HoloGene: Learning to Lift Sliced Spatial Transcriptomics to Holistic 3D Gene Fields

Yumin Zheng[1,2,*] **Qingtian Zhu**[3,*] **Ziyan Zhu**[4]
Kailu Song[1,2] **Yinqiang Zheng**[3,†] **Jun Ding**[1,2,5,†]
[1]McGill University [2]MUHC Research Institute [3]The University of Tokyo
[4]Carnegie Mellon University [5]Mila – Quebec AI Institute

Abstract

Spatial transcriptomics (ST) provides spatially resolved gene expression measurements, but most 3D ST datasets are acquired as independently processed 2D sections. Computational approaches span slice registration and integration, explicit or generative interpolation, and coordinate-field modeling. Within the last family, implicit neural representations (INRs) offer a continuous, coordinate-queryable parameterization, but standard isotropic coordinate encodings do not explicitly reflect that serial-section measurements are dense in-plane and sparse along the axial direction. We present **HoloGene**, an anisotropic conditional INR framework for estimating continuous 3D gene-expression fields from pre-registered ST sections. Its central contribution is an acquisition-aware factorization that couples dense in-plane synthesis with sparse axial evolution using established GAE, FiLM, pseudo-label supervision, and CORAL components. HoloGene compresses high-dimensional expression profiles into a graph-autoencoder latent manifold, uses pseudo-label-derived signature conditioning to provide coarse intra-section context, and applies depth-dependent FiLM modulation to model axial variation. A signature-stratified correlation-alignment regularizer encourages cross-slice covariance alignment in latent features. Using held-out measured sections as cross-sectional references, HoloGene improves reconstruction accuracy and biological preservation metrics compared with general-purpose INRs and adapted spatial-transcriptomics baselines. These results support anisotropic conditional modeling as a practical strategy for estimating 3D expression fields from discontinuous ST sections. Code is available at `https://github.com/Szym29/HoloGene`.

1 Introduction

Spatial Transcriptomics (ST) transforms biological tissue analysis by mapping high-dimensional gene expression profiles directly onto their physical coordinates. Advancing sequencing technologies are shifting the research focus from isolated 2D slices toward the construction of comprehensive 3D tissue atlases. This transition is essential for studying biological processes that are inherently three-dimensional, such as organogenesis and the volumetric organization of complex tissues (Chen et al., 2022; Wang et al., 2018; Zhang et al., 2023).

Computational approaches to 3D ST include slice registration and integration (Zeira et al., 2022; Qiu et al., 2024), explicit or generative interpolation (Lin et al., 2025; Li et al., 2024), and Implicit Neural Representations (INRs) (Zhu et al., 2025; Luo et al., 2025). Among these complementary families, INRs parameterize biological signals as continuous functions encoded within the weights of a neural network, typically a multilayer perceptron (MLP). By learning a mapping from spatial coordinates to gene expression profiles, INRs support continuous coordinate queries without requiring a predefined

* Equal contribution. † Corresponding authors: `yqzheng@ai.u-tokyo.ac.jp`, `jun.ding@mcgill.ca`.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

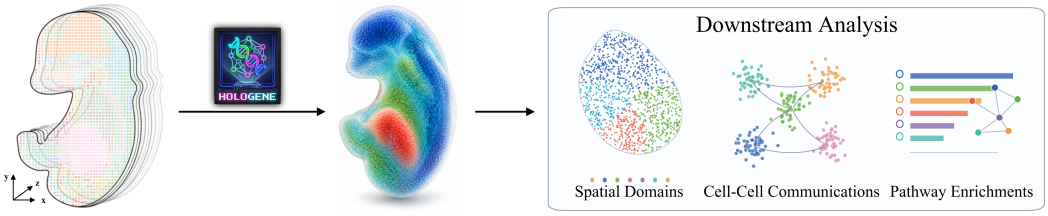


Figure 1: HoloGene reconstructs holistic 3D gene fields from discontinuous ST sections. By lifting sparsely sampled 2D slices into a continuous 3D field, HoloGene enables volumetric downstream analysis, including spatial-domain recovery, cell–cell communication inference, and pathway enrichment.

output grid. This is an output parameterization choice rather than a substitute for registration, integration, or experimentally measured intermediate sections.

However, reconstructing such a continuous 3D gene expression field is hindered by a duality of structural bottlenecks: **physical anisotropy** and **data heterogeneity**. The prevailing data acquisition workflow relies on mechanical sectioning, producing measurements that are dense and quasi-continuous in the XY-plane but sparse and discrete along the Z-axis. The physical spacing between slices can range from a few micrometers to hundreds of micrometers, depending on the experimental design, whereas measurements within each slice are much denser. This axial under-sampling creates a strong scale mismatch between the dense in-plane measurements and the sparse depth direction. This severe scale disparity induces topological discontinuities, since biological structures may shift nonlinearly between adjacent sections. Compounding this challenge are **inter-slice batch effects**: because slices are sequenced independently, technical variations across sections can masquerade as genuine biological signals. Consequently, standard coordinate INRs that encode (x, y, z) symmetrically (Sitzmann et al., 2020; Tancik et al., 2020) do not explicitly reflect the sampling asymmetry of serial ST, while models without a cross-slice alignment objective may retain slice-specific technical variation.

To address these acquisition-specific limitations, we propose **HoloGene**, an anisotropic conditional INR framework tailored to serial-section ST, where dense intra-slice measurements coexist with sparse, independently acquired sections along the depth axis. The central contribution is an acquisition-aware factorization that integrates established GAE, FiLM, pseudo-label supervision, and CORAL components around serial-section geometry. The synthesis network treats (x, y) as coordinates of a dense in-plane field, while z supplies layer-wise conditioning for axial variation. A graph autoencoder provides compact latent regression targets. The Pioneer Signature Network distills Leiden partitions derived from training-section expression into a coordinate-predictable conditioning signal, and signature-stratified correlation alignment reduces covariance shifts among comparable coarse tissue states across slices. This design produces coordinate-queryable estimates within the supplied pre-registered domain. Predictions at held-out measured sections can be compared with observed expression, whereas queries between measured sections are model-based interpolation estimates. Under this evaluated protocol and across the included spatial technologies and tissue types, HoloGene achieves the best result on most aggregate reconstruction and biological-preservation metrics among the compared methods.

2 Related Works

2.1 3D Spatial Transcriptomics Reconstruction

Reconstructing volumetric tissue architecture from anisotropic serial sections is a fundamental inverse problem in spatial biology. Traditional approaches primarily rely on rigid or non-rigid registration algorithms to align adjacent slices. Zeira et al. (2022) formulate slice alignment as an Optimal Transport (OT) problem in PASTE, matching spots across layers based on transcriptional similarity and spatial proximity. More recent frameworks shift toward graph-based integration and generative interpolation to bridge this dimensional gap. For example, Qiu et al. (2024) employ Graph Neural Networks (GNNs) for multi-slice alignment in Spateo, whereas Lin et al. (2025) fill gaps along the Z-axis in SpatialZ and construct dense 3D cell atlases by generating virtual intermediate slices. These approaches provide complementary strengths in registration, integration, or virtual-slice generation. Coordinate-field models use a different output parameterization, supporting expression queries at

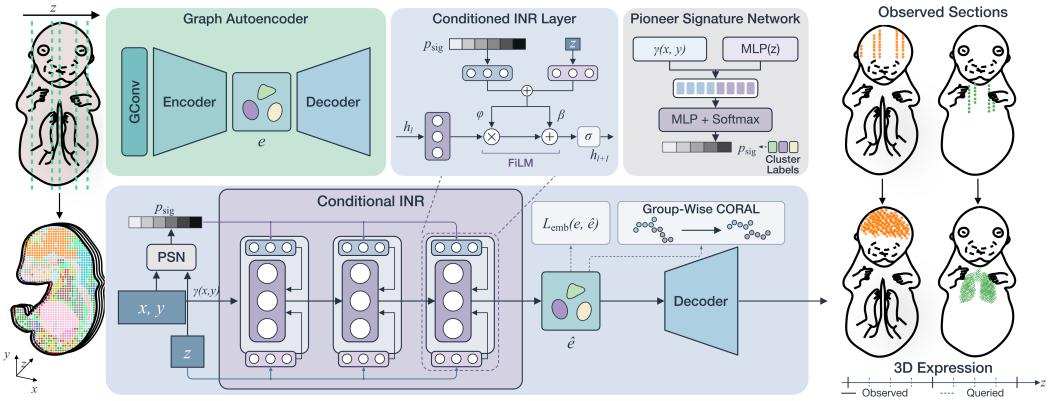


Figure 2: Overview of HoloGene. We first compress high-dimensional gene expression into latent embeddings via a graph autoencoder, then use PSN to distill training-expression partitions into coarse signature guidance, and synthesize a continuous 3D latent field using a decoupled, FiLM-conditioned INR with CORAL regularization to handle inter-slice anisotropy and regularize slice-specific representation variation.

locations within an aligned coordinate domain. This capability complements rather than replaces upstream registration or evidence from experimentally measured sections. Furthermore, standard batch correction techniques (*e.g.*, Harmony (Korsunsky et al., 2019), Seurat (Hao et al., 2021)) are typically decoupled from the spatial geometry but applied as preprocessing steps, potentially discarding spatially meaningful heterogeneity.

2.2 Representation Learning for ST

Owing to the high dimensionality and zero-inflated nature of ST data, representation learning plays a central role in extracting robust biological signals. Discriminative models, such as SpaGCN (Hu et al., 2021) and GraphST (Long et al., 2023), are widely adopted, leveraging spatial neighborhood graphs to smooth gene expression and identify spatial domains. Beyond discriminative models, generative approaches are increasingly used for tasks such as data imputation and super-resolution. Notably, Li et al. (2024) employ generative modeling in STAGE for conditional recovery of high-density gene expression from low-resolution measurements, while Zhu et al. (2025) use Graph Autoencoders (GAEs) in SUICA to compress whole-transcriptome profiles into compact latent embeddings. Our framework builds upon these representation-learning advances by adopting GAE-based compression to obtain compact latent regression targets. STAGE, SUICA, and related approaches use different supervision schemes and output parameterizations for their respective reconstruction settings. Our focus is the complementary setting of a specimen-specific coordinate field whose in-plane synthesis and axial conditioning reflect the acquisition geometry of registered serial sections.

2.3 Implicit Neural Representations in Biology

INRs parameterize signals as continuous functions $\Phi(\mathbf{x})$, where $\mathbf{x}$ denotes coordinates. While this paradigm is widely used for 3D reconstruction in computer vision through coordinate-based MLPs, spatial biology introduces task-specific acquisition and technical-variation considerations. More recently, this paradigm extends to biological data. Works such as SUICA (Zhu et al., 2025) and STINR (Luo et al., 2025) explore modeling gene expression fields with neural networks, supporting coordinate queries without a predefined output grid. Directly extending a coordinate INR to multi-slice ST leaves two acquisition-specific issues implicit:

- **Isotropic Bias:** Symmetric (x, y, z) coordinate encodings do not explicitly represent the asymmetry between the dense, quasi-continuous XY-plane and the sparse, discrete Z-axis.
- **Batch Sensitivity:** A coordinate regression objective alone does not explicitly distinguish independently acquired slice effects from biological variation.

HoloGene addresses these issues through decoupled anisotropic conditioning and pseudo-label-stratified cross-slice alignment.

3 Method

3.1 Overview

Reconstructing a continuous 3D gene expression field from sliced ST data presents two fundamental challenges: the *curse of dimensionality* inherent in whole-transcriptome profiles and the *intrinsic data anisotropy* arising from sparse sampling along the Z-axis.

To address these, we propose HoloGene. As illustrated in Figure 2, the framework consists of three integrated components:

1. **Manifold Compression:** We project high-dimensional, sparse gene expression vectors into compact, topology-preserving latent embeddings using a GAE, following SUICA (Zhu et al., 2025).
2. **Pioneer Signature Learning:** We predict training-expression-derived pseudo-label groups from 3D coordinates and use their probabilities to guide latent-field synthesis.
3. **Anisotropic Latent Synthesis:** We synthesize a continuous 3D embedding field using a decoupled INR that conditions in-plane biological variation on depth-dependent axial context.

3.2 Manifold Compression

Directly regressing high-dimensional, zero-inflated gene counts with an INR is computationally prohibitive and highly susceptible to noise overfitting. Therefore, we adopt a manifold learning strategy inspired by SUICA (Zhu et al., 2025) to compress the data into a dense latent representation.

We construct a spatial neighbor graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ over the *training slices only*, where $\mathcal{V}$ and $\mathcal{E}$ denote the node and edge sets, respectively, and each node v_i represents a spot with raw gene expression $\mathbf{g}_i \in \mathbb{R}^G$. Test slices are entirely excluded from graph construction and GCN message passing, ensuring no information leakage. A GCN encoder aggregates local neighborhood information to map $\mathbf{g}_i$ to a low-dimensional embedding $\mathbf{e}_i \in \mathbb{R}^D$ (typically $D = 64$), preserving local tissue topology. The network is pre-trained by minimizing the reconstruction loss. For subsequent INR training, we treat these learned embeddings $\{\mathbf{e}_i\}$ as latent regression targets, transforming the task from regressing sparse counts to fitting a smooth, denoised latent manifold that preserves the underlying tissue topology. Note that similar to SUICA (Zhu et al., 2025), the training of GAE is performed as a separate step.

3.3 Pioneer Signature Learning

Before generating detailed tissue structures, we employ a Pioneer Signature Network (PSN) to map 3D coordinates to coarse tissue-state groups derived from training expression. This module distills coordinate-predictable structure from expression-derived partitions into a conditioning signal for the subsequent synthesis phase; it does not discover biological states independently of expression.

PSN takes the full 3D coordinates $\mathbf{c} = (x, y, z)$ as input. Concretely, the in-plane coordinates (x, y) are first encoded via Gaussian/Fourier positional encoding to obtain spatial features, while z is passed through a small MLP to obtain depth features. The two feature vectors are concatenated and fed into a two-layer MLP that outputs biological-signature logits, followed by softmax to produce the Signature Probability Distribution $\mathbf{p}_{\text{sig}} \in \mathbb{R}^K$:

$$\mathbf{p}_{\text{sig}} = \text{softmax}\big(\text{MLP}\big[\gamma(x, y) \,\|\, \text{MLP}_z(z)\big]\big), \tag{1}$$

where $\gamma(\cdot)$ denotes the positional encoding and $\|$ denotes concatenation. Here, K is determined by the Leiden algorithm (Traag et al., 2019) applied to training-slice expression profiles (using the default resolution of 1.0 in scanpy). We use these Leiden-derived pseudo-labels for PSN supervision across all datasets and do not use dataset-provided cell-type annotations as supervision targets, avoiding label leakage from curated biological annotations. We use the standard Cross-Entropy Loss ($\mathcal{L}_{\text{CE}}$) that enforces consistency with pseudo-labels for supervision. This pseudo-label supervision provides coarse biological-state guidance for latent-field synthesis.

3.4 Anisotropic Latent Synthesis

The core of HoloGene is a conditional INR designed to handle the anisotropy of ST data. Standard isotropic coordinate INRs do not explicitly encode the axial–lateral sampling imbalance of serial-section ST. We propose a Decoupled Conditional Strategy that separates intra-slice manifold generation from inter-slice geometric evolution. Yin et al. (2023) explore a related decoupling strategy in DINo, where the temporal variable is modeled independently.

3.4.1 Decoupled Inputs and Conditioning

We restrict the main input of the synthesis network to intra-slice coordinates (x, y). The Z-axis information is exclusively provided via conditioning, which encourages the network to learn a canonical 2D manifold that is continuously modulated to match the structure at depth z.

We utilize Feature-wise Linear Modulation (FiLM) to inject the decoupled conditions directly into the intra-slice spatial embeddings. Specifically, we apply FiLM at every hidden layer. For a layer of width H_l, let $\mathbf{d}_z^{(l)} = \psi_l(z) \in \mathbb{R}^{2H_l}$ denote its learned depth condition, where ψ_l is a layer-specific affine map. This condition is distinct from $\mathrm{MLP}_z(z)$ used by PSN. The depth condition and a learned embedding of $\mathbf{p}_{\mathrm{sig}}$ jointly produce the layer-wise scale and shift parameters. At the l-th layer, we modulate the hidden feature $\mathbf{h}_l$ to obtain $\mathbf{h}_{l+1}$:

$$\mathbf{h}_{l+1} = \sigma\big(\boldsymbol{\phi}^{(l)} \odot \mathbf{h}_l + \boldsymbol{\beta}^{(l)}\big), \tag{2}$$

where $\mathbf{h}_l$ denotes the hidden representation at layer l, $\sigma(\cdot)$ denotes the non-linear activation, and the layer-wise modulation parameters $\boldsymbol{\phi}^{(l)}$ (scaling) and $\boldsymbol{\beta}^{(l)}$ (shifting) combine the following two conditions:

- **Signature Condition ($\mathbf{p}_{\mathrm{sig}}$):** The PSN probability distribution is mapped to a learned signature embedding that contributes to both scale and shift.
- **Z-Evolution Condition ($\mathbf{d}_z^{(l)}$):** The layer-specific encoding $\psi_l(z)$ contributes depth-dependent scale and shift, allowing the network to represent axial variation between measured slices.

3.4.2 Synthesis Objective

The synthesis network $\mathcal{S}_\theta$ is trained to regress the continuous latent manifold, which is later recovered to the raw gene space by the pre-trained decoder of GAE. Since the target embeddings $\mathbf{e}_i$ lie in a continuous Euclidean space, we optimize the network parameters θ by minimizing the Mean Squared Error (MSE) over all valid spots Ω:

$$\mathcal{L}_{\mathrm{emb}} = \frac{1}{|\Omega|} \sum_{i \in \Omega} \|\mathcal{S}_\theta(x_i, y_i \mid \mathbf{p}_{\mathrm{sig}}^{(i)}, z_i) - \mathbf{e}_i\|_2^2, \tag{3}$$

where $\mathbf{e}_i$ denotes the latent regression target for spot i derived from the pre-trained GAE. This objective forces the network to reconstruct a spatially continuous field that respects both the intra-slice topological structure (encoded in $\mathbf{e}_i$) and the inter-slice evolutionary consistency (enforced by the conditioning).

A critical hurdle in 3D ST is the "batch effect"—systematic technical variations between physical slices. If left unaddressed, the network may learn slice-specific artifacts as biological signals. To mitigate this, we also introduce a Correlation Alignment (CORAL) Loss, inspired by Deep CORAL (Sun and Saenko, 2016). This loss minimizes discrepancies between second-order statistics (covariance) of latent feature distributions for the same signature across different slices. Let $\mathbf{C}_{k,s} \in \mathbb{R}^{D \times D}$ denote the covariance matrix of D-dimensional latent features belonging to signature k in slice s. For each mini-batch, let $\mathcal{P}$ be the set of valid triples (k, s, s') for which pseudo-label group k contains at least the configured minimum number of spots in both slices s and s'. The covariance component of the batch-correction loss is

$$\mathcal{L}_{\mathrm{CORAL}}^{\mathrm{cov}} = \frac{1}{|\mathcal{P}| D^2} \sum_{(k,s,s') \in \mathcal{P}} \|\mathbf{C}_{k,s} - \mathbf{C}_{k,s'}\|_F^2 . \tag{4}$$

In practice, we discard a signature group k if it contains fewer than $2 \times$ `min_n` spots across all slices in the mini-batch, and exclude any slice-specific subset containing fewer than `min_n` spots. Thus,

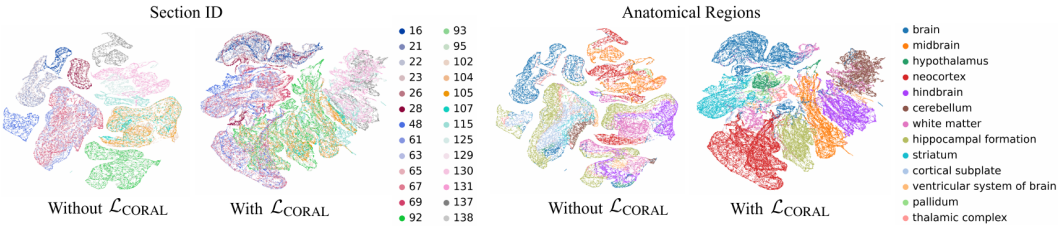


Figure 3: Ablation on $\mathcal{L}_{\text{CORAL}}$: UMAP visualizations colored by section ID and anatomical regions on the whole mouse brain data (Zhang et al., 2023). Enabling $\mathcal{L}_{\text{CORAL}}$ leads to better cross-slice mixing and less visual separation by section ID.

only groups shared by at least two sufficiently populated slices contribute to $\mathcal{P}$. This conservative filtering prevents CORAL from forcing alignment using poorly estimated covariance matrices from nearly empty subsets. By minimizing $\mathcal{L}_{\text{CORAL}}^{\text{cov}}$, we reduce cross-slice covariance discrepancies within each signature group and discourage slice-specific variation in the latent features. As illustrated in Figure 3, enabling $\mathcal{L}_{\text{CORAL}}$ produces better cross-slice mixing in the embedding space and less visual separation by slice identity. Together with the quantitative probes in Table 19, this supports the narrower conclusion that CORAL reduces measured slice-identity leakage on the evaluated whole mouse brain data. Once trained, HoloGene enables resolution-independent querying. For any queried coordinate in the pre-registered tissue domain, the network predicts an expression vector in $\mathbb{R}^G$. Predictions on held-out measured sections can be evaluated against observed expression, while queries between measured sections are model-based interpolation estimates.

4 Experiments

4.1 Datasets & Metrics

We report the main results on two representative 3D spatial transcriptomics datasets: MOSTA E16.5 (Chen et al., 2022), a volumetric mouse embryo atlas at embryonic day 16.5, and a whole mouse brain dataset (Zhang et al., 2023). Additional results on 3D Visual Cortex (Wang et al., 2018) and Breast Cancer (Kuett et al., 2022) are provided in Appendix A. For MOSTA E16.5, we use Spateo (Qiu et al., 2024) for spatial alignment in 2D across slices and place the slices in 3D space by directly adopting the z-axis coordinates from the original study (Qiu et al., 2024) to arrange slices in 3D; the other datasets provide pre-registered 3D coordinates directly.

For quantitative evaluation, we follow SUICA (Zhu et al., 2025) and report mean squared error (MSE), cosine similarity (Cosine), Pearson and Spearman correlations, and adjusted Rand index (ARI). For expression reconstruction, MSE, Cosine, and Spearman are computed only at entries with ground-truth expression greater than zero (masked metrics, denoted by subscript m), whereas Pearson uses all entries without masking. MSE/Cosine measure numerical fidelity, Pearson and Spearman measure correlation, and ARI compares clusters inferred from predicted expressions on held-out slices with matched dataset-provided cell-type or anatomical annotations. These annotations are used only for evaluation, while PSN is supervised by Leiden pseudo-labels from training-slice expressions. We further report annotation-referenced biological preservation metrics, including DEG F1, Cell-Cell Communications (CCC) Pearson/Top-F1 (based on CellTalkDB (Shao et al., 2021)), and pathway F1 using pathway enrichment resources; definitions are in Appendix K. For all metrics, higher values indicate better performance, except for MSE, where lower values are preferred ($\downarrow$).

4.2 Evaluation Protocol

We adopt a specimen-specific, slice-wise measured-section holdout protocol to evaluate within-volume axial generalization in a pre-registered tissue coordinate domain. HoloGene estimates expression in the supplied coordinate system and does not jointly solve slice registration.

Concretely, we randomly split sections along the z-axis into 80% for training and 20% for testing. All spots from held-out measured sections are excluded from model training, GAE pre-training, graph construction, and GCN message passing. The z-coordinates of test sections are never observed during training. We then evaluate whether the learned continuous estimator can recover the measured

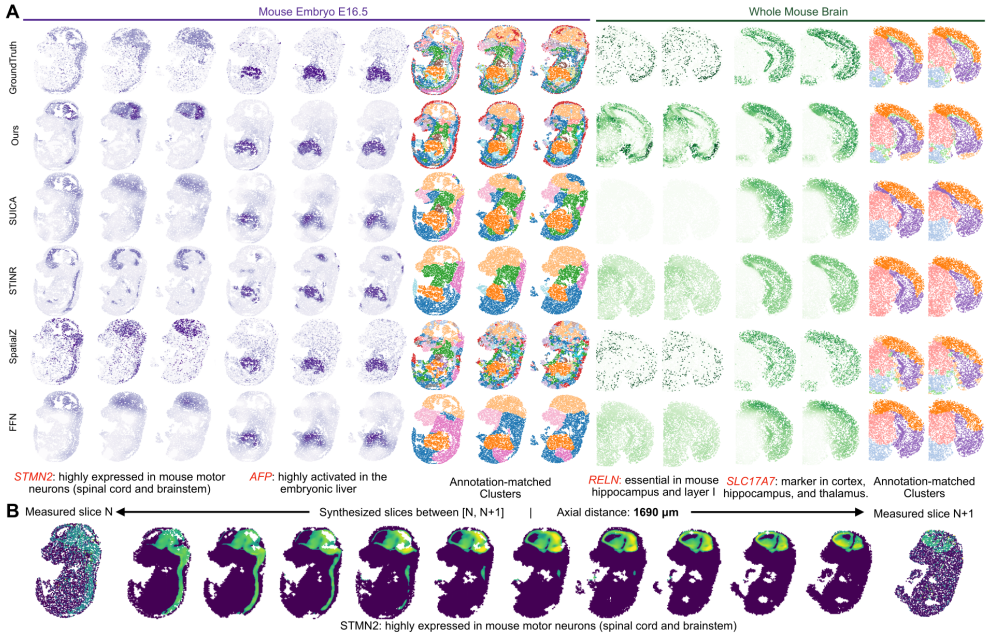


Figure 4: Held-out section synthesis and inter-section querying. A) Marker-gene and domain-map reconstruction on held-out measured sections. Domain maps show ground-truth annotations in the first row and predicted-expression clusters mapped to annotation colors in the remaining rows. See full results in Appendix A. B) Querying HoloGene between measured sections provides model-based interpolation estimates with smooth STMN2 variation. See full results in Appendix C.

expression profiles, marker-gene patterns, annotation-consistent domains, and downstream biological readouts of these unseen physical sections. The split is performed before every learned stage: the GAE graph and pretraining, Leiden pseudo-label construction, PSN supervision, latent synthesis, and CORAL regularization all use training sections only. Held-out measured sections are used only for evaluation and visualization.

4.3 Baselines

We compare HoloGene against both general-purpose INRs and recent task-specific spatial transcriptomics methods. Specifically, we include Fourier Feature Networks (FFN) (Tancik et al., 2020) and SIREN (Sitzmann et al., 2020) as representative INR baselines. We also compare with recent spatial transcriptomics methods, including STAGE (Li et al., 2024), SpatialZ (Lin et al., 2025), STINR (Luo et al., 2025), and SUICA (Zhu et al., 2025). Note that STAGE and SUICA are originally designed for 2D spatial data. For STAGE, we change the coordinate-matched latent bottleneck and its supervision target from (x, y) to registered (x, y, z) while retaining the original encoder, decoder, and objectives. For SUICA, we change the coordinate-MLP input from $\mathbb{R}^2$ to $\mathbb{R}^3$ and query registered (x, y, z) while retaining its GAE, decoder, and objectives. All methods receive the same aligned coordinates and use the same training and held-out physical sections and gene set; predictions are evaluated in the original expression space. For INR-based baselines, we grid-search learning rate $\in \{10^{-3}, 5 \times 10^{-4}, 10^{-4}\}$, batch size $\in \{2048, 4096, 8192\}$, and latent dimension $\in \{32, 64, 128\}$; STAGE follows its original training strategy. For SpatialZ, which is not a coordinate-INR baseline, we use the authors' default configuration from the released implementation. In ablations, we use the same pretrained GAE latent targets and decoder to isolate the effects of anisotropic FiLM conditioning, PSN guidance, and CORAL alignment.

4.4 Implementation Details

We implement all methods in PyTorch and run experiments on a single NVIDIA A100 GPU. For graph-based components, the k-NN graph uses $k = 5$ including self-loops. Baseline-specific training schedules follow the original implementations when available; for INR-based baselines, we use the grid search described above. Runtime is reported in Appendix I.

Table 1: Quantitative benchmarking of virtual slice synthesis. Subscript m denotes metrics evaluated only where ground-truth expression is positive. Bold/underline denote best/second-best based on unrounded values.

Method	MOSTA E16.5					Whole Mouse Brain				
	$\mathrm{MSE}_m \downarrow$	$\mathrm{Cos}_m \uparrow$	Pearson↑	$\mathrm{Spear}_m \uparrow$	ARI↑	$\mathrm{MSE}_m \downarrow$	$\mathrm{Cos}_m \uparrow$	Pearson↑	$\mathrm{Spear}_m \uparrow$	ARI↑
FFN (Tancik et al., 2020)	0.2588	0.7630	0.6525	0.5034	0.1009	<u>2.5616</u>	0.7443	0.4414	0.4009	<u>0.4275</u>
SIREN (Sitzmann et al., 2020)	0.2835	0.7373	0.6165	0.4695	0.0187	2.5778	0.7717	<u>0.4624</u>	<u>0.4262</u>	0.3837
STAGE (Li et al., 2024)	0.2987	0.7234	0.6655	0.5237	0.1213	2.9103	0.7356	0.4020	0.3662	0.3313
SpatialZ (Lin et al., 2025)	0.3145	0.7029	0.5315	0.4722	0.1188	3.3409	0.5279	0.2335	0.2070	0.0221
STINR (Luo et al., 2025)	<u>0.2332</u>	<u>0.7971</u>	<u>0.6986</u>	0.5308	<u>0.1422</u>	2.7411	<u>0.7782</u>	0.4289	0.3895	0.2586
SUICA-3D (Zhu et al., 2025)	0.2972	0.7233	0.6772	<u>0.5412</u>	0.1393	2.7776	**0.7803**	0.4130	0.3839	0.3618
HoloGene (Ours)	**0.2303**	**0.8054**	**0.7146**	**0.5503**	**0.1637**	**2.4835**	0.7748	**0.4648**	**0.4315**	**0.4670**

Table 2: Downstream biological-readout preservation on MOSTA E16.5 and whole mouse brain. Bold/underline denote best/second-best based on unrounded values.

Method	MOSTA E16.5				Whole Mouse Brain			
	Path. F1↑	CCC Pearson↑	CCC Top-F1↑	DEG F1↑	Path. F1↑	CCC Pearson↑	CCC Top-F1↑	DEG F1↑
FFN (Tancik et al., 2020)	0.5348	0.8654	<u>0.8892</u>	0.4389	0.3462	0.9708	0.8800	0.8346
SIREN (Sitzmann et al., 2020)	0.3419	0.8444	0.8851	0.4056	<u>0.5538</u>	<u>0.9767</u>	0.8956	0.8615
STAGE (Li et al., 2024)	0.4068	0.8986	0.8880	0.2861	0.4242	0.8977	0.5444	0.6269
SpatialZ (Lin et al., 2025)	<u>0.6490</u>	0.8806	0.8642	<u>0.5421</u>	0.4935	0.9752	<u>0.8991</u>	<u>0.8731</u>
STINR (Luo et al., 2025)	0.5674	0.8786	0.8823	0.4306	0.4545	0.9618	0.8489	0.7192
SUICA-3D (Zhu et al., 2025)	0.5163	<u>0.9122</u>	0.8509	0.4167	0.3529	0.9245	0.8222	0.6692
HoloGene (Ours)	**0.7239**	**0.9342**	**0.9138**	**0.6278**	**0.5714**	**0.9804**	**0.9022**	**0.8808**

4.5 Results

Main benchmark on held-out measured slices. Tables 1 and 2 and Figure 4 summarize the main performance comparison. HoloGene achieves the lowest MSE and the highest Pearson, masked Spearman, and ARI scores on both MOSTA E16.5 and whole mouse brain data, as well as the highest masked Cosine on MOSTA E16.5. It also achieves the best scores on all four biological preservation metrics on both datasets; whole-brain masked Cosine is slightly below SUICA-3D and STINR. On whole mouse brain, HoloGene raises ARI from the strongest baseline score of 0.428 to 0.467, indicating better recovery of annotation-consistent tissue domains from held-out slices. Additional results on 3D Visual Cortex and Breast Cancer are reported in Appendix A, where HoloGene remains competitive or best on representative reconstruction and biological preservation metrics.

Marker-level and domain-level reconstruction. The qualitative comparisons in Figure 4 illustrate marker- and domain-level behavior across the two primary datasets. The selected markers cover distinct anatomical programs. In MOSTA E16.5, *STMN2* (San Juan et al., 2022) marks neuronal and motor-neuron-enriched structures, while *AFP* (Kwon et al., 2006) highlights the embryonic liver region. In the whole mouse brain, *RELN* (Vílchez-Acosta et al., 2022; Fatemi, 2011) localizes to hippocampal and layer-I cortical structures, and *SLC17A7* (Herzog et al., 2011) marks glutamatergic excitatory neuronal populations in cortex, hippocampus, and thalamus. The predictions preserve these region-specific patterns across multiple held-out z positions, and the annotation-matched domain maps show that expression-derived tissue organization remains spatially coherent after synthesis. Individual marker and section outcomes can differ across methods; our comparative conclusions are based on aggregate evaluation over all held-out measured sections. See full results in Appendix A.

Preservation of downstream biological readouts. Beyond marker-level reconstruction, we further evaluated whether downstream biological readouts remain recoverable from the synthesized expression fields. In Table 2, HoloGene achieves the best CCC Pearson and CCC Top-F1 on both datasets, indicating that ligand–receptor communication patterns between tissue domains remain recoverable after virtual-slice synthesis. As an annotation-referenced case study on MOSTA E16.5 data, communication scores computed from predicted expressions closely match those from ground-truth expressions, with small residual differences and similar top sender-to-receiver connections (Figure 5, see full details in Appendix B). These results suggest that HoloGene preserves not only local marker-gene expression but also higher-order interaction structure inferred from spatially organized cell populations.

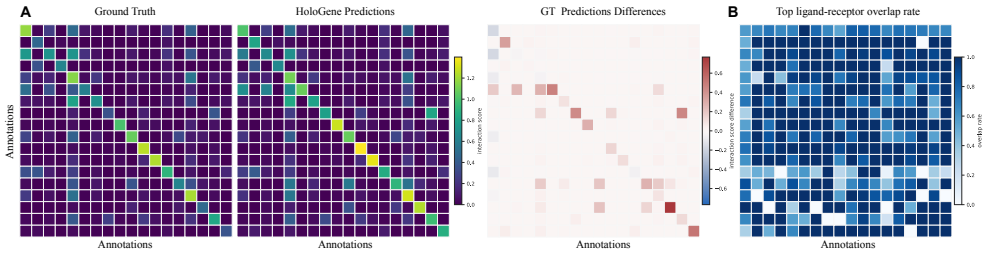


Figure 5: HoloGene preserves CCC patterns in MOSTA E16.5. (A) Ground-truth and HoloGene-predicted communication scores with their differences in the testing set. (B) Overlap rate of top significant ligand-receptor interactions recovered by HoloGene across annotation pairs. Darker blue indicates higher recovery.

Table 3: Component ablations. All variants use the same pretrained GAE. Bold/underline denote best/second-best based on unrounded values; full results are in Appendices D and K.

(a) Reconstruction and annotation-consistent clustering

Method	MOSTA E16.5			Whole Mouse Brain		
	$\text{MSE}_m \downarrow$	Pearson↑	ARI↑	$\text{MSE}_m \downarrow$	Pearson↑	ARI↑
Isotropic XYZ	0.2413	0.6869	0.1310	2.5460	0.4446	0.4155
FiLM-only	0.2354	0.7132	0.1454	2.5041	0.4636	0.4231
FiLM + PSN	0.2308	0.7134	0.1440	2.4899	0.4636	<u>0.4398</u>
FiLM + CORAL	**0.2235**	**0.7154**	<u>0.1569</u>	**2.4715**	<u>0.4644</u>	0.4354
Full HoloGene	<u>0.2303</u>	<u>0.7146</u>	**0.1637**	<u>2.4835</u>	**0.4648**	**0.4670**

(b) Biological preservation

Method	MOSTA E16.5			Whole Mouse Brain		
	Path. F1↑	CCC Pearson↑	DEG F1↑	Path. F1↑	CCC Pearson↑	DEG F1↑
Isotropic XYZ	0.5810	0.8937	0.4583	0.5000	0.9671	0.8538
FiLM-only	0.6667	0.9287	0.5944	0.4000	0.9792	<u>0.8692</u>
FiLM + PSN	0.6853	0.9233	<u>0.6167</u>	<u>0.5352</u>	<u>0.9818</u>	0.8654
FiLM + CORAL	<u>0.6854</u>	<u>0.9290</u>	0.6056	0.5231	**0.9829**	<u>0.8692</u>
Full HoloGene	**0.7239**	**0.9342**	**0.6278**	**0.5714**	0.9804	**0.8808**

Continuous inter-section evolution. Beyond held-out measured-section reconstruction, HoloGene can be queried at intermediate z-positions between measured sections. Because these locations do not have paired experimental measurements, we interpret the results as model-based interpolation estimates. Figure 4(B) shows that the estimated slices vary smoothly along the axial direction while preserving the *STMN2*-enriched motor-neuron pattern observed in neighboring measured sections; the full two-marker visualization is provided in Appendix C.

Ablations. Table 3 analyzes the contribution of anisotropic depth conditioning, PSN guidance, and CORAL alignment under a shared setup where all variants use the same GAE latent targets and decoder. Comparing Isotropic XYZ with FiLM-only evaluates anisotropic decoupling: replacing direct (x, y, z) input with (x, y)-based synthesis modulated by z improves MSE and Pearson on both primary datasets. In the FiLM-only versus FiLM+CORAL comparison, CORAL reduces whole-brain slice leakage from 0.1966 to 0.1770 and increases annotation retention from 0.4929 to 0.5102; label-independent diagnostics also show improved gene-level spatial autocorrelation and z-gradient preservation (Table 19). PSN contributes complementary signature-conditioned guidance, with metric-dependent benefits for biological preservation and annotation-consistent domain recovery. Full HoloGene does not maximize every metric or diagnostic independently; instead, it balances reconstruction, domain consistency, and biological preservation by combining FiLM-based depth conditioning, CORAL-based alignment, and PSN-based signature guidance. Complete ablations, label-independent diagnostics, representation probes, split-ratio analysis, and runtime results are provided in Appendices D, E, I and K.

5 Conclusion

We presented HoloGene, a manifold-aware conditional INR framework for estimating continuous 3D expression fields from discontinuous spatial transcriptomics sections. By combining GAE-based latent compression, FiLM-based depth conditioning, PSN-based signature guidance, and CORAL-based cross-slice alignment, HoloGene addresses the anisotropic sampling and slice-specific variation inherent to serial-section ST data. Under a slice-wise measured-section holdout protocol, HoloGene achieves strong reconstruction and biological-preservation performance across multiple 3D ST datasets, while ablations show that the proposed components provide complementary benefits for depth-aware synthesis, cross-slice alignment, and biological structure preservation.

Limitations. HoloGene estimates expression over a pre-registered tissue coordinate domain rather than directly measuring a dense 3D transcriptome. Intermediate queries between measured sections should therefore be interpreted as model-based estimates, and downstream biological conclusions require independent experimental validation. The current model produces point estimates without uncertainty quantification and requires pre-registered coordinates; the stress tests show performance degradation under larger residual misalignment and very sparse axial coverage. Finally, PSN supervision relies on Leiden-derived pseudo-labels. Reconstruction is stable over the evaluated resolution and label-corruption ranges, while annotation-referenced domain recovery is more sensitive; behavior beyond these ranges remains uncharacterized.

Acknowledgments

This work was supported by AMED [JP23tm0524003]; JSPS KAKENHI [24KK0209]; JST-Mirai Program [JPMJMI23G1]; the Canadian Institutes of Health Research (CIHR) [PJT-180505]; the Fonds de recherche du Québec - Santé (FRQS) [295298, 295299, 366764]; the Natural Sciences and Engineering Research Council of Canada (NSERC) [RGPIN2022-04399]; and the Meakins-Christie Chair in Respiratory Research. Funded by the European Union (AlRls, Grant Agreement No. 101289094). The authors declare no competing interests.

References

Ayushi Agrawal, Hasan Balcı, Kristina Hanspers, Susan L Coort, Marvin Martens, Denise N Slenter, Friederike Ehrhart, Daniela Digles, Andra Waagmeester, Isabel Wassink, et al. Wikipathways 2024: next generation pathway database. *Nucleic acids research*, 52(D1):D679–D689, 2024.

Ao Chen, Sha Liao, Mengnan Cheng, Kailong Ma, Liang Wu, Yiwei Lai, Xiaojie Qiu, Jin Yang, Jiangshan Xu, Shijie Hao, et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using dna nanoball-patterned arrays. *Cell*, 185(10):1777–1792, 2022.

The Gene Ontology Consortium. The gene ontology knowledgebase in 2026. *Nucleic Acids Research*, 54(D1):D1779–D1792, 01 2026. ISSN 1362-4962. doi: 10.1093/nar/gkaf1292. URL `https://doi.org/10.1093/nar/gkaf1292`.

S Hossein Fatemi. Reelin, a marker of stress resilience in depression and psychosis. *Neuropsychopharmacology*, 36(12):2371–2372, 2011.

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

Etienne Herzog, Fabien Nadrigny, Katlin Silm, Christoph Biesemann, Imke Helling, Tiphaine Bersot, Heinz Steffens, Richard Schwartzmann, U Valentin Nägerl, Salah El Mestikawy, et al. In vivo imaging of intersynaptic vesicle exchange using vglut1venus knock-in mice. *Journal of Neuroscience*, 31(43):15544–15559, 2011.

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

Minoru Kanehisa, Miho Furumichi, Yoko Sato, Masayuki Kawashima, and Mari Ishiguro-Watanabe. Kegg for taxonomy-based analysis of pathways and genomes. *Nucleic acids research*, 51(D1): D587–D592, 2023.

Liis Kolberg, Uku Raudvere, Ivan Kuzmin, Priit Adler, Jaak Vilo, and Hedi Peterson. g:profiler: interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). *Nucleic Acids Research*, 51(W1):W207–W212, 2023. doi: 10.1093/nar/gkad347.

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

Laura Kuett, Raúl Catena, Alaz Özcan, Alex Plüss, Peter Schraml, Holger Moch, Natalie de Souza, and Bernd Bodenmiller. Three-dimensional imaging mass cytometry for highly multiplexed molecular and cellular mapping of tissues and the tumor microenvironment. *Nature cancer*, 3(1): 122–133, 2022.

Gloria S Kwon, Stuart T Fraser, Guy S Eakin, Michael Mangano, Joan Isern, Kenneth E Sahr, Anna-Katerina Hadjantonakis, and Margaret H Baron. Tg (afp-gfp) expression marks primitive and definitive endoderm lineages during mouse development. *Developmental dynamics: an official publication of the American Association of Anatomists*, 235(9):2549–2558, 2006.

Shang Li, Kuo Gai, Kangning Dong, Yiyang Zhang, and Shihua Zhang. High-density generation of spatial transcriptomics with stage. *Nucleic Acids Research*, 52(9):4843–4856, 2024.

Senlin Lin, Zhikang Wang, Yan Cui, Qi Zou, Chuangyi Han, Rui Yan, Zhidong Yang, Wei Zhang, Rui Gao, Jiangning Song, et al. Bridging the dimensional gap from planar spatial transcriptomics to 3d cell atlases. *Nature Methods*, pages 1–13, 2025.

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

Yisi Luo, Xile Zhao, Kai Ye, and Deyu Meng. Stinr: Deciphering spatial transcriptomics via implicit neural representation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25930–25939, 2025.

Xiaojie Qiu, Daniel Y Zhu, Yifan Lu, Jiajun Yao, Zehua Jing, Kyung Hoi Min, Mengnan Cheng, Hailin Pan, Lulu Zuo, Samuel King, et al. Spatiotemporal modeling of molecular holograms. *Cell*, 187(26):7351–7373, 2024.

Eliot Ragueneau, Chuqiao Gong, Pierre Sinquin, Cristoffer Sevilla, Deidre Beavers, Alexander Grentner, Johannes Griss, Gregory F J Hogue, Nancy T Li, Lisa Matthews, Bruce May, Marija Milacic, Helia Mohammadi, Robert Petryszak, Karen Rothfels, Veronica Shamovsky, Ralf Stephan, Krishna Tiwari, Joel Weiser, Adam Wright, Marc Gillespie, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The reactome knowledgebase 2026. *Nucleic Acids Research*, 54(D1):D673–D681, 01 2026. ISSN 1362-4962. doi: 10.1093/nar/gkaf1223. URL `https://doi.org/10.1093/nar/gkaf1223`.

Irune Guerra San Juan, Leslie A Nash, Kevin S Smith, Marcel F Leyton-Jaimes, Menglu Qian, Joseph R Klim, Francesco Limone, Alexander B Dorr, Alexander Couto, Greta Pintacuda, et al. Loss of mouse stmn2 function causes motor neuropathy. *Neuron*, 110(10):1671–1688, 2022.

Xuesong Shao, Jian Liao, Chen Li, Xiaoyan Lu, Jie Cheng, and Xiaohui Fan. Celltalkdb: a manually curated database of ligand-receptor interactions in humans and mice. *Briefings in Bioinformatics*, 22(4):bbaa269, 2021. doi: 10.1093/bib/bbaa269.

Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in neural information processing systems*, 33:7462–7473, 2020.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016.

Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33:7537–7547, 2020.

Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12, 2019.

Alba Vílchez-Acosta, Yasmina Manso, Adrián Cárdenas, Alba Elias-Tersa, Magdalena Martínez-Losa, Marta Pascual, Manuel Álvarez-Dolado, Angus C Nairn, Víctor Borrell, and Eduardo Soriano. Specific contribution of reelin expressed by cajal–retzius cells or gabaergic interneurons to cortical lamination. *Proceedings of the National Academy of Sciences*, 119(37):e2120079119, 2022.

Xiao Wang, William E Allen, Matthew A Wright, Emily L Sylwestrak, Nikolay Samusik, Sam Vesuna, Kathryn Evans, Cindy Liu, Charu Ramakrishnan, Jia Liu, et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*, 361(6400):eaat5691, 2018.

Yuan Yin, Matthieu Kirchmeyer, Jean-Yves Franceschi, Alain Rakotomamonjy, and Patrick Gallinari. Continuous pde dynamics forecasting with implicit neural representations. In *The Eleventh International Conference on Learning Representations*, 2023.

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

Meng Zhang, Xingjie Pan, Won Jung, Aaron R Halpern, Stephen W Eichhorn, Zhiyun Lei, Limor Cohen, Kimberly A Smith, Bosiljka Tasic, Zizhen Yao, et al. Molecularly defined and spatially resolved cell atlas of the whole mouse brain. *Nature*, 624(7991):343–354, 2023.

Qingtian Zhu, Yumin Zheng, Yuling Sang, Yifan Zhan, Ziyan Zhu, Jun Ding, and Yinqiang Zheng. Suica: Learning super-high dimensional sparse implicit neural representations for spatial transcriptomics. In *International Conference on Machine Learning*, pages 80448–80462. PMLR, 2025.

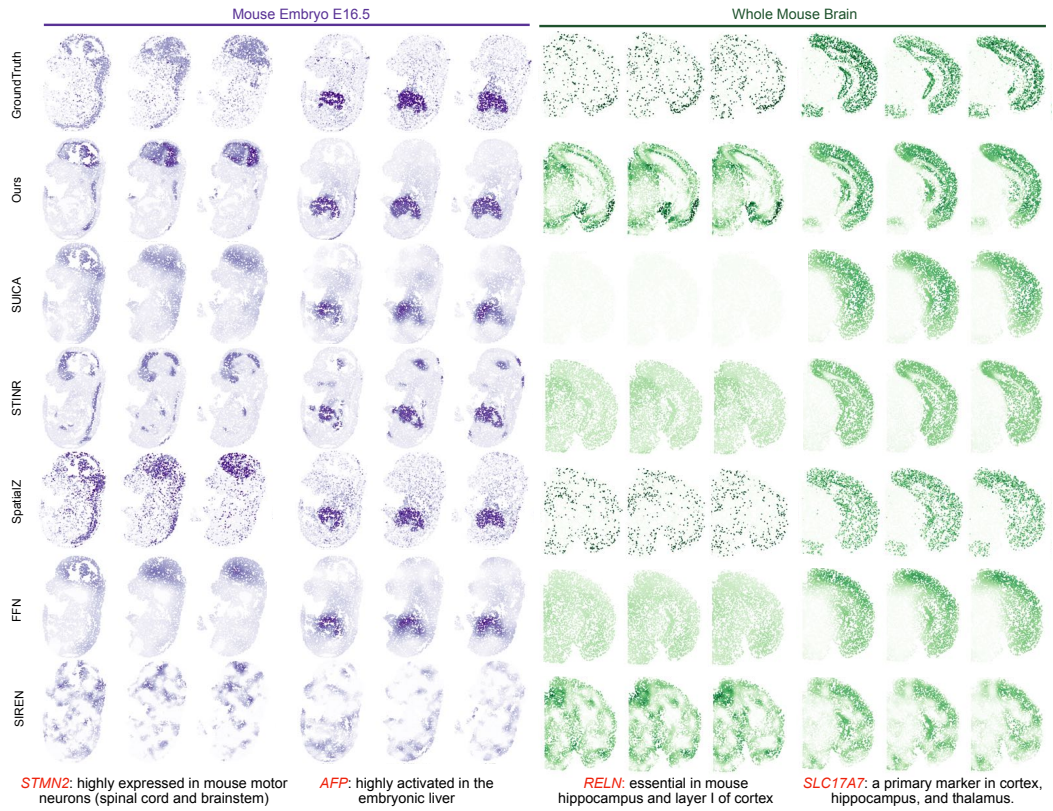


Figure 6: Qualitative results of virtual slice synthesis on representative 3D samples. We visualize measured reference slices and the corresponding predictions from different methods; marker-level outcomes vary across genes and sections, while aggregate comparisons are reported in the main text.

A Additional Results

We reported the full qualitative results of virtual slice synthesis on representative 3D samples in Figure 6. The Leiden-signature maps obtained from predicted gene expressions can be found in Figure 7.

We report additional benchmark results on 3D Visual Cortex (Wang et al., 2018) and Breast Cancer (Kuett et al., 2022) in Tables 4 and 5. For the breast cancer dataset, which contains a limited number of measured genes (25 genes), biological preservation metrics are not reported, as such metrics require broader transcriptomic coverage to be meaningful. These datasets are used as supplementary robustness checks, while the main paper focuses on MOSTA E16.5 and whole mouse brain to satisfy the main-text page budget.

Table 4: Additional reconstruction results on 3D Visual Cortex and Breast Cancer. Subscript m denotes masked metrics. Bold/underline denote best/second-best.

(a) 3D Visual Cortex

Method	$\mathrm{MSE}_m \downarrow$	$\mathrm{Cos}_m \uparrow$	Pearson↑	$\mathrm{Spear}_m \uparrow$	ARI↑
FFN (Tancik et al., 2020)	7.7323	0.8126	0.4156	0.4195	0.1492
SIREN (Sitzmann et al., 2020)	6.1583	0.8413	0.4518	0.5402	0.0747
STAGE (Li et al., 2024)	<u>4.9072</u>	0.8932	0.4989	0.6161	0.2134
SpatialZ (Lin et al., 2025)	8.3385	0.7488	0.3335	0.3554	0.0806
STINR (Luo et al., 2025)	4.9901	**0.9197**	<u>0.5421</u>	**0.6539**	0.1770
SUICA-3D (Zhu et al., 2025)	6.0817	0.8595	0.5186	0.5634	<u>0.3677</u>
HoloGene (Ours)	**4.8994**	<u>0.9050</u>	**0.5499**	<u>0.6531</u>	**0.4198**

(b) Breast Cancer

Method	$\mathrm{MSE}_m \downarrow$	$\mathrm{Cos}_m \uparrow$	Pearson↑	$\mathrm{Spear}_m \uparrow$
FFN (Tancik et al., 2020)	1.9190	0.9507	0.8179	0.8026
SIREN (Sitzmann et al., 2020)	0.8155	0.9793	0.9032	0.8754
STAGE (Li et al., 2024)	2.1255	0.9463	0.7632	0.7245
SpatialZ (Lin et al., 2025)	1.2619	0.9683	0.8557	0.8309
STINR (Luo et al., 2025)	<u>0.5906</u>	<u>0.9851</u>	<u>0.9203</u>	<u>0.8958</u>
SUICA-3D (Zhu et al., 2025)	1.8693	0.9520	0.8059	0.7815
HoloGene (Ours)	**0.5134**	**0.9871**	**0.9317**	**0.9087**

Table 5: Additional biological preservation metrics on 3D Visual Cortex. Pathway F1 is omitted because the 155-gene panel is too small for stable pathway enrichment.

Method	CCC Pearson↑	CCC Top-F1↑	DEG F1↑
FFN (Tancik et al., 2020)	0.8486	0.5000	0.2833
SIREN (Sitzmann et al., 2020)	<u>0.8928</u>	<u>0.5263</u>	0.4111
STAGE (Li et al., 2024)	0.7982	0.4342	0.2396
SpatialZ (Lin et al., 2025)	0.8783	0.4737	0.4401
STINR (Luo et al., 2025)	0.8215	0.4737	0.2563
SUICA-3D (Zhu et al., 2025)	0.8906	0.4079	<u>0.5111</u>
HoloGene (Ours)	**0.9002**	**0.5789**	**0.5905**



Figure 7: Leiden-signature maps obtained from predicted gene expressions. HoloGene yields more spatially coherent tissue regions with fewer isolated mislabeled spots compared with baselines.

B Cell Cell Communication reconstructions on the testing set

To further interpret the CCC preservation metrics, we provide annotation-referenced CCC visualizations in Figures 8 and 10. For each dataset, anatomical annotations are used only for post hoc sender–receiver grouping. We compute ligand–receptor communication scores from the ground-truth and HoloGene-predicted expression matrices using the same CCC procedure as in the main evaluation. Figure 10 compares the full annotation-pair communication score matrices, their prediction-minus-ground-truth residuals, and the overlap rates of top ligand–receptor interactions. Figure 8 provides a complementary network view by visualizing the strongest sender-to-receiver communication edges. Across both MOSTA E16.5 and whole mouse brain, HoloGene largely preserves the dominant communication topology, with small residual differences and high overlap of top interactions, supporting the CCC Pearson and CCC Top-F1 improvements reported in Table 2.

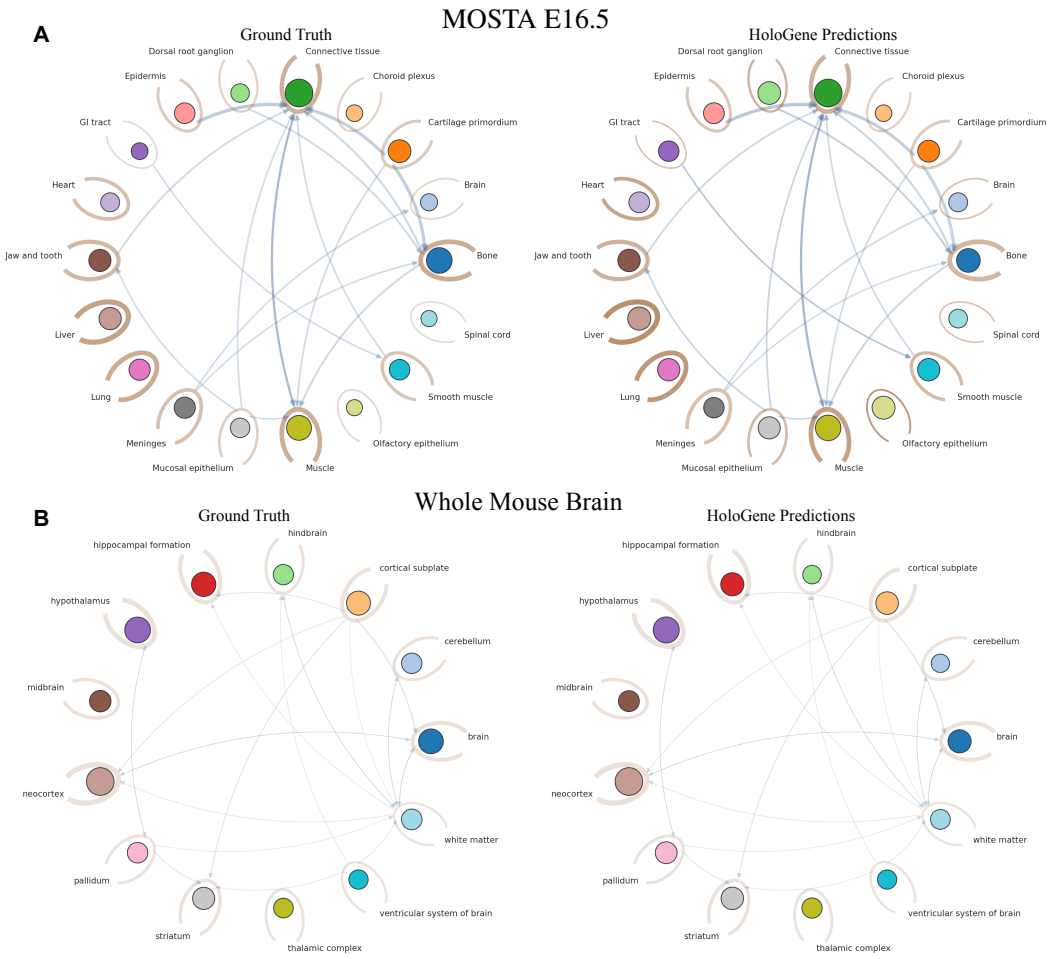
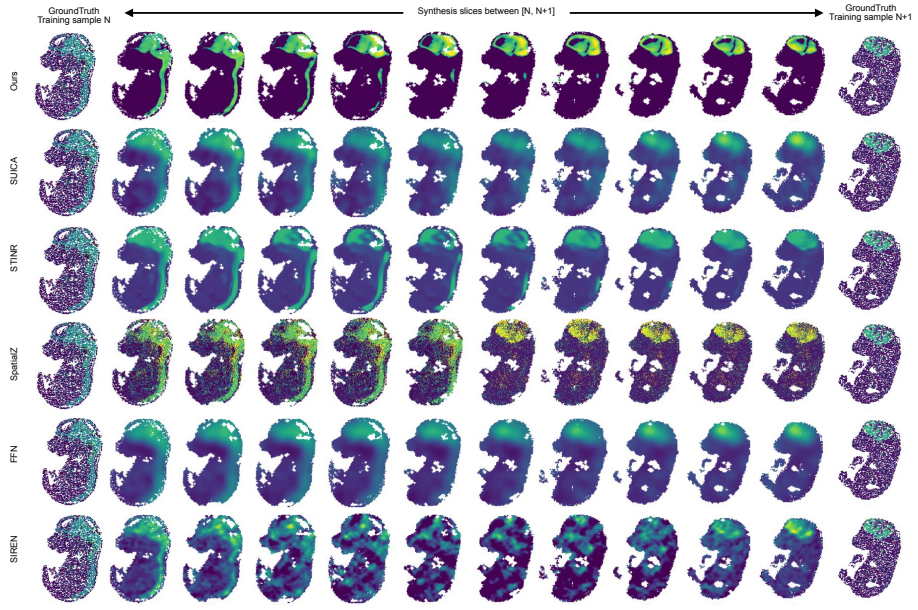


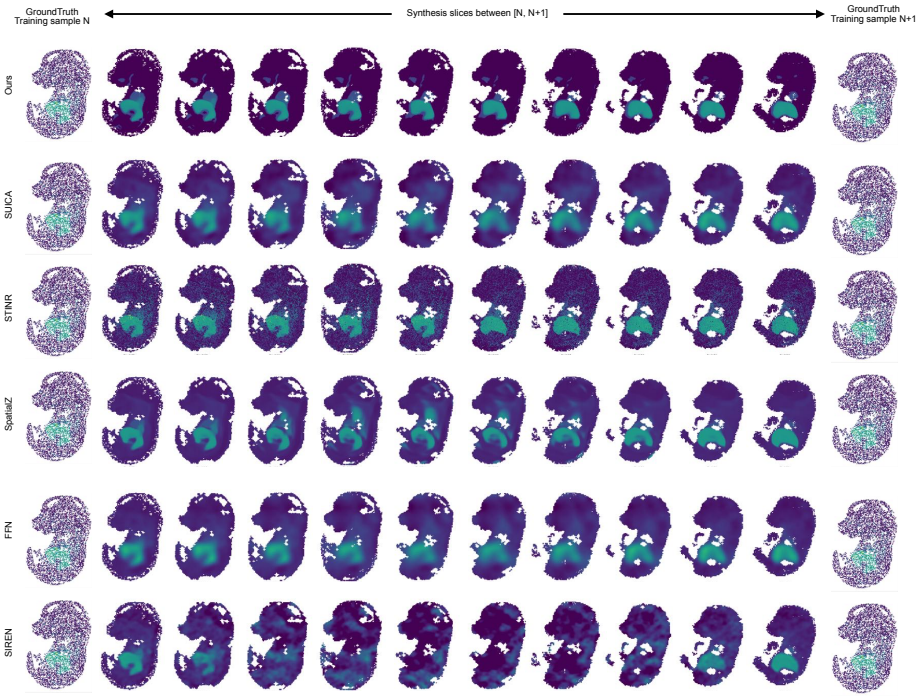
Figure 8: Annotation-referenced CCC network preservation on (A) MOSTA E16.5 and (B) whole mouse brain. Nodes denote anatomical annotations, and directed edges show top-scoring sender-to-receiver ligand–receptor communication connections. Ground-truth and HoloGene-predicted networks show similar high-confidence communication topology, indicating that virtual-slice synthesis preserves higher-order interaction structure.

C Visualization of Virtual Slice Evolution

We visualize the evolution of virtual slices between two measured slices for two gene markers in Figure 9. From left to right (or top to bottom), the leftmost and rightmost panels show the measured slices, while the intermediate panels show synthesized virtual slices, providing an intuitive assessment of the continuity and plausibility of the marker-expression patterns.



(a) STMN2: highly expressed in mouse motor neurons (spinal cord and brainstem)



(b) AFP: highly activated in the embryonic liver

Figure 9: Visualization of virtual-slice evolution for two gene markers. The leftmost and rightmost panels show measured slices, and the intermediate panels show synthesized virtual slices.

D Additional Ablations

We report unified ablations across reconstruction, clustering, and biological preservation metrics in Tables 6 to 8. We use the following variants. Isotropic XYZ uses the same GAE compression and decoder as HoloGene, but removes FiLM conditioning, PSN, and CORAL, and directly feeds (x, y, z) coordinates to the INR. FiLM only uses the decoupled FiLM-conditioned INR without PSN or CORAL. FiLM+PSN adds the PSN signature condition and supervision to the FiLM-only variant, whereas FiLM+CORAL adds only the CORAL loss, using the Leiden pseudo-label groups to define signature-conditioned covariance alignment when PSN is absent; evaluation metrics remain annotation-referenced as described in Appendix K. Full HoloGene uses FiLM, PSN, and CORAL jointly.

Table 6: MOSTA E16.5 ablation study. Subscript m denotes masked metrics. Higher is better except MSE.

Model	Path. F1↑	CCC Pearson↑	CCC Top-F1↑	DEG F1↑	$\mathrm{MSE}_m \downarrow$	$\mathrm{Cos}_m \uparrow$	Pearson↑	$\mathrm{Spear}_m \uparrow$	ARI↑
Isotropic XYZ	0.5810	0.8937	0.8865	0.4583	0.2413	0.7833	0.6869	0.5282	0.1310
FiLM-only	0.6667	0.9287	0.9056	0.5944	0.2354	0.7996	0.7132	0.5501	0.1454
FiLM + PSN	0.6853	0.9233	0.9070	0.6167	0.2308	0.8040	0.7134	0.5495	0.1440
FiLM + CORAL	0.6854	0.9290	0.9097	0.6056	**0.2235**	**0.8067**	**0.7154**	**0.5512**	0.1569
Full HoloGene	**0.7239**	**0.9342**	**0.9138**	**0.6278**	0.2303	0.8054	0.7146	0.5503	**0.1637**

Table 7: Whole mouse brain ablation study. Subscript m denotes masked metrics. Higher is better except MSE.

Model	Path. F1↑	CCC Pearson↑	CCC Top-F1↑	DEG F1↑	$\mathrm{MSE}_m \downarrow$	$\mathrm{Cos}_m \uparrow$	Pearson↑	$\mathrm{Spear}_m \uparrow$	ARI↑
Isotropic XYZ	0.5000	0.9671	0.8889	0.8538	2.5460	0.7519	0.4446	0.4125	0.4155
FiLM-only	0.4000	0.9792	0.8911	0.8692	2.5041	0.7718	0.4636	0.4309	0.4231
FiLM + PSN	0.5352	0.9818	**0.9022**	0.8654	2.4899	0.7720	0.4636	0.4305	0.4398
FiLM + CORAL	0.5231	**0.9829**	0.8933	0.8692	**2.4715**	0.7733	0.4644	0.4297	0.4354
Full HoloGene	**0.5714**	0.9804	**0.9022**	**0.8808**	2.4835	**0.7748**	**0.4648**	**0.4315**	**0.4670**

Table 8: 3D Visual Cortex ablation study. Subscript m denotes masked metrics. Higher is better except MSE. Pathway F1 is omitted because the 155-gene panel is too small for stable pathway enrichment.

Model	CCC Pearson↑	CCC Top-F1↑	DEG F1↑	$\mathrm{MSE}_m \downarrow$	$\mathrm{Cos}_m \uparrow$	Pearson↑	$\mathrm{Spear}_m \uparrow$	ARI↑
Isotropic XYZ	0.8975	0.5263	0.4958	5.1350	0.8842	0.5133	0.6161	0.2986
FiLM-only	0.8988	0.5132	0.5237	4.9825	0.9016	0.5496	0.6505	0.3819
FiLM + PSN	0.8985	0.5658	0.5627	4.9596	0.9038	**0.5541**	**0.6572**	0.3269
FiLM + CORAL	0.8978	0.5263	0.5404	4.9018	0.9000	0.5465	0.6491	**0.4211**
Full HoloGene	**0.9002**	**0.5789**	**0.5905**	**4.8994**	**0.9050**	0.5499	0.6531	0.4198

Overall, these ablations indicate a multi-objective trade-off rather than monotonic improvement from every added component. Compared with FiLM-only, FiLM+CORAL improves the measured spatial and representation diagnostics, whereas PSN improves selected signature-aware reconstruction and biological-preservation readouts. Full HoloGene offers a favorable balance across the primary reconstruction, clustering, and downstream biological metrics, although it is not optimal for every diagnostic individually.

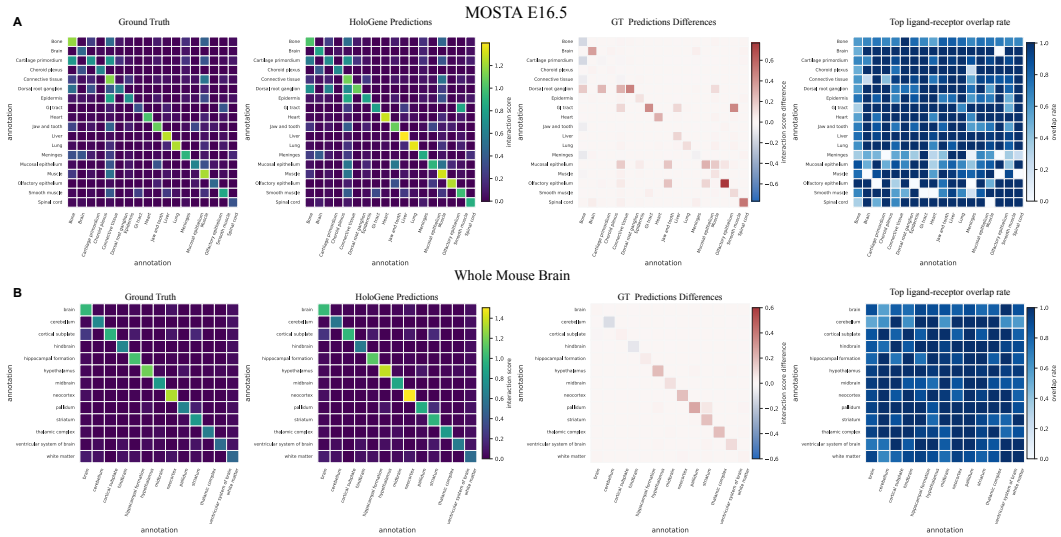


Figure 10: Annotation-referenced cell–cell communication (CCC) preservation on the two primary datasets. Rows and columns denote anatomical annotations used for post hoc sender–receiver grouping. For each dataset, we compare ground-truth CCC interaction scores, HoloGene-predicted CCC interaction scores, prediction-minus-ground-truth residuals, and top ligand–receptor interaction overlap rates. (A) MOSTA E16.5. (B) Whole mouse brain. HoloGene recovers similar label-pair communication structures to the ground truth, with small residual differences and high overlap of top ligand–receptor interactions across most annotation pairs.

E Effect of Train/Test Split Ratio

We study the performance of HoloGene under different train/test split ratios with MOSTA E16.5 (Chen et al., 2022). For each split, we train the model on the training slices and evaluate it on held-out test slices using the same protocol as in the main experiments. Table 9 summarizes reconstruction, clustering, and biological preservation metrics under each split.

Table 9: MOSTA E16.5 performance under different train/test split ratios. The 8:2 split corresponds to the default setting. Subscript m denotes metrics evaluated only where ground-truth expression is positive. Higher is better except MSE.

Train:Test	$\text{MSE}_m \downarrow$	$\text{Cos}_m \uparrow$	Pearson↑	$\text{Spear}_m \uparrow$	ARI↑	Path. F1↑	CCC Pearson↑	CCC Top-F1↑	DEG F1↑
8:2	0.2303	0.8054	0.7146	0.5503	0.1637	0.7239	0.9342	0.9138	0.6278
6:4	0.2266	0.8039	0.7131	0.5485	0.0988	0.6388	0.9187	0.8996	0.5024
5:5	0.3360	0.7310	0.6821	0.5246	0.0854	0.5422	0.9252	0.8651	0.3643
4:6	0.2770	0.7490	0.6808	0.5312	0.0981	0.5996	0.9254	0.8894	0.4341
2:8	0.2588	0.7782	0.6927	0.5384	0.0870	0.5170	0.9185	0.9036	0.3667

F Robustness to Axial Sampling Sparsity

The random train/test splits in Table 9 vary the number of observed sections but do not isolate the axial spacing between them. We therefore conduct a controlled-stride analysis in which every k-th measured section is observed and all remaining sections are held out. This directly controls the axial gap while retaining evaluation against measured expression on held-out sections. On MOSTA E16.5, strides $k = 2, 3, 4$ retain 7/13, 5/13, and 4/13 sections, respectively. The MOSTA results in Table 10 are averaged over three training seeds using the same stride layout. The 150-section whole mouse brain volume supports a wider range of strides, so we additionally evaluate $k = 2, 4, 6, 8, 10$, with three training seeds at each fixed stride layout.

Table 10: Controlled axial-stride evaluation. Every k-th section is observed and the remaining measured sections are held out. Results are mean $\pm$ standard deviation over three training seeds at each fixed stride layout for both datasets. Higher is better except MSE.

Dataset	Stride	$\mathrm{MSE}_m \downarrow$	Pearson↑	ARI↑	Path. F1↑
MOSTA E16.5	$k=2$	0.2344 ± 0.0018	0.7216 ± 0.0007	0.1225 ± 0.0014	0.6870 ± 0.0102
	$k=3$	0.2488 ± 0.0018	0.7160 ± 0.0009	0.1034 ± 0.0039	0.6916 ± 0.0121
	$k=4$	0.2250 ± 0.0019	0.7064 ± 0.0004	0.0976 ± 0.0039	0.6086 ± 0.0384
Whole mouse brain	$k=2$	2.5275 ± 0.0100	0.4657 ± 0.0004	0.3524 ± 0.0336	0.5483 ± 0.1100
	$k=4$	2.5003 ± 0.0385	0.4543 ± 0.0007	0.3363 ± 0.0110	0.5057 ± 0.0866
	$k=6$	2.5073 ± 0.0061	0.4476 ± 0.0001	0.3078 ± 0.0460	0.5387 ± 0.1810
	$k=8$	2.5592 ± 0.0163	0.4352 ± 0.0002	0.2679 ± 0.0146	0.4377 ± 0.0388
	$k=10$	2.5838 ± 0.0455	0.4280 ± 0.0026	0.2811 ± 0.0019	0.4983 ± 0.0126

On MOSTA E16.5, Pearson decreases from 0.722 at $k=2$ to 0.706 at $k=4$, and ARI decreases from 0.1225 to 0.0976; MSE and Pathway F1 are not monotonic across the tested strides. On whole mouse brain, Pearson decreases from 0.466 at $k=2$ to 0.428 at $k=10$, while ARI decreases from 0.352 to 0.281. The other metrics fluctuate across strides, reflecting the different held-out sections as well as axial spacing.

To separate aggregate stride performance from position within a missing interval, we reuse each trained stride model and group its held-out sections by distance to the nearest observed section, measured in section intervals. No model is retrained for an individual distance group. At $k=2$ and $k=3$ on MOSTA E16.5, all held-out sections have distance 1; $k=4$ is the first nontrivial case and contains distances 1 and 2. For the longer whole mouse brain volume, $k=6,8,10$ provide distances up to 5.

Table 11: Within-stride performance stratified by distance to the nearest observed section. Results are mean $\pm$ standard deviation over the same three training seeds for both datasets. Higher is better except MSE.

Dataset	Stride	Distance	$\mathrm{MSE}_m \downarrow$	Pearson↑	ARI↑	Path. F1↑
MOSTA E16.5	$k=4$	1	0.2315 ± 0.0019	0.7084 ± 0.0004	0.1036 ± 0.0048	0.6351 ± 0.0278
		2	0.2114 ± 0.0020	0.7024 ± 0.0006	0.0958 ± 0.0050	0.6790 ± 0.0171
Whole mouse brain	$k=6$	1	2.5449 ± 0.0172	0.4497 ± 0.0004	0.3128 ± 0.0340	0.5363 ± 0.1363
		2	2.4081 ± 0.0152	0.4526 ± 0.0004	0.3100 ± 0.0331	0.6237 ± 0.0538
		3	2.6961 ± 0.0179	0.4329 ± 0.0010	0.3097 ± 0.0306	0.6254 ± 0.1478
	$k=8$	1	2.6804 ± 0.0204	0.4367 ± 0.0001	0.2719 ± 0.0196	0.6608 ± 0.0220
		2	2.5129 ± 0.0208	0.4380 ± 0.0007	0.2817 ± 0.0204	0.7265 ± 0.1123
		3	2.5870 ± 0.0224	0.4293 ± 0.0011	0.2804 ± 0.0200	0.4490 ± 0.0630
		4	2.3339 ± 0.0210	0.4413 ± 0.0012	0.2783 ± 0.0185	0.3274 ± 0.0439
	$k=10$	1	2.5715 ± 0.0366	0.4458 ± 0.0015	0.2961 ± 0.0130	0.5739 ± 0.0609
		2	2.4384 ± 0.0366	0.4407 ± 0.0016	0.2951 ± 0.0110	0.3842 ± 0.0730
		3	2.6017 ± 0.0383	0.4175 ± 0.0018	0.2912 ± 0.0110	0.5640 ± 0.0423
		4	2.5466 ± 0.0391	0.4199 ± 0.0023	0.2786 ± 0.0067	0.5640 ± 0.0147
		5	2.8928 ± 0.0436	0.4004 ± 0.0026	0.2649 ± 0.0105	0.2219 ± 0.0085

For MOSTA E16.5 at $k=4$, the mid-gap distance 2 group has lower MSE (0.211 versus 0.232) and higher Pathway F1 (0.679 versus 0.635), but lower Pearson and ARI than distance 1. Whole mouse brain also exhibits non-monotonic distance effects at $k=6$ and $k=8$. The $k=10$ midpoint is a clearer limitation: at distance 5, MSE reaches 2.893, Pearson is 0.400, and Pathway F1 is 0.222. Different distance groups contain different anatomical sections, so distance alone does not determine performance.

Finally, we quantify variability due to section selection in the random-split setting of Table 9. At each held-out ratio from 20% to 60%, we sample three independent section partitions. The results in Table 12 therefore summarize partition variability rather than training-seed variability.

Table 12: Variability across random MOSTA E16.5 section partitions. Results are mean ± standard deviation over three independently sampled partitions at each held-out ratio. Higher is better except MSE.

Held out	Observed	$\mathrm{MSE}_m \downarrow$	Pearson↑	ARI↑	Path. F1↑
20%	10	0.2403 ± 0.0377	0.7200 ± 0.0122	0.1548 ± 0.0142	0.7289 ± 0.0295
30%	9	0.2379 ± 0.0331	0.7237 ± 0.0118	0.1346 ± 0.0139	0.7072 ± 0.0488
40%	7	0.2470 ± 0.0201	0.7061 ± 0.0183	0.1026 ± 0.0086	0.6339 ± 0.0043
50%	6	0.2813 ± 0.0474	0.6948 ± 0.0149	0.0995 ± 0.0144	0.5971 ± 0.0523
60%	5	0.2539 ± 0.0211	0.7021 ± 0.0185	0.0941 ± 0.0003	0.6229 ± 0.0595

Together, the controlled-stride, distance-stratified, and repeated-partition analyses distinguish the effect of axial gap from the number and identity of observed sections. Pearson and domain recovery generally decline over the broader stride range, but the individual metrics need not vary monotonically. The most sparsely observed whole-brain midpoint illustrates a practical limitation under very sparse axial coverage.

G Robustness to Registration and Pseudo-Label Perturbations

Residual registration errors. Starting from the supplied alignment of MOSTA E16.5, we independently perturb every training and held-out section while keeping expression values and the section split fixed. All spots in one section receive the same rigid transform, whereas perturbation directions and angles are sampled independently across sections. Translation magnitudes follow the percentage settings used in the perturbation runs, rotations are measured in degrees, and axial jitter is expressed relative to the inter-section spacing. We then repeat the original training and held-out evaluation protocol. Table 13 reports mean ± standard deviation over three random seeds for all conditions, including the unperturbed reference.

Table 13: Residual-registration stress test on MOSTA E16.5. All results, including the unperturbed reference, are mean ± standard deviation over three random seeds. Higher is better except MSE.

Perturbation	$\mathrm{MSE}_m \downarrow$	Pearson↑	ARI↑
None	0.2369 ± 0.0063	0.7115 ± 0.0003	0.1567 ± 0.0010
Translation 1%	0.2340 ± 0.0016	0.7091 ± 0.0011	0.1515 ± 0.0045
Translation 5%	0.2460 ± 0.0049	0.6877 ± 0.0081	0.1031 ± 0.0210
Translation 10%	0.2556 ± 0.0088	0.6719 ± 0.0141	0.0780 ± 0.0078
Rotation 1°	0.2318 ± 0.0012	0.7123 ± 0.0010	0.1553 ± 0.0033
Rotation 5°	0.2362 ± 0.0030	0.7028 ± 0.0026	0.1404 ± 0.0092
Rotation 10°	0.2417 ± 0.0043	0.6938 ± 0.0050	0.1185 ± 0.0096
Axial jitter $0.25\times$ spacing	0.2346 ± 0.0094	0.7109 ± 0.0014	0.1631 ± 0.0167
Axial jitter $1\times$ spacing	0.2391 ± 0.0121	0.7109 ± 0.0036	0.1582 ± 0.0138
Translation 3% + rotation 3°	0.2397 ± 0.0035	0.6973 ± 0.0053	0.1194 ± 0.0133
Translation 10% + rotation 10°	0.2575 ± 0.0044	0.6713 ± 0.0144	0.0737 ± 0.0152

Performance generally degrades as translation or rotation magnitude increases, while small perturbations remain close to the unperturbed reference. Pearson correlation is comparatively stable under axial jitter. This experiment probes residual errors around a supplied registration and does not establish robustness to completely unregistered or non-rigidly distorted data.

Pseudo-label resolution and corruption. The Leiden groups used to supervise PSN and stratify CORAL are constructed from training-section expression only. They are distinct from the dataset-provided annotations used to compute ARI after prediction. We first vary the Leiden resolution used for these training-only groups. As shown in Table 14, reconstruction metrics vary within a narrow range, while annotation-referenced ARI is more sensitive to group granularity.

Table 14: Sensitivity to the Leiden resolution used for training-only PSN and CORAL groups on MOSTA E16.5. Higher is better except MSE.

Resolution	$\mathrm{MSE}_m \downarrow$	Pearson↑	ARI↑
0.5	0.2307	0.7133	0.1594
0.8	0.2351	0.7129	0.1657
1.0	0.2303	0.7146	0.1637
1.2	0.2354	0.7106	0.1590
1.5	0.2330	0.7150	0.1573

We next randomly shuffle 3%, 5%, or 10% of training pseudo-label assignments while fixing the section split and training seed. Nonzero corruption conditions are averaged over independent corruption shuffles. The uncorrupted reference and corruption runs do not share the same repetition source: the uncertainty available for the former comes from training-seed variation, whereas the latter varies corruption shuffles at a fixed seed. Table 15 reports mean $\pm$ standard deviation, with these distinct repetition sources identified explicitly.

Table 15: Sensitivity to shuffled training pseudo-label assignments on MOSTA E16.5. Results are mean $\pm$ standard deviation. The uncorrupted reference varies training seeds; nonzero corruption conditions vary label shuffles at a fixed training seed. Higher is better except MSE.

Corrupted labels	$\mathrm{MSE}_m \downarrow$	Pearson↑	ARI↑
0%	0.2278 ± 0.0032	0.7140 ± 0.0022	0.1691 ± 0.0049
3%	0.2306 ± 0.0000	0.7140 ± 0.0005	0.1590 ± 0.0013
5%	0.2337 ± 0.0014	0.7134 ± 0.0006	0.1529 ± 0.0051
10%	0.2325 ± 0.0028	0.7136 ± 0.0013	0.1626 ± 0.0004

Across the evaluated resolution and corruption ranges, expression-level reconstruction remains comparatively stable, while annotation-referenced domain recovery varies more. These controlled tests do not imply robustness beyond the evaluated pseudo-label settings.

H Repeated-Run Robustness

We separate two sources of variability on MOSTA E16.5. For training stochasticity, every method uses one fixed section split and is trained with five seeds. For section-selection variability, every method uses the same five independently sampled train/held-out section partitions and one fixed training seed. Thus panels (a) and (b) of Table 16 should not be interpreted as repeated estimates of the same randomness source.

Table 16: Reconstruction and clustering robustness on MOSTA E16.5. Results are mean ± standard deviation over five training seeds at a fixed split or five section splits at a fixed training seed. Higher is better except MSE.

(a) Five training seeds, fixed section split

Method	$\mathrm{MSE}_m \downarrow$	Pearson↑	ARI↑
FFN	0.3131 ± 0.0025	0.6583 ± 0.0134	0.1235 ± 0.0084
SIREN	0.2764 ± 0.0093	0.6657 ± 0.0041	0.1092 ± 0.0113
STAGE	0.3100 ± 0.0038	0.5788 ± 0.0106	0.1087 ± 0.0140
SpatialZ	0.3131 ± 0.0091	0.5308 ± 0.0021	0.1468 ± 0.0041
STINR	$\mathbf{0.2246 \pm 0.0032}$	0.6950 ± 0.0112	0.1339 ± 0.0102
SUICA-3D	0.3016 ± 0.0016	0.6711 ± 0.0107	0.1352 ± 0.0023
HoloGene	0.2333 ± 0.0067	$\mathbf{0.7121 \pm 0.0008}$	$\mathbf{0.1533 \pm 0.0065}$

(b) Five section splits, fixed training seed

Method	$\mathrm{MSE}_m \downarrow$	Pearson↑	ARI↑
FFN	0.3070 ± 0.0093	0.6663 ± 0.0114	0.0989 ± 0.0323
SIREN	0.2777 ± 0.0204	0.6697 ± 0.0123	0.1096 ± 0.0145
STAGE	0.3037 ± 0.0145	0.5708 ± 0.0101	0.1141 ± 0.0119
SpatialZ	0.3243 ± 0.0096	0.5296 ± 0.0069	0.1287 ± 0.0210
STINR	0.2495 ± 0.0138	0.7086 ± 0.0085	0.1284 ± 0.0101
SUICA-3D	0.3294 ± 0.0145	0.6725 ± 0.0094	0.1381 ± 0.0041
HoloGene	$\mathbf{0.2361 \pm 0.0109}$	$\mathbf{0.7191 \pm 0.0092}$	$\mathbf{0.1547 \pm 0.0149}$

Table 17: Biological-readout robustness over five training seeds at a fixed section split on MOSTA E16.5. Results are mean ± standard deviation.

Method	Path. F1↑	CCC Pearson↑	CCC Top-F1↑	DEG F1↑
FFN	0.113 ± 0.019	0.871 ± 0.033	0.802 ± 0.074	0.341 ± 0.012
SIREN	0.404 ± 0.014	0.901 ± 0.009	0.688 ± 0.011	0.410 ± 0.031
STAGE	0.384 ± 0.052	0.885 ± 0.020	0.890 ± 0.012	0.280 ± 0.022
SpatialZ	0.673 ± 0.019	0.902 ± 0.013	0.889 ± 0.025	0.552 ± 0.012
STINR	0.656 ± 0.041	0.870 ± 0.016	0.868 ± 0.016	0.497 ± 0.008
SUICA-3D	0.509 ± 0.012	0.847 ± 0.004	0.849 ± 0.001	0.402 ± 0.010
HoloGene	$\mathbf{0.681 \pm 0.037}$	$\mathbf{0.928 \pm 0.004}$	$\mathbf{0.903 \pm 0.006}$	$\mathbf{0.600 \pm 0.015}$

Table 18: Biological-readout robustness over five section splits at a fixed training seed on MOSTA E16.5. Results are mean ± standard deviation.

Method	Path. F1↑	CCC Pearson↑	CCC Top-F1↑	DEG F1↑
FFN	0.327 ± 0.219	0.872 ± 0.041	0.854 ± 0.049	0.358 ± 0.013
SIREN	0.430 ± 0.047	0.915 ± 0.026	0.710 ± 0.041	0.433 ± 0.028
STAGE	0.382 ± 0.014	0.916 ± 0.034	0.905 ± 0.013	0.306 ± 0.019
SpatialZ	0.642 ± 0.030	0.896 ± 0.026	0.868 ± 0.016	0.576 ± 0.053
STINR	0.638 ± 0.030	0.909 ± 0.030	0.890 ± 0.012	0.459 ± 0.020
SUICA-3D	0.526 ± 0.006	0.869 ± 0.021	0.859 ± 0.020	0.407 ± 0.009
HoloGene	$\mathbf{0.703 \pm 0.042}$	$\mathbf{0.948 \pm 0.017}$	$\mathbf{0.919 \pm 0.009}$	$\mathbf{0.588 \pm 0.030}$

Across both analyses, HoloGene has the highest Pearson and ARI and the strongest four biological readouts; STINR has the lowest MSE under fixed-split training-seed variation, while HoloGene has the lowest MSE across section splits. These results characterize the two evaluated variability sources separately rather than implying invariance to either source.

I Runtime

On MOSTA E16.5 (downsampled to 10% per slice, single A100), training times (seconds) are: FFN 1320, SIREN 1305, STAGE 542, SpatialZ 791, STINR 2262, SUICA 2083, and HoloGene 2040 for the PSN/CORAL/FiLM synthesis stage. GAE pretraining takes 636s. HoloGene's main synthesis stage is comparable to STINR and SUICA, and the additional GAE pretraining stage is a one-time cost for latent manifold construction.

J Dataset Licenses and Availability

We use previously released datasets and cite the corresponding original publications in the main paper. The MOSTA E16.5 dataset is distributed under CC BY 4.0. The whole mouse brain dataset is distributed under CC BY 4.0. The Breast Cancer dataset from Kuett *et al*. is distributed under CC BY 4.0. The 3D Visual Cortex dataset is publicly available, but we did not identify an official explicit license statement. We do not redistribute the original datasets as new assets in this submission.

K Biological Preservation Metrics

This section details the biological preservation readouts used in Table 2. For all metrics, prediction-derived statistics are compared with the matched ground-truth expression matrix on held-out test cells. Higher is better for all biological preservation metrics. ARI uses dataset-provided cell-type or anatomical annotations as held-out evaluation targets. DEG F1, CCC metrics, and pathway F1 are also computed based on these dataset-provided annotations. These annotations are used only for evaluation and visualization; they are not used for model training, PSN supervision, CORAL regularization, model selection, or hyperparameter tuning.

Pathway enrichment resources. Pathway enrichment is performed with the g:Profiler g:GOSt API using the mouse organism setting (`mmusculus`) (Kolberg et al., 2023). We use GO Biological Process (Consortium, 2026), Reactome (Ragueneau et al., 2026), WikiPathways (Agrawal et al., 2024), and KEGG (Kanehisa et al., 2023) with FDR correction at threshold 0.05. For each method and dataset, the background is the evaluated gene universe rather than the whole mouse genome. We retain terms with adjusted $p < 0.05$, term size at most 1000, intersection size at least 15, and at most the top 20 enriched terms per label; labels are retained only when the ground truth has at least three significant pathways. Pathway F1 is reported for MOSTA E16.5 and whole mouse brain, but not for 3D Visual Cortex because its 155-gene panel is too small for stable enrichment.

Ligand-receptor resource. Cell-cell communication (CCC) metrics use the mouse ligand-receptor table from CellTalkDB v1.0 (Shao et al., 2021). The source database contains 2033 mouse ligand-receptor pairs. For each dataset, we retain only pairs whose ligand and receptor genes are both present in the evaluated gene panel.

Top20 DEG micro F1. For each dataset-provided annotation label l, we perform one-vs-rest differential expression analysis separately on the ground-truth expression matrix and the predicted expression matrix. We retain the top 20 up-regulated marker genes per label, denoted G_l^{GT} and G_l^{Pred}. The micro-averaged precision and recall are

$$\text{Precision}_{micro} = \frac{\sum_l |G_l^{Pred} \cap G_l^{GT}|}{\sum_l |G_l^{Pred}|}, \qquad \text{Recall}_{micro} = \frac{\sum_l |G_l^{Pred} \cap G_l^{GT}|}{\sum_l |G_l^{GT}|}. \tag{5}$$

The reported score is the corresponding micro F1. This metric measures whether the model preserves label-specific marker genes.

CCC interaction Pearson. For each dataset-provided annotation label l and gene g, let $\mu_{l,g}^M$ be the mean expression for method $M \in \{GT, Pred\}$ among cells with label l. For sender annotation s, receiver annotation r, and ligand-receptor pair $p = (g_L, g_R)$, the communication score is

$$S_{s,r,p}^M = \mu_{s,g_L}^M \cdot \mu_{r,g_R}^M \cdot A_{s,r}, \tag{6}$$

where $A_{s,r}$ is the spatial adjacency weight estimated from the spatial nearest-neighbor graph. CCC interaction Pearson is the Pearson correlation between flattened predicted and ground-truth CCC score tensors, computed over the union of interactions with nonzero predicted or ground-truth scores.

CCC top interaction F1. We also evaluate whether the strongest communication events are recovered. The candidate universe contains all tuples $u = (s, r, p)$, where s and r are dataset-provided annotation labels and p is a retained ligand-receptor pair. Let U be the set of candidates with nonzero predicted or ground-truth communication scores. We set

$$K = \min(5000,\ \max(25,\ \lceil 0.05|U| \rceil)). \tag{7}$$

Ground truth and prediction are ranked independently over the same candidate universe, and the top-K sets T_K^{GT} and T_K^{Pred} are compared by set-overlap F1. The same rule for K is used for all methods within a dataset.

Pathway micro F1. For each dataset-provided annotation label l, we compare the set of significant enriched pathway terms in ground truth and prediction:

$$P_l^{GT} = \{(\text{direction}, \text{source}, \text{term id})\}_{GT,l}, \qquad P_l^{Pred} = \{(\text{direction}, \text{source}, \text{term id})\}_{Pred,l}. \tag{8}$$

Micro precision and recall pool pathway overlaps across retained labels, and the reported pathway score is the corresponding micro F1.

Label-independent spatial diagnostics. To complement the annotation-referenced biological preservation metrics above, we also report label-independent diagnostics on the whole mouse brain dataset in Table 19. Gene Moran Pearson computes Moran's I for each gene on the ground-truth and predicted expression fields and reports the Pearson correlation across genes, measuring whether spatial autocorrelation patterns are preserved without using cluster labels. Gene gradient-z Pearson computes, for each gene, the Pearson correlation between expression and the z-coordinate, then correlates these per-gene scores between ground truth and prediction. Pathway mean Pearson computes per-pathway Pearson correlation of cell-wise z-scored Reactome pathway scores between ground truth and prediction; pathway neighborhood Pearson applies the same comparison after 3D spatial neighborhood averaging with a $k = 15$ nearest-neighbor graph.

The same comparison also includes two representation probes. Slice leakage is the linear-probe accuracy for predicting slice identity within annotation groups; lower values indicate less slice-identity information. Annotation retention is the linear-probe accuracy for predicting anatomical annotation; higher values indicate greater retained biological information. These annotations are used only for post hoc probing, not to train the representations.

Table 19: Spatial biological and representation diagnostics on whole mouse brain. Slice leakage is lower-is-better; the other metrics are higher-is-better. Bold denotes the better score.

Model	Moran Pear.↑	Gene Grad. z ↑	Path. Mean↑	Path. Nbhd.↑	Slice Leak↓	Ann. Ret.↑
FiLM-only	0.6040	0.9220	0.2655	0.6227	0.1966	0.4929
FiLM + CORAL	**0.6103**	**0.9326**	**0.2662**	**0.6244**	**0.1770**	**0.5102**

Axial-specific preservation diagnostics. To test whether reduced slice leakage is accompanied by loss of axial biology, we isolate CORAL in the PSN-conditioned whole-mouse-brain model: "HoloGene w/o CORAL" uses FiLM+PSN, whereas full HoloGene uses FiLM+PSN+CORAL. Gene-z Spearman measures preservation of per-gene monotonic axial profiles. Within each cell type, pathway axial direction agreement compares the direction of pathway activation changes along z, pathway axial trajectory Pearson compares the full axial trajectories, and pathway axial signal ratio measures retained variance in the fitted linear axial component.

Table 20: Axial-specific preservation diagnostics on whole mouse brain. All metrics are higher-is-better.

Diagnostic	HoloGene w/o CORAL	Full HoloGene
Gene-z Spearman	0.9200	**0.9479**
Pathway axial direction agreement	0.8169	**0.8251**
Pathway axial trajectory Pearson	0.4629	**0.4740**
Pathway axial signal ratio	0.4893	**0.5145**

All four measured scores improve when CORAL is enabled in this comparison. Separately, the FiLM-only versus FiLM+CORAL comparison in Table 19 improves all four label-independent spatial diagnostics, reduces slice leakage, and increases annotation retention. These complementary comparisons support CORAL on the evaluated tissue without implying a guarantee for other tissues or sampling regimes.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and introduction describe the method scope, the slice-wise 3D ST setting, and the empirical claims evaluated in the experiments. The conclusion also states the main assumptions and limitations.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: The conclusion includes a limitations paragraph covering model-based virtual slices, lack of uncertainty quantification, the pre-registration assumption, slice spacing, and dependence on clustering quality.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A].

Justification: The paper proposes and evaluates an empirical method and does not claim new theoretical results or include theorem-style proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes].

Justification: The paper and appendix describe the method and experimental setup. Code is available at `https://github.com/Szym29/HoloGene`.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code is available at `https://github.com/Szym29/HoloGene`. The paper uses existing public 3D spatial transcriptomics datasets, cites their original publications, and summarizes dataset licenses and availability in Appendix J.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: The experimental section specifies the slice-wise 80/20 split, baseline adaptations, k-NN graph setting, and the search ranges for learning rate, batch size, and latent dimension.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: On MOSTA E16.5, Appendix H reports mean $\pm$ standard deviation for reconstruction, clustering, and biological readouts over five training seeds at a fixed section split and five section splits at a fixed training seed. The registration stress test in Appendix G also reports mean $\pm$ standard deviation over three random seeds. These quantify variability in the specified experiments; not all results across all datasets are repeated, and we do not report formal significance tests.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes].

   Justification: The experiments state that runs were conducted on a single NVIDIA A100 GPU and report representative training times for all compared methods on E16.5. Exact memory usage is not separately profiled.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The work uses previously released spatial transcriptomics datasets and performs computational method development and benchmarking. We are not aware of any aspect of the work that deviates from the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [No].

Justification: The paper does not include a dedicated broader impacts discussion. It focuses on the technical method and reports methodological limitations, including that reconstructed signals are model-based estimates requiring independent experimental validation.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A].

Justification: The paper does not release a high-risk model, scraped dataset, or generative system with evident dual-use safety concerns.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes].

Justification: The datasets and baseline methods are credited through citations, and dataset license/availability information is summarized in Appendix J. MOSTA E16.5, whole mouse brain, and the Kuett *et al*. Breast Cancer dataset are under CC BY 4.0; 3D Visual Cortex is publicly available, but an official explicit license statement was not identified.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The HoloGene code repository is available at `https://github.com/Szym29/HoloGene`.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A].

Justification: The paper does not involve crowdsourcing experiments or direct interaction with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: The study is a secondary computational analysis of previously released datasets and does not recruit participants or collect new human-subject data.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A].

    Justification: LLM-based tools assisted with manuscript preparation and the implementation, debugging, and execution of experimental code. They are not components of the proposed HoloGene model. The authors reviewed the resulting code and experimental outputs and take responsibility for the reported results.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.