gene expression prediction by integrating multi-resolution features. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11591–11600. IEEE, 2024.

Xiaohang Fu, Yue Cao, Beilei Bian, Chuhan Wang, Dinny Graham, Nirmala Pathmanathan, Ellis Patrick, Jinman Kim, and Jean Yee Hwa Yang. Spatial gene expression at single-cell resolution from histology using deep learning with GHIST. *Nat. Methods*, 22(9):1900–1910, September 2025.

Aniruddha Ganguly, Debolina Chatterjee, Wentao Huang, Jie Zhang, Alisa Yurovsky, Travis Steele Johnson, and Chao Chen. MERGE: Multi-faceted hierarchical graph-based GNN for gene expression prediction from whole slide histopathology images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2412.02601.

Bryan He, Ludvig Bergenstråhle, Linnea Stenbeck, Abubakar Abid, Alma Andersson, Åke Borg, Jonas Maaskola, Joakim Lundeberg, and James Zou. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature biomedical engineering*, 4(8):827–834, 2020.

Tinglin Huang, Tianyu Liu, Mehrtash Babadi, Wengong Jin, and Rex Ying. Scalable generation of spatial transcriptomics from histology images via whole-slide flow matching. In *International Conference on Machine Learning (ICML)*, 2025. Spotlight.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier Hénaff, Matthew M Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver IO: A general architecture for structured inputs & outputs. *arXiv [cs.LG]*, July 2021.

Guillaume Jaume, Paul Doucet, Andrew H Song, Ming Y Lu, Cristina Almagro-Pérez, Sophia J Wagner, Anurag J Vaidya, Richard J Chen, Drew F K Williamson, Ahrong Kim, and Faisal Mahmood. HEST-1k: A dataset for spatial transcriptomics and histology image analysis, June 2024.

Bo Li, Yong Zhang, Qing Wang, Chengyang Zhang, Mengran Li, Guangyu Wang, and Qianqian Song. Gene expression prediction from histology images via hypergraph neural networks. *Brief. Bioinform.*, 25(6):bbae500, September 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. *Nat. Rev. Genet.*, 22(10):627–644, October 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Enikő Lázár, Raphaël Mauron, Žaneta Andrusivová, Julia Foyer, Mengxiao He, Ludvig Larsson, Nick Shakari, Sergio Marco Salas, Christophe Avenel, Sanem Sariyar, Jan Niklas Hansen, Marco Vicari, Paulo Czarnewski, Emelie Braun, Xiaofei Li, Olaf Bergmann, Christer Sylvén, Emma Lundberg, Sten Linnarsson, Mats Nilsson, Erik Sundström, Igor Adameyko, and Joakim Lundeberg. Spatiotemporal gene expression and cellular dynamics of the developing human heart. *Nat. Genet.*, 57(11):2756–2771, November 2025.

Vivien Marx. Method of the year: spatially resolved transcriptomics. *Nature methods*, 18(1):9–14, 2021.

Taku Monjo, Masaru Koido, Satoi Nagasawa, Yutaka Suzuki, and Yoichiro Kamatani. Efficient prediction of a spatial transcriptomics profile better characterizes breast cancer tissue sections without costly experimentation. *Sci. Rep.*, 12(1):4133, March 2022.

Lambda Moses and Lior Pachter. Museum of spatial transcriptomics. *Nature Methods*, 19(5):534–546, 2022.

Yi Niu, Jiashuai Liu, Yingkang Zhan, Jiangbo Shi, Di Zhang, Marika Reinius, Ines Machado, Mireia Crispin-Ortuzar, Jialun Wu, Chen Li, and Zeyu Gao. PH2ST:ST-prompt guided histological hypergraph learning for spatial gene expression prediction. *arXiv [cs.CV]*, March 2025.

Oladapo Edward Olaniru, Ulrich Kadolsky, Shichina Kannambath, Heli Vaikkinen, Kathy Fung, Pawan Dhami, and Shanta J Persaud. Single-cell transcriptomic and spatial landscapes of the developing human pancreas. *Cell Metab.*, 35(1):184–199.e5, January 2023.

Cantin Ortiz, Jose Fernandez Navarro, Aleksandra Jurek, Antje Märtin, Joakim Lundeberg, and Konstantinos Meletis. Molecular atlas of the adult mouse brain. *Sci. Adv.*, 6(26):eabb3446, June 2020.

Shi Pan, Jianan Chen, and Maria Secrier. Teaching pathology foundation models to accurately predict gene expression with parameter efficient knowledge transfer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 605–613. Springer, 2025.

Daniela Ruiz, Paula Cárdenas, Leonardo Manrique, Daniela Vega, Gabriel M Mejia, and Pablo Arbeláez. Completing spatial transcriptomics data for gene expression prediction benchmarking. *Med. Image Anal.*, 106(103754):103754, December 2025.

Kristiyan Sakalyan, Alessandro Palma, Filippo Guerranti, Fabian Theis, and Stephan Günnemann. Modeling microenvironment trajectories on spatial transcriptomics with nicheflow. *Advances in Neural Information Processing Systems*, 38:103328–103371, 2026.

Mohammad Vali Sanian, Arshia Hemmat, Amirhossein Vahidi, Jonas Maaskola, Jimmy Tsz Hang Lee, Stanislaw Makarchuk, Yeliz Demirci, Nana-Jane Chipampe, Muzlifah Haniffa, Omer Bayraktar, Lassi Paavolainen, and Mohammad Lotfollahi. 3D-guided scalable flow matching for generating volumetric tissue spatial transcriptomics from serial histology. *arXiv [cs.CV]*, November 2025.

Patrik L Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O Westholm, Mikael Huss, Annelie Mollbrink, Sten Linnarsson, Simone Codeluppi, Åke Borg, Fredrik Pontén, Paul Igor Costea, Pelin Sahlén, Jan Mulder, Olaf Bergmann, Joakim Lundeberg, and Jonas Frisén. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294): 78–82, July 2016.

Chuhan Wang, Adam S Chan, Xiaohang Fu, Shila Ghazanfar, Jinman Kim, Ellis Patrick, and Jean Y H Yang. Benchmarking the translational potential of spatial gene expression prediction from histology. *Nat. Commun.*, 16(1):1544, February 2025.

Ronald Xie, Kuan Pang, Sai Chung, Catia Perciani, Sonya MacParland, Bo Wang, and Gary Bader. Spatially resolved gene expression prediction from histology images via bi-modal contrastive learning. *Advances in Neural Information Processing Systems*, 36:70626–70637, 2023.

Qiao Xuanyuan, Haojia Wu, Hemalatha Sundaramoorthi, Pierre Isnard, Changfeng Chen, Waleed Rahmani, and Benjamin D Humphreys. Multimodal spatial transcriptomic characterization of mouse kidney injury and repair. *Nat. Commun.*, 16(1):7567, August 2025.

Yuansong Zeng, Zhuoyi Wei, Weijiang Yu, Rui Yin, Yuchen Yuan, Bingling Li, Zhonghui Tang, Yutong Lu, and Yuedong Yang. Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. *Brief. Bioinform.*, 23(5):bbac297, September 2022.

Zhenyi Zhang, Yuhao Sun, Qiangwei Peng, Tiejun Li, and Peijie Zhou. Integrating dynamical systems modeling with spatiotemporal scrna-seq data analysis. *Entropy*, 27(5):453, 2025.

Peijie Zhou, Federico Bocci, Tiejun Li, and Qing Nie. Spatial transition tensor of single cells. *Nat. Methods*, 21(6):1053–1062, June 2024.

Junchao Zhu, Ruining Deng, Tianyuan Yao, Juming Xiong, Chongyu Qu, Junlin Guo, Siqi Lu, Mengmeng Yin, Yu Wang, Shilin Zhao, Haichun Yang, and Yuankai Huo. Asign: An anatomy-aware spatial imputation graphic network for 3d spatial transcriptomics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30829–30838, 2025.

A Training Procedure

The full TEMPUS training step is given in Algorithm 1. Each iteration samples a target time and a spatial crop, embeds target queries, performs per-step sparse selection and aggregation over the history buffer, fuses the per-step tokens via cross-attention, predicts the held-out content, and applies the combined MSE + cycle-consistency loss.

Algorithm 1 TEMPUS training

Require: Target time t, target crop $\mathcal{C}^{(t)}$, history $\mathcal{H}_t$

// Embed target queries
$\mathbf{u}_q \leftarrow \mathrm{MLP}_x(\mathbf{x}_q) + \mathrm{PE}_{\mathrm{sp}}(\mathbf{p}_q)$ for all $q \in \mathcal{C}^{(t)}$

// Per-step sparse selection and aggregation
for $\tau \in \mathcal{H}_t$ **do**
  Embed history: $\mathbf{u}_i, \mathbf{v}_i$ via MLP_x, MLP_y, $\mathrm{PE}_{\mathrm{sp}}$
  Score: $s_{q,i}^{(\tau)}$ via QK-LayerNorm with learnable σ_h
  Select: $\mathcal{I}_q^{(\tau)} \leftarrow \mathrm{argtopK}(\mathbf{s}_q^{(\tau)})$ ▷ STE backward
  Aggregate: $\mathbf{a}_q^{(\tau)} \leftarrow \mathrm{Agg}\big(\sum_{i \in \mathcal{I}} \alpha_{q,i}[\mathbf{u}_i; \mathbf{v}_i]\big) + \boldsymbol{\gamma}(\Delta\tau)$
end for

// Temporal fusion and prediction
$\mathbf{z}_q \leftarrow \mathrm{CrossAttnDecoder}\big(\mathbf{u}_q, \{\mathbf{a}_q^{(\tau)}\}_\tau\big)$
$\mathbf{z}_q \leftarrow \mathrm{TargetSelfAttn}(\{\mathbf{z}_q\}_q)$
$\hat{\mathbf{y}}_q \leftarrow h(\mathbf{z}_q)$

// Loss (Eq. 8)
$\mathcal{L} \leftarrow \mathrm{MSE} + \lambda_{\mathrm{cyc}}\,\mathcal{L}_{\mathrm{cyc}}$
Backpropagate $\mathcal{L}$; update parameters

B Inference Procedure

At inference, the target time-step is partitioned into overlapping spatial chunks to bound memory and preserve locality; each chunk is processed independently with its own spatially cropped history, and overlapping predictions are averaged. When rolling forward across multiple test time-steps, earlier model predictions replace missing ground truth in the history buffer; under in-slide imputation, the propagated representation mixes predicted content for held-out points with observed content elsewhere, matching the training protocol. The cycle-consistency decoder ψ is disabled at inference, since it serves only as a training regulariser. For tissue datasets, count-level metrics are computed by decoding $\hat{\mathbf{y}}_q$ through the frozen NB `CountDecoder` (§2.7); on synthetic data no decoder is used. The full procedure is given in Algorithm 2.

Algorithm 2 TEMPUS inference

Require: Target time t, observed history $\mathcal{T}_{\text{hist}}$, optional same-time context; spatial chunk centres
Ensure: Predicted content embeddings $\{\hat{\mathbf{y}}_q\}$ at held-out points
for each spatial chunk centred at $\mathbf{p}_c$ **do**
  Crop history steps around $\mathbf{p}_c$ (radius r_h)
  if autoregressive rollout **then**
    Substitute predicted $\hat{\mathbf{y}}$ for earlier test time steps
  end if
  if in-slide imputation **then**
    Prepend same-time context $\mathcal{S}^{(t,0)}$ with $\Delta\tau = 0$
  end if
  Run forward pass (Algorithm 1, lines 2–15) without cycle decoder
  Accumulate per-point predictions; average over overlapping tiles
end for
if tissue data **then**
  Decode $\hat{\mathbf{y}} \to \hat{\mathbf{c}}$ through frozen embedding-to-count decoder
end if
return $\{\hat{\mathbf{y}}_q\}$ (and $\{\hat{\mathbf{c}}_q\}$ on tissue datasets)

C Architecture Details

Input encoder (f_{enc}). In the tissue instantiation, patch-level image features are extracted from a frozen UNI (Chen et al., 2024) ViT-L/16 pretrained with DINOv2 on over 100M histology patches ($d_x = 1024$). Features are extracted once per dataset and cached; the vision encoder is never finetuned. On synthetic benchmarks, $\mathbf{x}$ is provided directly as a low-dimensional vector and no encoder is needed; $\mathbf{x}$ is fed as input wherever the real-data setup would consume the histology image embedding, both for TEMPUS and for every baseline, so the input modality is shared across all models on synthetic data.

Content encoder. In the tissue instantiation, per-cell gene expression is embedded by TERRA (Birk et al., 2026), a foundational spatial-transcriptomics encoder that produces a 384-dim per-cell embedding and a 384-dim per-neighbourhood embedding; we concatenate them to form the content vector $\mathbf{y} \in \mathbb{R}^{d_y}$ with $d_y = 768$. The encoder is run once per dataset and frozen; TEMPUS is trained entirely in this embedding space and only decoded back to counts at evaluation. On synthetic data, $\mathbf{y}$ is provided directly and no encoder/decoder is needed.

Modality projections (MLP_x, MLP_y). Each is a two-layer MLP with GELU activation: $\mathbb{R}^{d_x} \to \mathbb{R}^d \to \mathbb{R}^d$ and $\mathbb{R}^{d_y} \to \mathbb{R}^d \to \mathbb{R}^d$ respectively, with LayerNorm after each linear layer.

Spatial positional encoding (PE_{sp}). Fourier-feature encoding of normalised coordinates $\mathbf{p}/s_{\text{coord}}$ with K_f geometric frequencies $\pi \cdot 2^k$, $k = 0, \ldots, K_f - 1$. We use $K_f = 8$, producing a $4 \times K_f = 32$-dimensional feature (sin and cos for each of the two spatial coordinates at each frequency) that is projected to $\mathbb{R}^d$ by a linear layer. The normalisation constant s_{coord} is set once as $1.1 \times \max_{i,t} \|\mathbf{p}_i^{(t)}\|_\infty$ across the training set and frozen.

Temporal encoding (γ). Sinusoidal embedding of the relative offset $\Delta\tau = t - \tau$ using 64 frequencies, followed by a learned linear projection to $\mathbb{R}^d$.

Top-K attention heads. The scoring uses multi-head attention with n_h heads and per-head dimension $d_h = d/n_h$. Both queries and keys are LayerNorm-normalised before the dot product. Each head has an independent learnable inverse temperature σ_h initialised to 1.0.

Aggregator (Agg). A shared two-layer MLP with GELU: $\mathbb{R}^{2d} \to \mathbb{R}^d \to \mathbb{R}^d$, applied identically at every history step. The input dimension is $2d$ because it receives the concatenation of the input and content value channels $[\mathbf{u}_i; \mathbf{v}_i]$.

Cross-attention decoder. An L_c-layer transformer decoder where target queries cross-attend to the sequence of per-step aggregated tokens. Each layer consists of multi-head cross-attention, LayerNorm, a feed-forward block with hidden dimension $4d$ and GELU activation, and residual connections. No self-attention within the decoder layers; temporal interactions occur solely through cross-attention to the history token sequence.

Target self-attention. An L_s-layer self-attention stack applied after the cross-attention decoder, allowing target queries within the same spatial crop to exchange information. Each layer follows the standard pre-norm transformer block with multi-head self-attention, feed-forward, and residual connections.

Prediction head (h). A two-layer MLP: $\mathbb{R}^d \to \mathbb{R}^d \to \mathbb{R}^{d_y}$ with GELU activation and LayerNorm after the first layer.

Cycle decoder (ψ). A two-layer MLP: $\mathbb{R}^{d+d_\gamma} \to \mathbb{R}^d \to \mathbb{R}^d$ with GELU, where d_γ is the temporal encoding dimension. Takes as input the concatenation of the fused representation $\mathbf{z}_q$ and the relative time embedding $\boldsymbol{\gamma}(\Delta\tau)$. Instantiated only when $\lambda_{\text{cyc}} > 0$.

Per-query selection rather than a learned latent bottleneck. Distinct from prior latent-bottleneck architectures (Perceiver-IO and its descendants), our queries are the target points themselves; each point selects its own K history neighbours at every past time step. Selection thus depends jointly on the query's input embedding, its spatial coordinate, and the offset to the history time, exactly the inductive bias one wants on data whose spatial structure is non-stationary across time (tissue undergoing deformation, environmental sensors moving, evolving niches). A latent bottleneck would average over all queries before retrieval and lose this per-point specificity.

Two-stage training summary. The TEMPUS model is supervised end-to-end only with embedding-space MSE on $\mathbf{y}$. For each spatial-transcriptomics dataset, a separate small Negative-Binomial (NB) decoder is pre-trained on (TERRA embedding, raw count) pairs of that dataset's training split, then frozen. Architecture: a 2-layer MLP backbone (Linear→LayerNorm→SiLU→Dropout) feeding a softplus mean head and a per-gene learnable dispersion. Concretely, for embedding $\mathbf{y} \in \mathbb{R}^{d_y}$ and gene $g \in \{1, \dots, G\}$,

$$\boldsymbol{\mu}(\mathbf{y}) = \text{softplus}\big(W\,\text{MLP}(\mathbf{y})\big), \quad \theta_g = \text{softplus}(\theta_{0,g}), \tag{9}$$

with the NB likelihood

$$P(x_g \mid \mu_g, \theta_g) \;=\; \frac{\Gamma(x_g + \theta_g)}{\Gamma(\theta_g)\,\Gamma(x_g + 1)} \left(\frac{\theta_g}{\theta_g + \mu_g}\right)^{\theta_g} \left(\frac{\mu_g}{\theta_g + \mu_g}\right)^{x_g}, \tag{10}$$

and the training objective

$$\mathcal{L}_{\text{NB}} \;=\; -\,\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \sum_{g=1}^{G} \log P\big(x_{i,g} \,\big|\, \mu_g(\mathbf{y}_i),\, \theta_g\big), \tag{11}$$

fitted only on the training split. At evaluation, the temporal model's $\hat{\mathbf{y}}$ is passed through this frozen NB head to obtain count-space predictions $\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}(\hat{\mathbf{y}})$; reported count-space metrics use $\hat{\boldsymbol{\mu}}$ versus the ground-truth counts. This keeps every method comparable in the same count space. This decode back to counts step applies only to spatial-transcriptomics data; on synthetic datasets, $\mathbf{y}$ is the prediction target itself and no decoder is used.

D HYPERPARAMETERS

Inputs. For all real datasets, image embeddings are extracted once with a frozen UNI v1 ViT-L/16 (Chen et al., 2024) on 96×96 px patches centred on each spot or cell. Counts are normalised through the standardised count layer used by the shared decoder.

Table 4: Default hyperparameters for the real and synthetic regimes. Values that match across the two regimes are shown once.

Category	Parameter	Real	Synthetic
Model dimensions			
	Model dimension d	128	16
	Attention heads n_h	4	1
	Per-head dimension $d_h = d/n_h$	32	16
	FFN hidden dimension ($4d$)	512	64
	Fourier frequencies K_f	8	
Top-K selection			
	Selected entries K	32	24
	Straight-through estimator	on	
	Initial inverse temperature σ_h	1.0	
Architecture depth			
	Cross-attention decoder layers L_c	2	
	Target self-attention layers L_s	1	
	Bottleneck (aggregator) layers	2	
Spatial cropping			
	Target crop radius r_t	dataset-specific	—
	History crop radius r_h	dataset-specific	—
	Aligned history crops	on	—
	Minimum points per crop	50	—
Training			
	Optimiser	AdamW	
	Learning rate	3×10^{-4}	
	Weight decay	0.01	
	LR schedule	cosine annealing	
	Warmup steps	500	
	Gradient clipping	1.0	
	Epochs	10	50
	Dropout	0.1	
Loss weights			
	Cycle weight λ_{cyc}	0.5	

Optimisation. We train with AdamW (Loshchilov & Hutter, 2019) at lr 3×10^{-4}, weight decay 0.01, cosine annealing, and warmup of 500 steps. Real datasets: 10 epochs. Synthetic datasets: 50 epochs. Gradient clipping at norm 1.0.

Table 4 lists the defaults used across all experiments. The real/synthetic split reflects the two regimes: the real-data instantiation operates on 1024-dim UNI image embeddings and 768-dim TERRA content embeddings, so a wider d=128 is appropriate; the synthetic instantiation uses small low-dimensional vectors and runs at d=16. Values not in the table (Fourier-feature normaliser, FFN type, layernorms) follow the architecture spec in §C.

Crops. Target crop radius r_t tuned per dataset (Visium: 200 px; Xenium: $80\,\mu$m). History crop radius r_h=500 px on Visium, $200\,\mu$m on Xenium.

Evaluation chunking. Spatial chunking via bucket-sort (lexicographic on (g_y, g_x) integer-bucket coordinates) with optional overlapping-tile averaging at stride $r_t/2$. All temporal eval helpers run under `@torch.no_grad()`.

E TOY DATASETS IN DETAIL

Table 5 summarises both synthetic benchmarks at a glance; per-dataset construction details follow.

Table 5: Synthetic-dataset overview. Each timepoint is a single N-point cloud; defaults are listed. Generator equations are in Appendix E.

Dataset	N (cells/tp)	d (output dim)	T (timepoints)	Modes / niches	Train timepoints	Test timepoints
Phase Memory (PM)	500	10	10	3 modes	$t = 0, \ldots, 7$	$t = 8, 9$
Trajectory Memory (TM)	200	10	5	3 niches	$t = 0, \ldots, 3$	$t = 4$

Phase Memory. At each time $t \in \{0, \ldots, T-1\}$ we re-sample N point positions uniformly on $[0, 10]^2$, $\mathbf{p}_i^{(t)} \sim \mathcal{U}([0, 10]^2)$. Two ingredients of the generator are fixed across time and used by every time-step: a shared linear map $W_{\text{sh}} \in \mathbb{R}^{d\times d}$, and one per-mode linear map $W_m \in \mathbb{R}^{d\times d}$ for $m = 0, \ldots, M-1$. (Both are sampled once from $\mathcal{N}(0, 0.3^2)$.)

Input signal $\mathbf{x}$. Each point's $\mathbf{x}$ embedding is a fixed Fourier-style feature map of its coordinate, plus a small time drift:

$$\boldsymbol{\phi}(\mathbf{p}) \;=\; \big[\sin(\tfrac{p_x}{2}),\, \cos(\tfrac{p_x}{2}),\, \sin(\tfrac{p_y}{2}),\, \cos(\tfrac{p_y}{2}),\, \sin(0.3(p_x{+}p_y)),\, \cos(0.3(p_x{-}p_y)),\, \tfrac{p_x}{10},\, \tfrac{p_y}{10}\big] \in \mathbb{R}^8,$$

$$\mathbf{x}_i^{(t)} \;=\; B\,\boldsymbol{\phi}(\mathbf{p}_i^{(t)}) \;+\; 0.1\,\sin\!\big(\tfrac{2\pi t}{T}\big)\,\mathbf{d} \;+\; \boldsymbol{\eta}_x, \quad \boldsymbol{\eta}_x \sim \mathcal{N}(0, 0.02^2 I),$$

with fixed $B \in \mathbb{R}^{d\times 8}$ and a fixed drift direction $\mathbf{d} \in \mathbb{R}^d$. So $\mathbf{x}$ is essentially a static function of position, with a tiny periodic shift in t.

Rotating modes. Each point at time t is assigned to one of M angular sectors around the centre $\mathbf{c} = (5, 5)$, with a sector boundary that rotates at constant angular velocity $\theta_t = 2\pi t/T$:

$$m_i^{(t)} \;=\; \Big\lfloor \frac{\big(\angle(\mathbf{p}_i^{(t)} - \mathbf{c}) - \theta_t\big) \bmod 2\pi\big)}{2\pi/M} \Big\rfloor \in \{0, \ldots, M-1\}.$$

The same physical position therefore belongs to a different mode at different times.

Content target $\mathbf{y}$. A convex mixture of a shared $\mathbf{x}$-only map and a mode-specific $\mathbf{x}$-conditional map:

$$\mathbf{y}_i^{(t)} \;=\; \alpha\, W_{\text{sh}}\, \mathbf{x}_i^{(t)} \;+\; (1-\alpha)\, W_{m_i^{(t)}}\, \mathbf{x}_i^{(t)} \;+\; \boldsymbol{\eta}_y, \quad \boldsymbol{\eta}_y \sim \mathcal{N}(0, \sigma^2 I),$$

where $\alpha = \texttt{shared_weight} \in [0, 1]$ controls how much of $\mathbf{y}$ is recoverable from $\mathbf{x}$ alone: $\alpha{=}1$ makes history useless, $\alpha{=}0$ makes the mode (hence the rotation phase) strictly necessary.

Why memory is provably needed. At a held-out target time $t^\star$, the observed $\mathbf{x}_i^{(t^\star)}$ depends almost entirely on position; any two points at the same position over different times have nearly identical $\mathbf{x}$. Their $\mathbf{y}$'s differ because their modes differ, and the mode is determined by the rotation phase $\theta_{t^\star}$ which is not encoded in $\mathbf{x}$. To predict $\mathbf{y}_i^{(t^\star)}$ a model must read $\theta_{t^\star}$ off the past: i.e. recognise which past time-step's points share the current mode at the current position.

Defaults. $T{=}10$, $N{=}500$, $M{=}3$, $d{=}10$, $\alpha{=}0.3$, $\sigma{=}0.05$.

Biological analogue. A developmental tissue where the same morphological niche acquires a different transcriptomic identity across stages. Mode visualisation: Fig. 5.

Trajectory Memory. Each of the N points is permanently assigned to one of m niches; let $n(i) \in \{0, \ldots, m-1\}$ denote the niche of point i. The dataset has three coupled mechanisms: (i) niche centres rotate jointly on a circle, (ii) each niche has its own scalar dynamic in time, and (iii) the content target is the time-integral of $\mathbf{x}$.

Niche geometry (rotating circle). Niche centres sit on a circle of radius R around tissue centre $\mathbf{c} = (5, 5)$, with initial angles equispaced and a common angular speed $\omega_\circ = 2\pi/(T{-}1)$:

$$\theta_n(t) \;=\; \tfrac{2\pi n}{m} \;+\; t\,\omega_\circ, \qquad \text{ctr}_n(t) \;=\; \mathbf{c} \;+\; R\,(\cos\theta_n(t),\, \sin\theta_n(t)).$$

Each point's coordinate is its niche centre plus a fixed within-niche offset and a small linear velocity perturbation:

$$\mathbf{p}_i^{(t)} \;=\; \text{ctr}_{n(i)}(t) \;+\; \mathbf{o}_i \;+\; t\,\mathbf{v}_i, \quad \mathbf{o}_i \sim \mathcal{N}(0, \sigma^2_{\text{niche}} I),\; \mathbf{v}_i \sim \mathcal{N}(0, \sigma^2_{\text{vel}} I).$$

So the same physical position is occupied by different niches at different times.

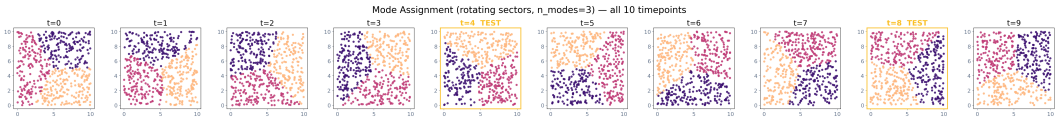

Figure 5: **Phase Memory.** Per-cell mode assignment over time. Each subplot is a time point; colour encodes mode. The mapping rotates with t so a memory-less predictor cannot distinguish modes from $\mathbf{x}^{(t)}$ alone.

Per-niche orthogonal directions. Each niche is given a unit vector $\mathbf{v}_n \in \mathbb{R}^d$. The set $\{\mathbf{v}_n\}_{n=1}^m$ is sampled once and made orthonormal via QR decomposition (when $m \leq d$); otherwise random unit vectors. Different niches therefore occupy disjoint subspaces of feature space: $\mathbf{v}_n^\top \mathbf{v}_{n'} \approx 0$ for $n \neq n'$.

Per-niche scalar trajectory. Each niche has its own oscillating-plus-drifting scalar:

$$g_n(t) \;=\; A_n \,\sin\!\big(\omega_n\, t + \phi_n\big) \;+\; \mu_n \,\tfrac{t}{T-1},$$

with $A_n, \omega_n, \phi_n, \mu_n$ sampled once per niche from fixed distributions.

Input signal $\mathbf{x}$*.* At each time-step the $\mathbf{x}$ embedding is the niche-direction vector scaled by the niche's current scalar:

$$\mathbf{x}_i^{(t)} \;=\; g_{n(i)}(t) \cdot \mathbf{v}_{n(i)} \;+\; \boldsymbol{\eta}_x, \quad \boldsymbol{\eta}_x \sim \mathcal{N}(0, \sigma^2 I).$$

Because niche directions are orthogonal, projecting $\mathbf{x}_i^{(t)}$ onto each $\mathbf{v}_n$ reveals the niche identity exactly, but nothing about past or future values.

Content target $\mathbf{y}$ *(time-integral).* The content is the running cumulative sum of $\mathbf{x}$ along time, which lives in the same niche subspace as $\mathbf{x}$ but encodes the integral of $g_{n(i)}$ rather than its instantaneous value:

$$\mathbf{y}_i^{(t)} \;=\; \sum_{s=0}^{t} \mathbf{x}_i^{(s)} \;+\; \boldsymbol{\eta}_y \;=\; \Big(\sum_{s=0}^{t} g_{n(i)}(s)\Big)\, \mathbf{v}_{n(i)} \;+\; \boldsymbol{\eta}_y.$$

(An optional $\frac{1}{t+1}$ normalisation is available; we leave it off in the headline numbers.)

Why memory and cross-cell attention are both required. At held-out time $t^\star$, observing $\mathbf{x}_i^{(t^\star)}$ reveals only $g_{n(i)}(t^\star)\,\mathbf{v}_{n(i)}$, hence niche identity and the instantaneous scalar; it gives no information about $\sum_{s<t^\star} g_{n(i)}(s)$. To recover $\mathbf{y}_i^{(t^\star)}$ a model must read past values $g_{n(i)}(s)$ from history. Because niches rotate spatially, those values cannot be read from points in the same physical neighbourhood; the model has to attend to past points of the same niche identity, which sit at different spatial positions in each past panel. Cross-niche attention is useless (orthogonal subspaces); cross-time within-niche attention is necessary. This is exactly the inductive bias top-K over the joint $\mathbf{x}$ + spatial scoring is designed to deliver.

Defaults. T=5, N=200, m=3, d=10, R=3, σ_{niche}=0.3, σ_{vel}=0.05, $\omega_n \sim \mathcal{U}(0.3, 1.5)$, $A_n \sim \mathcal{U}(0.5, 1.0)$, $\phi_n \sim \mathcal{U}(0, 2\pi)$, $\mu_n = 0$, σ=0.05. Latest time-step held out as test.

Biological analogue. Niche-specific gene programmes (e.g. pacemaker vs. working cardiomyocytes) whose expression trajectory is driven by an internal cumulative behavior that cannot be inferred from a single timepoint's morphology. Dataset overview: Fig. 6.

F Dataset overviews (real data)

Table 6 summarises the four real datasets at a glance; per-dataset details follow.

Gene panel size and metric scale. The model and the count-space metrics in Table 2 are computed across all G genes that pass each dataset's standard preprocessing filter, with G specific to each dataset (heart: 9,021; pancreas: 4,798; Developing human pancreas: 5,022; see Table 6). Because the gene sets and per-gene count magnitudes differ across datasets, count-space MSE and MAE values are not directly comparable across columns of Table 2; PCC is computed per-cell and is the metric that compares meaningfully across datasets.

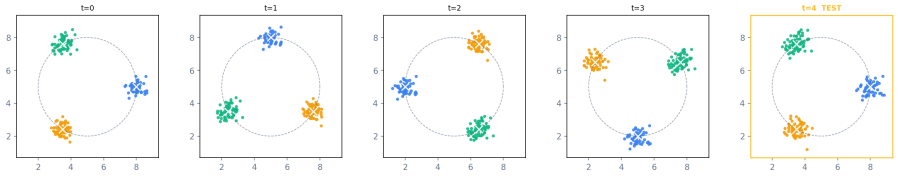

Figure 6: **Trajectory Memory.** Niche assignment, niche centre trajectories, and per-niche scalar trajectories $g_n(t)$. Niche directions $\mathbf{v}_n$ are approximately orthonormal so the niche identity is recoverable from $\mathbf{x}^{(t)}$, but the cumulative trajectory is not.

Table 6: Real-dataset overview. Sections, total cells (Xenium) or spots (Visium), and the train/test split used throughout the paper. The full timepoint list for each dataset is the union of its train and test columns. Kidney IRI sections are subsampled to 20,000 cells each (from 1,374,915 cells total) to keep all methods comparable in memory.

Dataset (platform)	# sections	# cells / spots	# genes	# tp	Train timepoints	Test timepoints
Human heart (Visium) (Lázár et al., 2025)	38	$\sim$64,316 spots	9,021	9	PCW 6.0, 7.0, 7.5, 8.0	PCW 9.0, 10.0, 10.5, 11.0, 12.0
Human fetal pancreas (Visium) (Olaniru et al., 2023)	7	$\sim$1,200–2,100/section	4,798	4	PCW 12, 15, 18	PCW 20
Developing human pancreas (Xenium)	12	1,779,792 cells	5,022	6	PCW 8, 11, 12	PCW 13, 14, 17
Kidney IRI (Xenium) (Xuanyuan et al., 2025)	12	240,000 (20,000/section)	299	6	sham, 4 h	12 h, 2 d, 14 d, 6 wk

F.1 Human heart (Visium)

38 sections (64,316 spots) across 9 timepoints PCW 6.0–12.0 with paired H&E and 10x Visium (Lázár et al., 2025). The dataset captures the establishment of the four-chamber heart, septation, and valve formation. 20× whole-slide images at $\sim$10,000 px per axis. Test split: `PCW_9.0,PCW_10.0,PCW_10.5,PCW_11.0,PCW_12.0` (5 timepoints; training uses PCW 6.0–8.0).

F.2 Human pancreas (Visium)

7 Visium sections across 4 timepoints (PCW 12, 15, 18, 20) of the human fetal pancreas profiling the emergence of endocrine and acinar lineages and the islet niche (Olaniru et al., 2023). Section sizes range from $\sim$1,200 to $\sim$2,100 spots, paired H&E. Test split: latest stages.

F.3 Developing human pancreas (Xenium)

12 sections, 1,779,792 cells in total, across 6 timepoints PCW 8–17 of human fetal pancreas development profiled with 10x Xenium. Aligned H&E at 0.5 μm/px.

G Baselines in detail

Hist2ST (Zeng et al., 2022). A per-spot regression model that predicts gene expression from a single tissue patch. The pipeline is a ConvMixer feature extractor on a 224×224 px patch centred on the spot, followed by a Transformer block with spatial position tokens and a graph neural network that lets nearby spots share representations through a learned adjacency. Output is a per-gene regression head trained with MSE on log-normalised counts. There is no cross-section memory: each section is processed independently. We use it as the simplest cross-modal baseline.

BLEEP (Xie et al., 2023). A bimodal contrastive learning method that jointly embeds H&E patches and gene-expression vectors of the corresponding spots into a shared space using an InfoNCE-style objective. At inference time, prediction proceeds by retrieval: the embedding of the test patch is queried against the training-set library of (image, expression) pairs, and the predicted expression is a k-nearest-neighbour weighted average of the library expressions. This keeps the training set as a non-parametric memory at inference but operates within a single section.

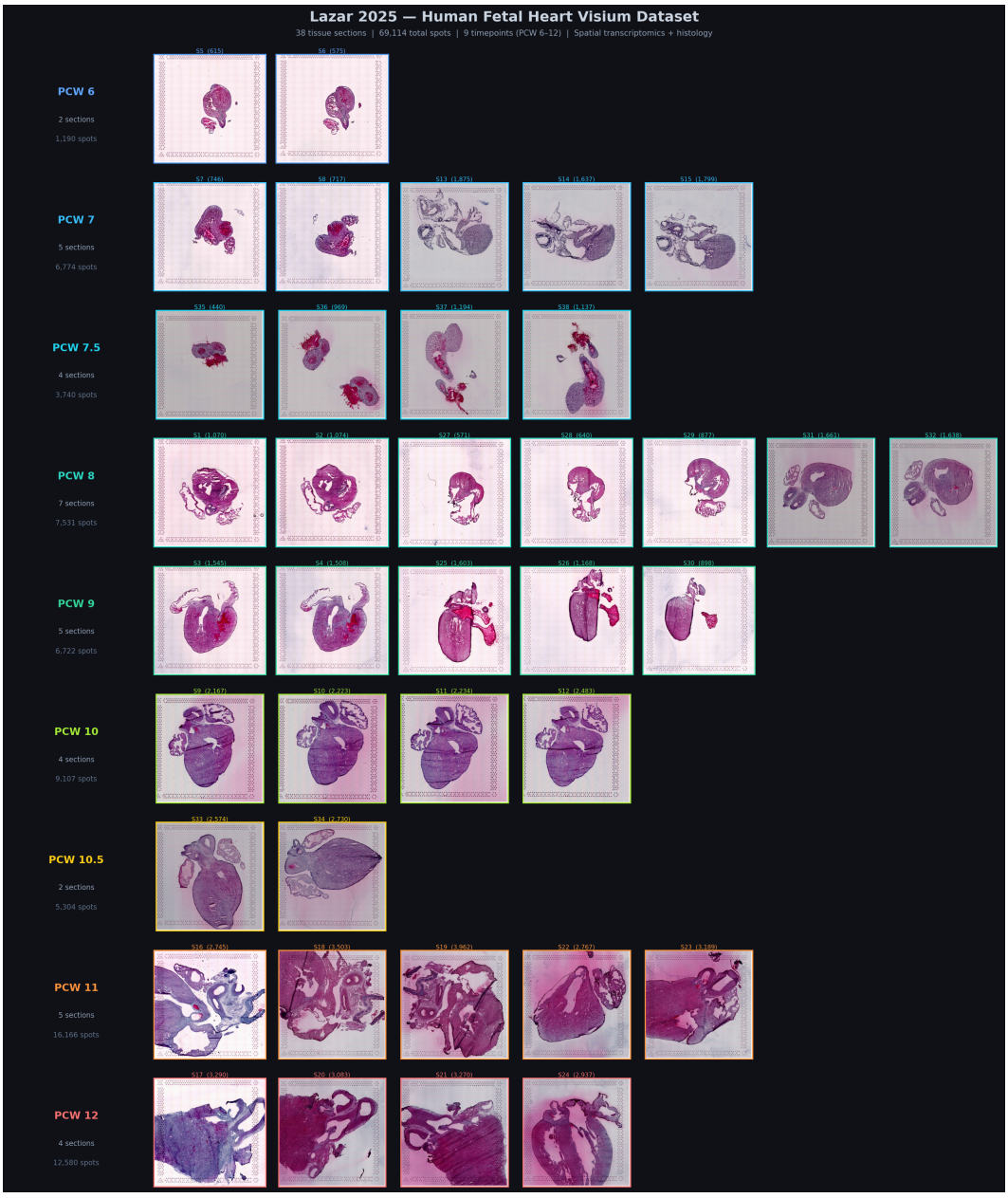

Figure 7: **Heart dataset overview.** Per-timepoint H&E with overlaid Visium spots; PCW6.0 (small, undivided heart tube) through PCW12.0 (four-chambered, septated heart).

MERGE (Ganguly et al., 2025). A hierarchical graph neural network for spatial gene-expression prediction from whole-slide images. The slide is partitioned into image patches that are clustered both spatially (geometric proximity) and morphologically (feature similarity), giving two tiers of edges: intra-cluster (local) and inter-cluster (long-range). A CNN encoder produces patch features; a GNN propagates information along both edge tiers; a count head predicts gene expression directly. The model also evaluates several gene-aware count-smoothing schemes to mitigate spatial-transcriptomics technical noise.

STFlow (Huang et al., 2025). A whole-slide flow-matching model that casts spot-wise expression prediction as conditional flow matching: starting from Gaussian noise, a learned vector field iteratively transports the sample to the target gene-expression distribution, conditioned on histology features and

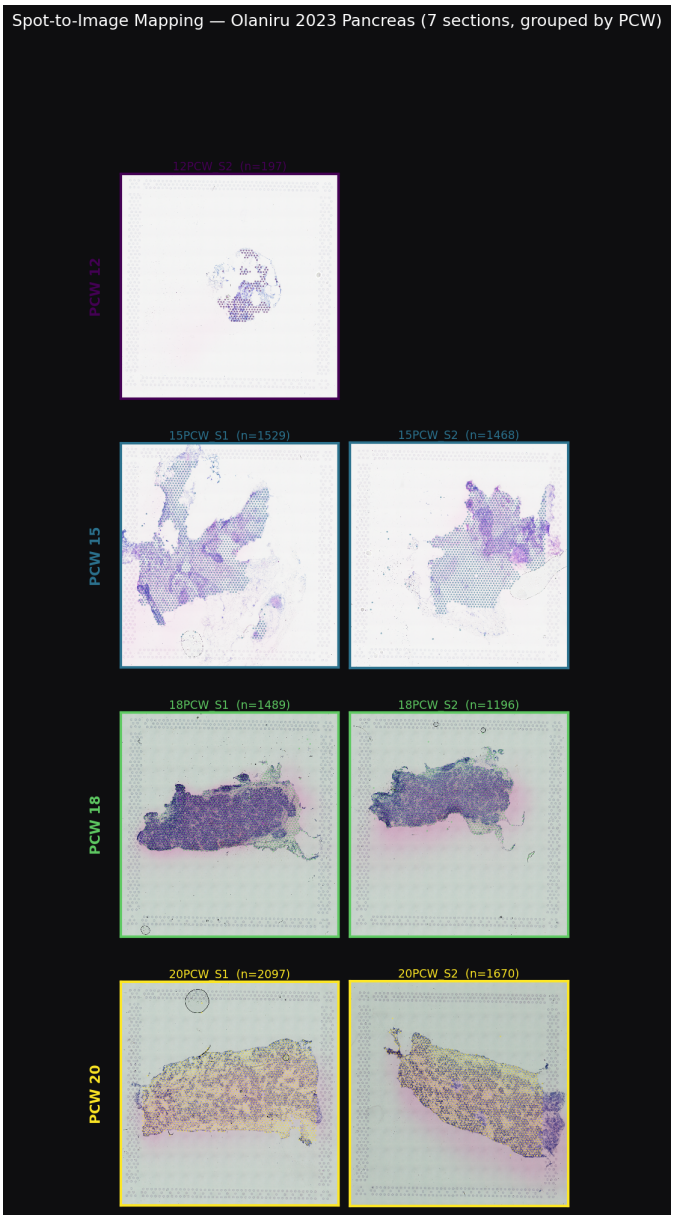

Figure 8: **Pancreas dataset overview.** All-spot overlays per section.

the spot's spatial neighbourhood. The architecture allows scalable inference at whole-slide resolution. STFlow has no temporal memory and is therefore the natural no-memory reference in our temporal protocol; we use the official implementation.

PH2ST (Niu et al., 2025). A prompt-guided hypergraph learning framework for spatial gene-expression prediction from histology, designed for in-slide imputation. Multi-scale histology representations from the test section are guided by a small set of ground-truth ST "prompts" (a few observed spots) through hypergraph message passing, allowing the model to handle realistic partial-observation regimes (missing-spot imputation, super-resolution, local-to-global). It is the strongest recent in-slide method we benchmark against.

NicheFlow (Sakalyan et al., 2026). A flow-matching model for the temporal evolution of spatial transcriptomics. NicheFlow learns a continuous-time velocity field over $(\mathbf{y}, \mathbf{p})$ pairs (the content



Figure 9: **Developing human pancreas (Xenium) overview.** Per-section cell-level expression overlays at sub-cellular resolution.

modality and its spatial coordinate) and integrates it forward in time to model microenvironment trajectories. Crucially, NicheFlow conditions only on the past content modality $\mathbf{y}$ (and position $\mathbf{p}$); it does not condition on the always-available histology modality $\mathbf{x}$. We use it directly to test how far a temporal flow-matching backbone goes without cross-modal conditioning.

NicheFlow+X (ours). To isolate the contribution of the second modality from the contribution of explicit history attention, we introduce an extension of NicheFlow that conditions the velocity field on the histology embedding $\mathbf{x}$ at each time-step (otherwise identical training and architecture). The gap between NicheFlow and NicheFlow+X measures how much the always-available $\mathbf{x}$ helps trajectory recovery; the gap between NicheFlow+X and TEMPUS measures the additional value of explicit, learned history attention on top of cross-modal conditioning.

All baseline methods are trained from scratch on the same slides and timepoints.

Table 7: Real-data results in count space on the Developing human pancreas (Xenium) dataset. Mean ± std over 3 seeds, averaged across held-out timepoints. Hist2ST, MERGE and STFlow rows are blank because their published implementations did not run at this dataset's per-slide cell count; the per-method analysis is in Appendix G.1.

	Developing human pancreas (Xenium)		
Method	PCC↑	MSE↓	MAE↓
Single-section H&E to expression (no temporal memory)			
Hist2ST (Zeng et al., 2022)	–	–	–
BLEEP (Xie et al., 2023)	0.478 ± 0.031	0.739 ± 0.213	0.139 ± 0.008
MERGE (Ganguly et al., 2025)	–	–	–
STFlow (Huang et al., 2025)	–	–	–
TEMPUS (no memory)	0.495 ± 0.004	0.708 ± 0.001	0.115 ± 0.002
Temporal models (use cross-time information)			
NicheFlow (Sakalyan et al., 2026)	0.408 ± 0.042	1.739 ± 0.321	0.339 ± 0.012
NicheFlow+X	0.438 ± 0.019	0.754 ± 0.056	0.161 ± 0.034
TEMPUS (Ours)	**0.535 ± 0.014**	**0.678 ± 0.012**	**0.111 ± 0.002**

Table 8: Per-slide scale of the datasets each baseline's published implementation was developed and validated on, compared to the per-section cell count in Developing human pancreas (Xenium). Hist2ST and MERGE were designed for hundreds of Visium spots per slide; STFlow's training corpus included Xenium but at slide sizes still $\sim 10\times$ smaller than Developing human pancreas.

Method	Original training datasets	Avg. spots / cells per slide
Hist2ST (Zeng et al., 2022)	HER2+ Visium, cSCC Visium	$\sim$180–500 spots
MERGE (Ganguly et al., 2025)	ST-Net (Visium), Her2ST (Visium), SCC (Visium)	$\sim$400–800 spots
STFlow (Huang et al., 2025)	Mixed Visium + Xenium (HEST corpus)	$\sim$4,200–19,000
Developing human pancreas (Xenium)		$\sim$50,000–400,000 cells

G.1 Developing human pancreas (Xenium): results and baseline applicability

Table 7 reports count-space metrics on the Developing human pancreas (Xenium) dataset, the third real benchmark referred to from §3.4 (Table 2). TEMPUS reaches PCC 0.535 vs. 0.438 for the next-best temporal model (NicheFlow+X, $+0.097$) and 0.495 for the strongest no-memory baseline (TEMPUS no memory, $+0.040$); MSE drops from 0.754 (NicheFlow+X) to 0.678 and MAE from 0.161 to 0.111.

The published implementations of Hist2ST, MERGE, and STFlow did not run on this dataset under our compute budget. The reason is a substantial mismatch between the per-slide spot/cell counts each method was designed for and the per-section cell counts in Developing human pancreas. Table 8 summarises the comparison.

Two specific failure modes follow from this gap. (i) *Compute scaling.* Hist2ST and MERGE both build a graph or dense self-attention over all spots in a slide; their compute and memory cost scales at least quadratically in the number of spots. At $\sim$50–400k cells per Developing human pancreas section, the published code paths run out of memory before the first forward pass completes. STFlow's flow-matching backbone tolerates more spots than Hist2ST or MERGE, but its largest validated Xenium slides averaged $\sim 19{,}000$ spots; Developing human pancreas sections are an order of magnitude larger and the official implementation hits OOM.

We therefore leave Hist2ST, MERGE, and STFlow blank in the Developing human pancreas column of Table 2 rather than report numbers from out-of-distribution adaptations. The relevant comparison on this dataset is between the temporal models (NicheFlow, NicheFlow+X, TEMPUS) and the no-history reference TEMPUS (no memory).

Controlled comparison at a scale every method supports. Leaving those rows blank means the full-section table cannot be read as a complete head-to-head, so we additionally run all methods on a

Table 9: Controlled comparison on the Developing human pancreas (Xenium) dataset with every method on the same footing. Sections are subsampled to 10,000 cells each so that Hist2ST, MERGE and STFlow fit in memory; all six timepoints are kept and the last four are held out (long-horizon forecast). All methods share the same UNI image encoder, TERRA embedding space, and frozen count decoder. Mean ± std over 3 seeds, count space.

	Developing human pancreas (Xenium), subsampled		
Method	PCC↑	MSE↓	MAE↓
Single-section H&E to expression (no temporal memory)			
Hist2ST (Zeng et al., 2022)	0.412 ± 0.038	0.785 ± 0.230	0.118 ± 0.011
BLEEP (Xie et al., 2023)	0.432 ± 0.034	0.748 ± 0.210	0.114 ± 0.008
MERGE (Ganguly et al., 2025)	0.414 ± 0.031	0.755 ± 0.205	0.113 ± 0.008
STFlow (Huang et al., 2025)	0.415 ± 0.028	0.745 ± 0.195	0.115 ± 0.007
Temporal models (use cross-time information)			
NicheFlow (Sakalyan et al., 2026)	0.393 ± 0.043	0.810 ± 0.240	0.111 ± 0.012
NicheFlow+X	0.420 ± 0.040	0.772 ± 0.226	0.121 ± 0.010
TEMPUS (Ours)	$\mathbf{0.478 \pm 0.022}$	$\mathbf{0.720 \pm 0.178}$	$\mathbf{0.108 \pm 0.005}$

subsampled version of the same dataset in which every implementation fits in memory. We subsample each section to $10{,}000$ cells, keep all six timepoints, and hold out the last four (a long-horizon forecast); every method uses the same UNI image encoder, the same TERRA embedding space, and the same frozen count decoder, so the only difference between rows is the model. Table 9 reports the result: TEMPUS leads on all three metrics (PCC 0.478 vs. 0.432 for the next-best method, BLEEP). Note that the absolute values are lower than in Table 7 because subsampling removes most of the cells that provide spatial context, so this table should be compared within itself rather than against the full-section numbers.

H Extended ablation results

The main text (§4) reports four headline ablations on Trajectory Memory: learned top-K vs. k-NN, top-K value, cycle-loss weight, and history depth. Here we report the real-data (heart) ablations.

H.1 Real-data ablations (heart)

Crop-radius ablations. The spatial-crop radii r_t (target) and r_h (history) are not interchangeable hyperparameters: r_t defines how many cells the model has to predict in one forward pass, and r_h defines how much spatial context is available from each past time-point. Panels (c)–(d) of Fig. 10 show the count-PCC sweep over both on the heart dataset (mean ± std over seeds, log-x axes). The history-crop radius r_h is mildly better at the smallest values ($r_h \in [100, 200]$ px) but the trend across two decades is shallow ($\sim 0.78 \to 0.77$); the model is fairly robust to history-crop size. The target-crop radius r_t is the more sensitive of the two: there is a clear sweet spot at $r_t = 100$ px (PCC 0.794), and PCC drops by ~ 0.05 as the crop is enlarged toward the full slide ($r_t \geq 500$). Two consequences: (i) larger target crops mix locally heterogeneous tissue within a single training pass, diluting the per-cell supervision signal; (ii) since each target crop pairs with a history crop, growing r_t also forces the history field to cover a region whose niche structure is no longer locally homogeneous, weakening the per-query top-K retrieval. The defaults $r_t = 200$ and $r_h = 500$ used elsewhere in the paper are deliberately conservative; tighter crops would gain a small amount on heart count-PCC at the cost of more crops per epoch.

H.2 Phase Memory ablations

We replicate the headline ablations of §4 on Phase Memory and additionally vary the number of past time-points fed to a fixed trained model at inference. Mean across 3 seeds with seed-std error bars in Fig. 11. Three observations carry over from the Trajectory Memory results: (i) the top-K value has a mid-range optimum, with K=64 giving the best PCC and K=4 being mildly unstable; the curve

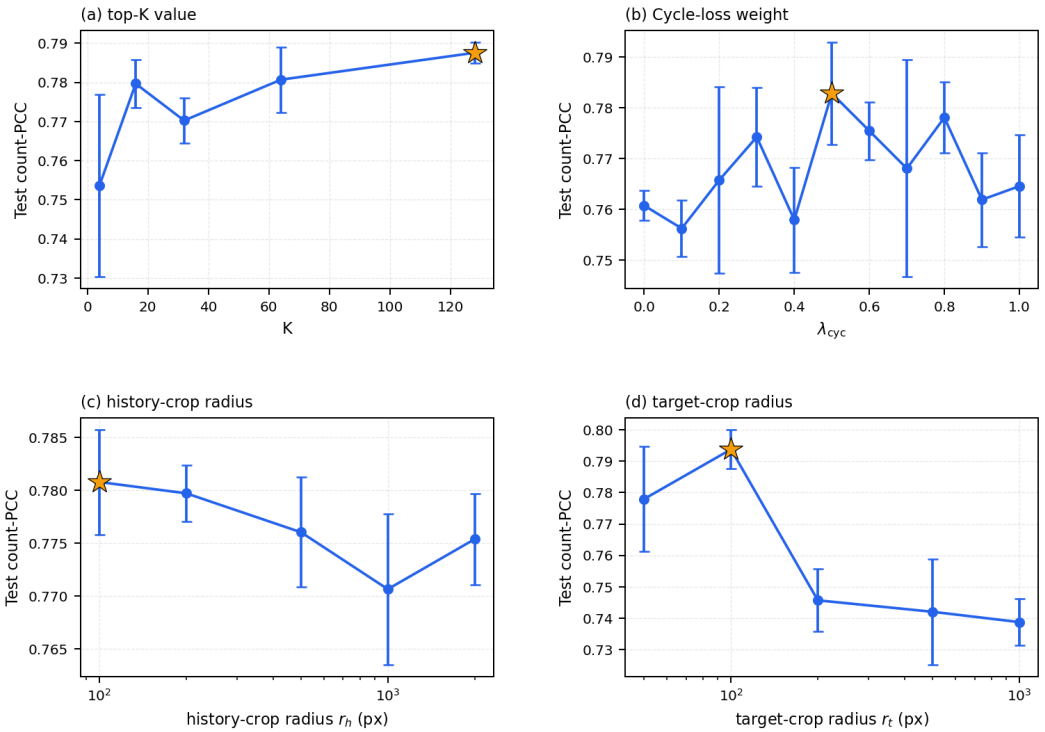

Figure 10: **Heart real-data ablations.** Same protocol as the toy headline figure (Fig. 4) but evaluated in count space (test count-PCC, mean±std over seeds); orange star marks the best setting. (a) On the heart dataset the model is largely insensitive to K across the swept range, with a mild best at the marked value, consistent with the dataset providing relatively dense and locally redundant history. (b) The cycle-consistency loss again prefers small non-zero weight, mirroring the Trajectory Memory trend in Fig. 4(c). (c) History-crop radius r_h: shallow trend, smallest radii are mildly better. (d) Target-crop radius r_t: sharp sweet spot at r_t=100 px, with PCC dropping ~ 0.05 for $r_t \geq 500$.

is broadly flat in $K \in [16, 128]$, indicating that the gain from adding more retrieved cells saturates well below the buffer size. (ii) The cycle-consistency loss is helpful at small non-zero weights and again at $\lambda_{\text{cyc}} \approx 1$, with a non-monotonic trough in between; this is the same shape we observed on Trajectory Memory and on heart, suggesting the cycle objective is most useful as either a light regulariser or a strong reconstruction signal, not somewhere between. (iii) The history-depth panel is monotonically improving across the entire range from 1 to 7 past time-points, with the largest single jump from $1 \to 2$ time-points (PCC $0.69 \to 0.76$) and continued improvement out to the longest history we tested. This is consistent with Phase Memory's structure: each additional past panel provides a new constraint on the rotation phase θ_t, so memory is not just helpful but compositionally additive. Together with the Trajectory Memory headline figure (Fig. 4) and the heart variant (Fig. 10), these results indicate that the same architectural choices win across datasets and that the model's behaviour is robust to dataset structure.

H.3 Naive temporal baselines

A model that merely interpolates between observed timepoints could achieve part of TEMPUS's gain without learning anything about tissue dynamics. We therefore compare against naive temporal baselines that share the same frozen encoders and the same frozen count decoder, so that the only difference is how the prediction is formed from the history: last-observed expression, temporal k-NN retrieval, linear and spline extrapolation in embedding space, and PCA interpolation. We additionally ablate TEMPUS's two inputs by removing the target H&E and by removing the past ST history. Table 10 shows that every one falls well short of TEMPUS, with the largest margins on the two synthetic benchmarks constructed so that no single timepoint suffices (0.871 vs. at best 0.427 on

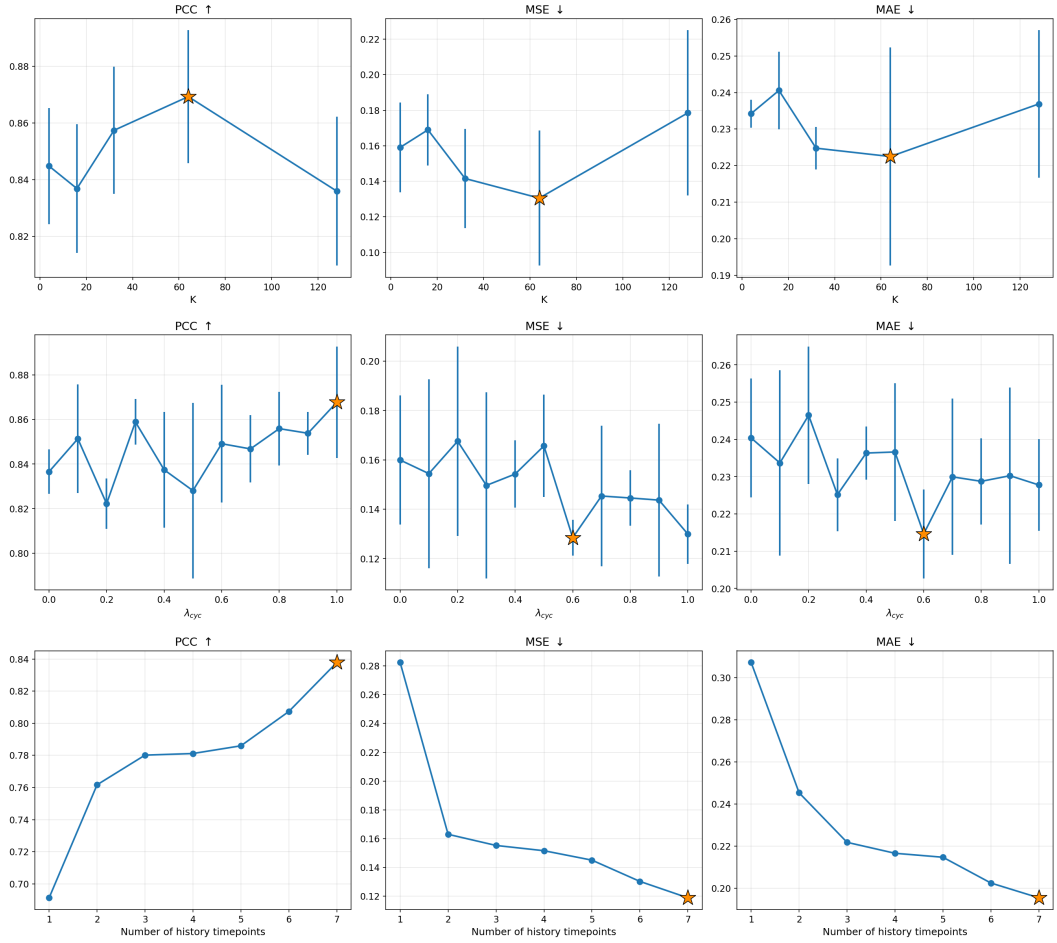

Figure 11: **Phase Memory ablations.** Mean across 3 seeds; bars are seed std. *Top row*: top-K value $K \in \{4, 16, 32, 64, 128\}$ (PCC / MSE / MAE). *Middle row*: cycle-consistency loss weight $\lambda_{\text{cyc}} \in \{0.0, \ldots, 1.0\}$. *Bottom row*: history-depth at inference, MSE / MAE / PCC against number of past time-points fed to the trained model.

Phase Memory; 0.883 vs. at best 0.512 on Trajectory Memory). Removing either input also degrades the model sharply, so both the target morphology and the past molecular history are load-bearing.

H.4 Complete results table

Table 11 in the main text reports the methods that were run on all three real datasets, so that every cell is filled. Table 11 gives the complete version, additionally including Hist2ST and STFlow on Human heart and Human pancreas.

H.5 Per-timepoint results on Human heart

Aggregate means can hide gains concentrated in a subset of sections, so we disaggregate the Human heart results by held-out timepoint (Table 12). Each held-out timepoint is a single section, so no averaging over sections occurs within a row and the final row is the aggregate reported in Table 2. TEMPUS is best at every timepoint and improves over its own no-memory variant at all five, so the gain is uniform rather than driven by a subset. It is also markedly steadier than the single-section baselines: TEMPUS varies by 0.143 PCC across timepoints (0.714–0.857) with seed s.d. ≤ 0.012, whereas MERGE spans 0.591–0.851 with seed s.d. up to 0.043.

Table 10: Naive temporal baselines (test PCC↑). Every row uses the same frozen UNI image encoder, the same TERRA embedding space, and the same frozen count decoder as TEMPUS, so the only difference is how the prediction is formed from the available history. The last two rows ablate TEMPUS's two inputs. MSE/MAE follow the same ordering.

Method	Phase Memory	Trajectory Memory	Xenium
Naive temporal baselines			
Last-observed mean	0.367	0.429	0.362
Temporal k-NN	0.352	0.512	0.454
Linear extrapolation	0.236	0.439	0.376
Spline extrapolation	0.310	0.423	0.337
PCA interpolation	0.236	0.439	0.382
Input ablations of TEMPUS			
TEMPUS − target H&E	0.341	0.434	0.412
TEMPUS − past ST	0.427	0.310	0.472
TEMPUS (Ours)	**0.871**	**0.883**	**0.535**

Table 11: Real-data results in count space on three datasets: Human heart, Human pancreas, and a kidney ischemia–reperfusion injury (IRI) time course. Mean ± std over 3 seeds, averaged across held-out timepoints. Baseline citations are given in Table 1. Entries marked "–" were not run on the kidney IRI time course.

	Human heart			**Human pancreas**			**Kidney IRI**		
Method	PCC↑	MSE↓	MAE↓	PCC↑	MSE↓	MAE↓	PCC↑	MSE↓	MAE↓
Single-section H&E to expression (no temporal memory)									
Hist2ST	0.034 ± 0.005	7.965 ± 0.000	1.028 ± 0.000	0.244 ± 0.004	26.61 ± 0.01	1.142 ± 0.002	–	–	–
BLEEP	0.671 ± 0.011	4.828 ± 0.101	0.369 ± 0.002	0.805 ± 0.005	17.72 ± 0.20	**0.448 ± 0.004**	0.474 ± 0.010	0.142 ± 0.005	0.180 ± 0.004
MERGE	0.698 ± 0.026	4.681 ± 0.253	0.371 ± 0.008	0.805 ± 0.001	18.92 ± 1.84	0.561 ± 0.024	0.461 ± 0.011	0.147 ± 0.004	0.184 ± 0.005
STFlow	0.494 ± 0.008	6.747 ± 0.022	0.342 ± 0.000	0.719 ± 0.005	22.53 ± 0.24	0.465 ± 0.010	–	–	–
TEMPUS (no memory)	0.738 ± 0.008	4.218 ± 0.081	0.331 ± 0.001	0.788 ± 0.004	17.05 ± 0.33	0.490 ± 0.022	0.480 ± 0.003	0.132 ± 0.003	0.171 ± 0.003
Temporal models (use cross-time information)									
NicheFlow	0.613 ± 0.037	5.930 ± 0.360	0.365 ± 0.004	0.760 ± 0.008	27.19 ± 1.20	0.638 ± 0.012	0.431 ± 0.013	0.149 ± 0.006	0.193 ± 0.006
NicheFlow+X	0.705 ± 0.003	4.529 ± 0.137	0.353 ± 0.002	0.781 ± 0.003	22.96 ± 1.30	0.594 ± 0.018	0.463 ± 0.014	0.143 ± 0.005	0.185 ± 0.005
TEMPUS (Ours)	**0.794 ± 0.006**	**3.680 ± 0.082**	**0.318 ± 0.002**	**0.812 ± 0.006**	**16.32 ± 0.07**	0.477 ± 0.018	**0.512 ± 0.008**	**0.129 ± 0.003**	**0.165 ± 0.003**

H.6 Encoder-matched comparison

A natural concern is that TEMPUS benefits from its frozen foundation-model embeddings rather than from temporal modelling. To remove that confound we place every method on one pipeline: the same UNI image encoder, the same TERRA embedding space as the prediction target, and the same frozen count decoder, so that the only difference between rows is the model itself. Table 13 reports the result on Human heart. TEMPUS retains its lead (PCC 0.794 vs. 0.732 for the strongest matched baseline, STFlow–TERRA), so the advantage is attributable to the temporal model and not to the representation pipeline.

H.7 Component ablation

Table 14 removes one component at a time, holding everything else fixed. Each part earns its place. Replacing the frozen embeddings with direct count prediction costs the most on real data ($0.794 \to 0.664$ with an MSE loss, 0.581 with an NB loss). Removing the cross-attention decoder collapses Trajectory Memory entirely ($0.883 \to -0.050$), which is expected: TM's target is a per-niche cumulative trajectory that cannot be recovered without fusing information across time-points. Replacing top-K selection with full attention over all history tokens is *worse*, not merely slower ($0.794 \to 0.709$ on heart; $0.883 \to 0.805$ on TM), so the selection mechanism improves quality rather than only cost.

Table 12: Per-timepoint results on Human heart (count-space PCC↑, mean ± std over 3 seeds; train PCW 6.0–8.0, test PCW 9.0–12.0). Each held-out timepoint is a single section, so these are per-section values and their mean is the aggregate reported in Table 2. TEMPUS is best at every timepoint and improves over its no-memory variant at all five.

Timepoint (= section)	TEMPUS (Ours)	TEMPUS (no mem.)	MERGE	BLEEP
PCW 9.0	**0.788 ± 0.012**	0.762 ± 0.008	0.687 ± 0.016	0.706 ± 0.006
PCW 10.0	**0.857 ± 0.010**	0.774 ± 0.011	0.591 ± 0.043	0.624 ± 0.021
PCW 10.5	**0.855 ± 0.004**	0.816 ± 0.004	0.851 ± 0.005	0.777 ± 0.004
PCW 11.0	**0.714 ± 0.008**	0.692 ± 0.005	0.660 ± 0.042	0.655 ± 0.010
PCW 12.0	**0.754 ± 0.003**	0.647 ± 0.018	0.700 ± 0.040	0.592 ± 0.046
Mean	**0.794 ± 0.006**	0.738 ± 0.008	0.698 ± 0.026	0.671 ± 0.011

Table 13: Encoder-matched comparison on Human heart. All methods use the same UNI image encoder, predict into the same TERRA embedding space, and decode through the same frozen count decoder; the only difference between rows is the temporal model. Count space, mean ± std over 3 seeds.

Method (all on UNI + TERRA + shared decoder)	PCC↑	MSE↓	MAE↓
BLEEP–TERRA (Xie et al., 2023)	0.716 ± 0.100	3.996 ± 0.200	0.365 ± 0.009
Hist2ST–TERRA (Zeng et al., 2022)	0.724 ± 0.011	3.928 ± 0.012	0.353 ± 0.001
MERGE–TERRA (Ganguly et al., 2025)	0.707 ± 0.200	4.533 ± 0.023	0.388 ± 0.001
STFlow–TERRA (Huang et al., 2025)	0.732 ± 0.100	3.910 ± 0.014	0.340 ± 0.001
TEMPUS (Ours)	**0.794 ± 0.006**	**3.680 ± 0.082**	**0.318 ± 0.004**

H.8 Inference-time complexity and scaling

Because retrieval is confined to bounded local crops, inference is linear in cell count with memory independent of slide size: $\mathcal{O}(N\,T\,(M{+}K)\,d)$ compute and $\mathcal{O}(N\,(K{+}T)\,d)$ memory with $K{=}32$. Table 15 confirms this empirically: peak memory stays flat at 0.12 GB from $10{,}000$ to $190{,}000$ cells while runtime grows linearly, so a whole $190{,}000$-cell Xenium section completes in ~ 3.6 s at a memory footprint where the single-section graph baselines run out of memory.

H.9 Reconstruction ceiling of the frozen count decoder

All count-space metrics in this paper pass through one frozen Negative-Binomial decoder, shared by every method and ablation. Because it is shared, decoder error is a common constant across rows and cannot explain differences between them. To bound its effect we decode *ground-truth* embeddings to counts and measure the resulting PCC, which upper-bounds what any method predicting in this space can reach (Table 16). On Human heart the test ceiling (0.871) is close to the train ceiling (0.909), a gap of ≈ 0.04 PCC, indicating the decoder generalises rather than overfitting the training sections. TEMPUS (0.794) sits below the ceiling with 0.077 PCC of headroom rather than saturating it, so the decoder is not the source of the reported gains.

I Inspecting the attention map of a target cell on Phase Memory

We pick one held-out target cell at the test time-point and inspect its attention map: which past cells the model attended to, and at which past time-points. The Phase Memory benchmark is convenient because its data-generating process gives an unambiguous notion of useful history, so we can compare the model's attention against what the dataset structure actually requires.

What the model should do, from the dataset construction. On Phase Memory (§E), the content target is

$$\mathbf{y}_i^{(t)} \;=\; \alpha\, W_{\mathrm{sh}}\, \mathbf{x}_i^{(t)} \;+\; (1-\alpha)\, W_{m_i^{(t)}}\, \mathbf{x}_i^{(t)} \;+\; \boldsymbol{\eta}_y,$$

Table 14: Component ablation. Each row removes a single component from the full model, keeping everything else fixed. Left: developing Human heart (9,021 genes, count space). Right: Trajectory Memory. Mean ± std over 3 seeds where available.

	Human heart			Trajectory Memory		
Ablation	PCC↑	MSE↓	MAE↓	PCC↑	MSE↓	MAE↓
Full TEMPUS	**0.794 ± 0.006**	**3.680 ± 0.082**	**0.318 ± 0.002**	**0.883 ± 0.025**	**0.023 ± 0.008**	**0.117 ± 0.021**
− Frozen embeddings (count, MSE loss)	0.664	5.465 ± 0.087	0.332 ± 0.004	–	–	–
− Frozen embeddings (count, NB loss)	0.581	6.013 ± 0.096	0.348 ± 0.005	–	–	–
− Learned top-K retrieval	0.742 ± 0.007	4.07 ± 0.090	0.347 ± 0.003	0.849 ± 0.025	0.031 ± 0.009	0.144 ± 0.022
− Cross-attention decoder	0.713 ± 0.008	4.39 ± 0.105	0.347 ± 0.006	−0.050± 0.020	0.211 ± 0.010	0.376 ± 0.020
− Cycle-consistency	0.735 ± 0.007	4.20 ± 0.092	0.352 ± 0.003	0.823 ± 0.025	0.025 ± 0.008	0.124 ± 0.021
− Top-K selection (full attention)	0.709 ± 0.008	4.40 ± 0.100	0.349 ± 0.004	0.805 ± 0.025	0.038 ± 0.009	0.158 ± 0.022

Table 15: Inference-time scaling. Retrieval is performed inside bounded local crops, so compute is $\mathcal{O}(N\,T\,(M{+}K)\,d)$ and memory $\mathcal{O}(N\,(K{+}T)\,d)$ with $K{=}32$: peak memory is independent of slide size while runtime grows linearly in cell count. Measured on one A100 with $d{=}128$, $T{=}3$, and 1000-cell crops.

Slide size (cells)	Crops	Total time	Peak GPU memory
10,000	10	57 ms	0.12 GB
50,000	50	0.28 s	0.12 GB
190,000 (whole Xenium section)	190	3.6 s	0.12 GB

where $m_i^{(t)}$ is the mode of point i at time t and $\alpha = 0.3$. The mode depends on the angle of the point relative to the slide centre $\mathbf{c}$ and on the rotation phase $\theta_t = 2\pi t/T$:

$$m_i^{(t)} \;=\; \Big\lfloor \frac{\big((\angle(\mathbf{p}_i^{(t)} - \mathbf{c}) - \theta_t) \bmod 2\pi\big)}{2\pi/M} \Big\rfloor .$$

The embedding $\mathbf{x}_i^{(t)}$ encodes only position, not the rotation phase, so a model that sees $\mathbf{x}^{(t^\star)}$ alone cannot recover $m^{(t^\star)}$ and therefore cannot resolve the mode-specific term $(1-\alpha)\,W_{m_i^{(t^\star)}}\,\mathbf{x}_i^{(t^\star)}$. Two pieces of information from past observations identify $\theta_{t^\star}$ uniquely: (i) the canonical mode field at $t{=}0$, where $\theta_0 = 0$, anchors the rotation in absolute terms; (ii) one or more later observations let the model estimate the rotation rate $\omega = 2\pi/T$. Equivalently, looking at past cells in a fixed spatial neighbourhood and tracking how the mode at that location evolves through time gives a local time-series of mode labels from which $\theta_{t^\star}$ can be read off. A useful attention map at a held-out target cell should reflect one or both of these strategies.

What the attention map shows. Figure 12 shows the attention map of a single held-out target cell at $t^\star = 8$, on which the model achieves a test PCC of 0.96 and MSE of 0.0205. The time-attention plot (right) is biased toward the earliest history step: $t{=}0$ carries the largest mass (0.191), $t{=}1$ next (0.161), $t{=}2$ third (0.128), and the remaining steps $t{=}3,\ldots,7$ collapse to a near-uniform ~ 0.10 each. The spatial attention map (left) is localised: across all eight past panels, the cells the model attended to concentrate on a region rather than spreading across the slide.

The localised spatial attention amounts to building a time-series of mode labels at the target cell's position, an alternative path to recovering $\theta_{t^\star}$ that is internally consistent with the temporal anchoring.

Why this matters. The fact that the attention shape adapts to the task structure ($t{=}0$-anchored on Phase Memory, recency-weighted on Trajectory Memory) is interesting. It is evidence that the model is responding to the data-generating process rather than defaulting to a fixed temporal prior.

J Failure-mode attention on Trajectory Memory

We complement the best-query analysis in §3.5 with the corresponding failure case. Within the same held-out evaluation pool on Trajectory Memory, we pick the query with the lowest cosine similarity between TEMPUS prediction and ground truth and inspect its top-K attention. Figure 13 shows the

Table 16: Reconstruction ceiling of the frozen count decoder on Human heart. The ceiling is obtained by decoding *ground-truth* embeddings to counts, so it upper-bounds what any method predicting in this embedding space can reach. The train/test gap is small (≈ 0.04 PCC), indicating the decoder generalises rather than overfitting the training sections, and TEMPUS sits below the ceiling with positive headroom rather than saturating it.

Dataset	Ceiling (train)	Ceiling (test)	TEMPUS (test)	Headroom
Human heart	0.909	0.871	0.794	0.077

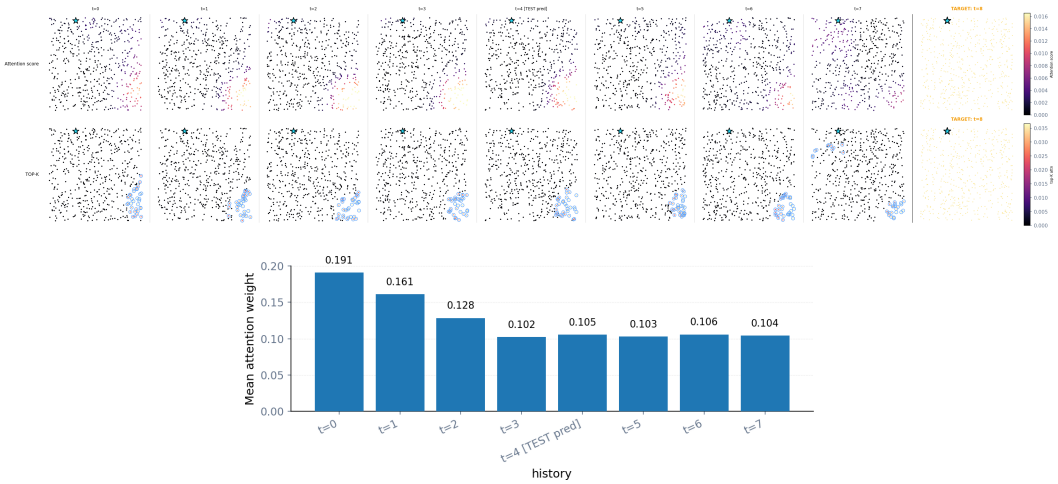

Figure 12: **Attention map of a single held-out target cell on Phase Memory.** The model achieves a test PCC of 0.96 and MSE of 0.0205 on this query. Spatial attention (top): across the eight past panels, the cells the model attended to concentrate on a localised region rather than spreading across the slide. Time attention (bottom): the model places the dominant weight on the earliest past step (t=0, mass 0.191), with secondary weight on t=1 (0.161), and a near-uniform ~ 0.10 weight thereafter. Both are consistent with the dataset's structure: t=0 is the canonical reference frame for the rotation phase θ_t, and a localised spatial neighbourhood gives a time-series of mode labels from which $\theta_{t^\star}$ can be inferred (§E).

worst-cosine held-out query, on which the model achieves a test PCC of 0.21 and MSE of 0.0355. The query (cyan star) sits in the blue niche at the target time, but the top-K circles at each past panel land on a cluster that is orange at that past panel, not blue. The time-attention plot (right) reflects this collapse: instead of the recency-weighted profile of the best query, the four past times each receive roughly the same mass ($\sim$0.21–0.30, almost flat). When the spatial pick is wrong, the model has no informative signal in any single past time and effectively averages uniformly over history, a hallmark of a confused query.

Aggregate retrieval quality. The two preceding figures are single queries, so we also quantify retrieval over the whole held-out pool. Trajectory Memory provides ground-truth niche identity, which lets us check directly whether the learned top-K selects history cells from the correct niche: it does so for 74.3% of retrieved cells, against a 33% chance rate for the three niches. Retrieval quality also predicts accuracy across held-out queries, the fraction of same-niche retrievals correlates negatively with prediction error (Spearman $\rho = -0.58$, $p = 4 \times 10^{-8}$), i.e. queries whose retrieval is more often correct are the ones predicted best. This supports the interpretability claim at the level of the population rather than selected examples.

K Additional spatial-temporal gene expression results

Predicting exocrine dynamics. We additionally examined SPINK1 (serine peptidase inhibitor Kazal type 1), a marker of early exocrine progenitors and acinar differentiation that is highly

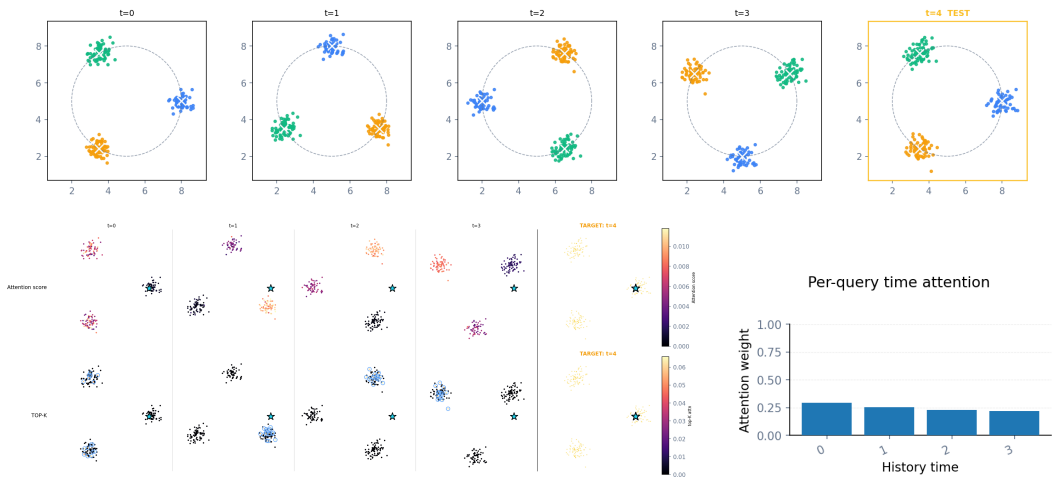

Figure 13: **Worst-cosine held-out query on Trajectory Memory.** *Top*: dataset reference (same as Fig. 2, top), shown again for comparison with the worst-case attention below. *Bottom*: spatial (left) and time (right) top-K attention for the held-out target query (cyan star).

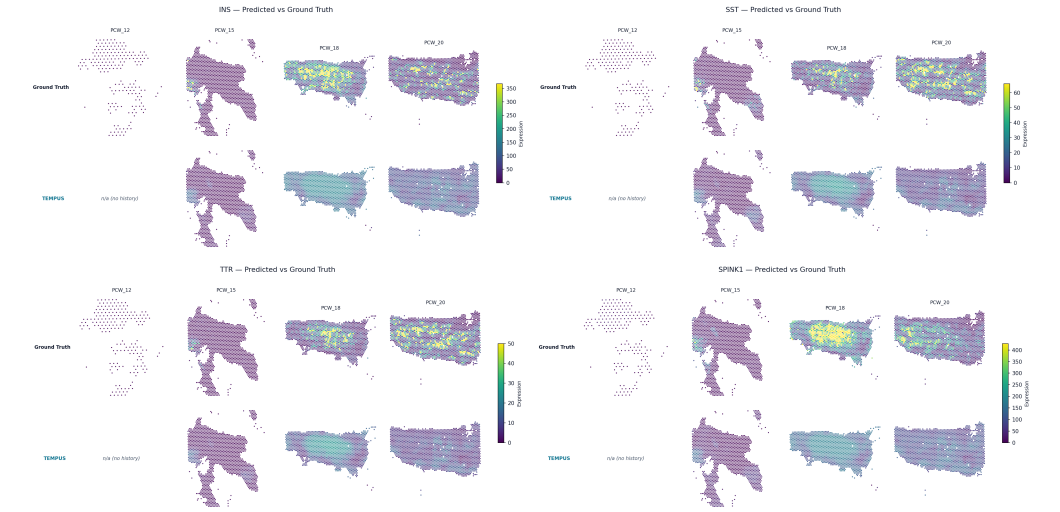

Figure 14: **Spatial expression maps of endocrine and exocrine markers across pancreatic development.** Each row shows the spatial distribution of gene expression across the tissue section at each timepoint. **Ground truth** (top): observed expression at PCW 12, 15, 18, and 20. **Model prediction** (bottom): predicted expression at the held-out PCW 20 slide, conditioned on histology image embeddings and temporal memory from PCW 12–18.

expressed throughout mid-gestation. Unlike the endocrine markers, SPINK1 represents a different tissue compartment, namely the acinar parenchyma, and exhibits sustained high expression rather than a rising trajectory. The model correctly predicts that SPINK1 remains highly expressed at PCW 20 (Fig. 3), and the spatial prediction maps (Fig. 14) place the predicted high-expression signal in the acinar region of the tissue, spatially distinct from the predicted endocrine signal. This demonstrates that the model has learned to distinguish cell-type-specific spatial niches from the histology image rather than producing a single undifferentiated prediction for the entire section. Taken together, these results show that our method captures both the endocrine maturation dynamics (lineage commitment and hormonal upregulation from PCW 12 to 20) and the sustained exocrine programme (acinar maintenance), and correctly resolves their distinct spatial territories, all from a held-out timepoint it was never trained on.

To complement the pancreas spatial maps in Figure 14, we additionally evaluate the per-spot spatial fidelity of TEMPUS on the human heart Visium dataset (Lázár et al., 2025). The model is trained on the early developmental window (PCW 6.0, 7.0, 7.5, 8.0) and queried on every later test timepoint (PCW 9.0, 10.0, 10.5, 11.0, 12.0). In Figure 15 each row corresponds to one of seven canonical heart-development markers; within a row the upper strip shows ground-truth expression and the lower strip the TEMPUS prediction at every timepoint. PCW 6.0 has no temporal history available and is shown as a blank panel; PCW 7.0–8.0 are training slides shown for reference, and PCW 9.0–12.0 are the held-out test slides predicted from histology embeddings and temporal memory of the training slides only. The seven genes span two complementary tissue compartments: *cardiomyocyte* markers (TNNI3, MYL2, IRX4) report contractile maturation and chamber identity, becoming progressively restricted to the developing ventricular wall; *stromal/extracellular-matrix* markers (COL1A2, ELN, PDGFRA) report fibroblast and matrix-deposition activity in the cardiac interstitium and outflow-tract mesenchyme; TMSB10 is included as a broadly expressed cytoskeletal marker that is expected to remain high across compartments. TEMPUS recovers the spatial localisation of each marker on the held-out timepoints rather than producing a uniform tissue-wide prediction: cardiomyocyte markers are concentrated in the ventricular myocardium, stromal markers occupy the surrounding mesenchyme and ECM-rich regions, and the predicted maps track the ground truth through the PCW 9.0–12.0 progression. This indicates that the same conditioning mechanism that resolves endocrine versus exocrine territory in the pancreas also separates cardiomyocyte from stromal territory in the heart, and that the temporal memory propagates spatial structure into developmental stages the model never observed during training.

L Reproducibility

Default hyperparameters are listed in Appendix D (Table 4); the architecture is described in Appendix C. All code, configs, and experiment scripts are submitted as a supplementary zip file alongside this manuscript. Each model is trained with three different random seeds and evaluated independently per seed; reported numbers are mean $\pm$ std over the three runs. Each run was performed on a single NVIDIA H200 GPU; per-seed wall-clock is approximately 10 minutes for the synthetic runs and 30 minutes for the real-data (heart/pancreas) runs.

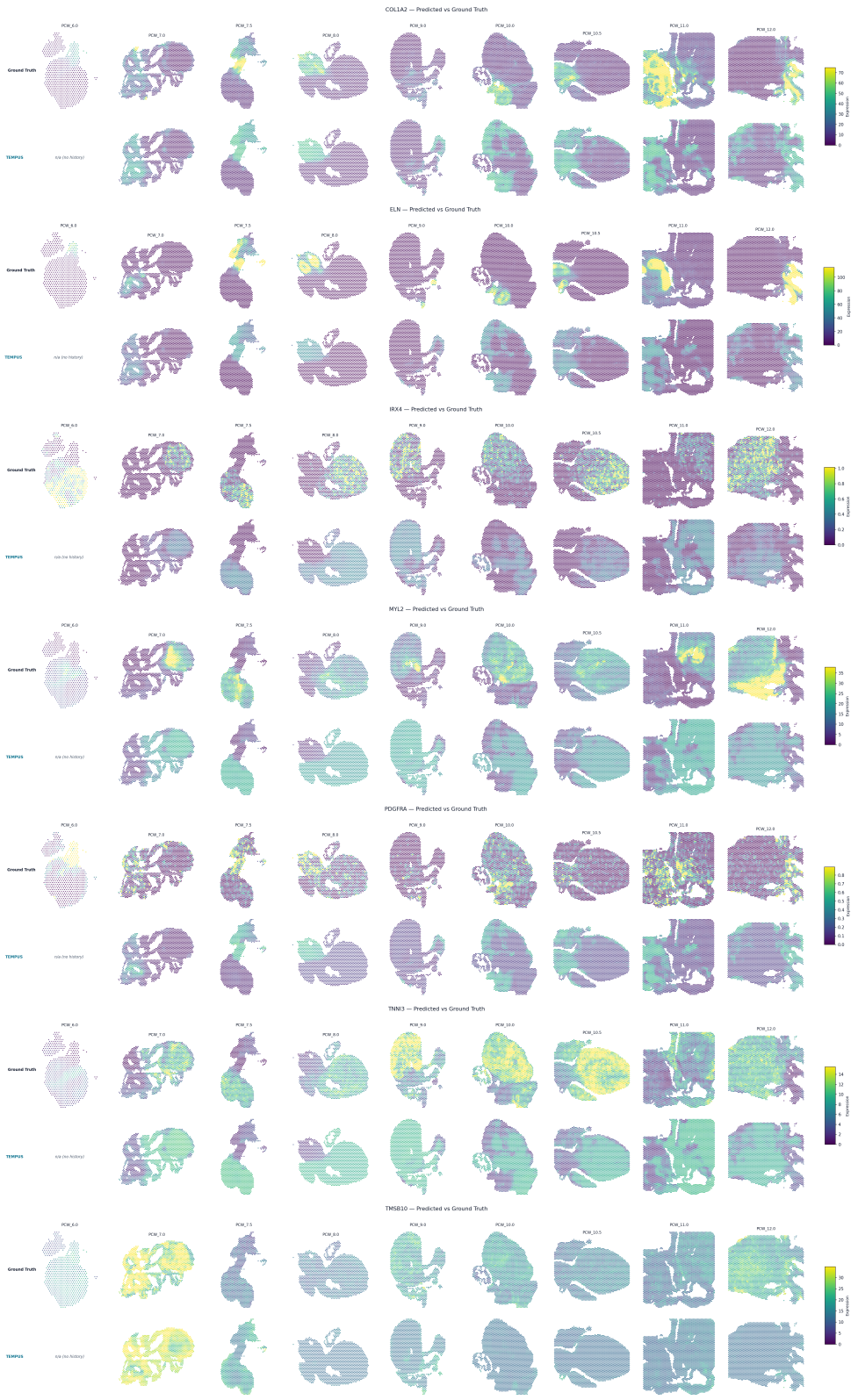

Figure 15: Spatial expression maps of seven heart-development markers (COL1A2, ELN, IRX4, MYL2, PDGFRA, TNNI3, TMSB10) on the human heart Visium dataset. Each row: ground truth (top strip) vs. TEMPUS prediction (bottom strip) across PCW 6.0–12.0. Train: PCW 6.0–8.0; held-out test: PCW 9.0–12.0. See §K for compartment grouping and discussion.