stDIFFUSION: A DIFFUSION BASED MODEL FOR GENERATIVE SPATIAL TRANSCRIPTOMICS

Sumeer A. Khan[1,2+] , Xabier Martínez-de-Morentin [1+] , Robert Lehmann[1], Vincenzo Lagani[1,2], Narsis A. Kiani[4], David Gomez-Cabrero[1], Jesper Tegnér[1,3,5,6 ‡]
[1]Biological and Environmental Science and Engineering Division (BESE)
King Abdullah University of Science and Technology (KAUST)
Thuwal,Saudi Arabia.
[4]Algorithmic Dynamic Lab, Department of Oncology and Pathology
Karolinska Institute, Stockholm, Sweden
{narsis.kiani}@ki.se

ABSTRACT

Spatial Transcriptomics (ST) allows deep characterization of the 2D organization of expression data within tissue slices. The ST technology provides a tissue contextualization of deep single-cell profiles. Recently, numerous computational and machine learning methods have addressed challenges such as data quality, augmentation, annotation, and the development of integrative platforms for data analysis. In contrast, here we ask whether unseen spatial transcriptomics data can be predicted and if we can interpolate novel transcriptomic slices. To this end, we adopt a denoising diffusion probabilistic-based model (DDPM) to demonstrate the learning of generative ST models for several tissues. Furthermore, our generative diffusion model interpolates (predicts) unseen slices located "between" the collected finite number of ST slices. This methodology sets the stage for learning predictive deep 3D models of tissues from a finite number of spatial transcriptomics slices, thus heralding the advent of AI-augmented spatial transcriptomics.

1 INTRODUCTION

The twin rise of single-cell genomics, producing high-resolution atlases of different cells, and spatial transcriptomics (ST), uncovering the cellular transcriptional organization within tissues, impacts fundamental cell biology and translational research (Vandereyken et al., 2023) . Recently, a wide range of computational and machine learning methods target challenges such as data quality and augmentation (imputation, normalization), annotation (deconvolution of cell types, clustering, resolution), and tissue analysis (spatial gradients, extracting hierarchical organization), and development of integrative platforms (software, databases, data management)(Zeng et al., 2022) (Fang et al., 2023) (Palla et al., 2022a).

While the bulk of the work has been 2D Spatial Transcriptomics at different length scales (Palla et al., 2022a), increasing efforts are homing to the potential of 3D Spatial Transcriptomics. Recently, progress has been achieved in aligning ST slices using probabilistic models (Zeira et al., 2022) or aligning slices using a 3D neighborhood graph into a local coordinate system (Fang et al., 2023) (Wang et al., 2023a).

In this work, we develop a generative ST model applicable to diverse tissue types. Specifically, for a given finite set of ST slices, we ask if we can predict or interpolate unseen ST slices beyond a finite set of ST slices (Figure 1a). To this end, we first assessed whether a denoising diffusion probabilistic

[*]S.A.K. and X.M.M. contributed equally to this work.

[†]Correspondence to: sumeer.khan@kaust.edu.sa, jesper.tegner@kaust.edu.sa.

[‡]SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence[2],CEMSE KAUST[3], Unit of Computational Medicine, Department of Medicine, Center for Molecular Medicine, Karolinska Institutet, Karolinska University Hospital, Stockholm, Sweden[5],
Science for Life Laboratory, Tomtebodavägen, Solna, Sweden.[6]

model (Ho et al., 2020a) (DDPM) could serve as a generative model for ST. The trilemma – the balancing act between quality, diversity, and speed - of generative models (Xiao et al., 2022) means that both variational encoders and normalizing flows suffer from lower quality. In contrast, Generative Adversarial Networks are prone to mode collapse, thus rendering them less suitable for producing diversity. Recently, the DDPM model (Ho et al., 2020a) has achieved a faster subsampling, becoming a practical model for learning via a forward noise process and neural network learning the reverse diffusion process (Figure 1b). We hypothesized that DDPM models are a suitable candidate for the reconstruction of ST slices as DDPM has recently achieved state-of-the-art in image generation (DALLE 2), super-resolution, image denoising, and inpainting (Yang et al., 2024).

2 RESULTS

To evaluate whether DDPM models are suitable for the reconstruction of ST slices, we trained a DDPM model (stDiffusion) (Figure 1b; see Methods) on MERFISH (Moffitt et al., 2018) slice data. For each spatial spot, we first extract its original gene–expression vector $x_0 \in \mathbb{R}^G$, compute a fixed cell–type embedding and project its two–dimensional coordinates into a spatial embedding (see Methods). During the forward diffusion pass, only x_0 is noised, while both cell–type embedding and spatial embedding remains unchanged. In the reverse (denoising) pass, we form the conditioned input(Figure, 1b) and feed it into our learnable network which predicts the noise at each timestep. We optimize the model by minimizing the mean–squared error(see Methods). We use spot-wise clustering to compare the structure of ground truth and generated data using the neighborhood enrichment test (Palla et al., 2022b) to identify spatially enriched clusters. Visualizing the clustering (Supp Fig S1a,b) demonstrates the model's ability to replicate the structure and relationships found in the reference data (Figure 1b). Embedding contextual features such as cell type and spatial coordinates enhances the model's predictive accuracy during reverse diffusion. Thus, we provide the first demonstration that a DDPM model can generate and reconstruct spatial transcriptomics data.

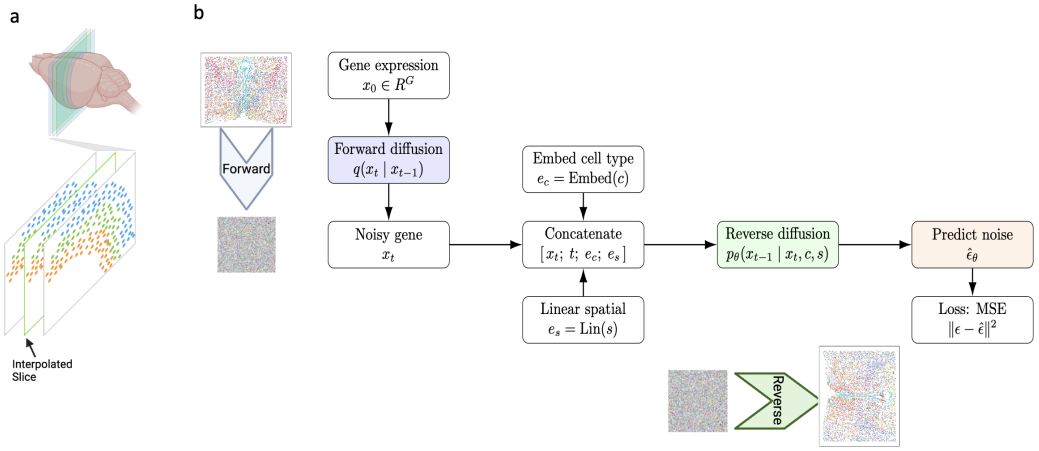

Figure 1: **Schematic of stDiffusion:** a) depiction of interpolation across slices of brain b) stDiffusion for in silico generation and interpolation of spatial transcriptomics data. Details in methods section.

To quantify how faithfully *stDiffusion* reconstructs spatial patterns, we apply Leiden clustering to both the ground–truth and generated spot profiles, and compute neighborhood–enrichment z–scores to capture spatial co–localization of clusters (Palla et al., 2022b). As shown in Supplementary Figure S1a–b and summarized in Figure 1b, *stDiffusion* successfully replicates the spatial organization of transcriptional clusters, demonstrating that embedding cell type and spatial context exclusively in the reverse diffusion steps substantially enhances reconstruction fidelity. This represents the first demonstration that a DDPM framework can both generate and accurately reconstruct spatially resolved transcriptomics data.

Next, we asked whether *stDiffusion* could generate unseen ST slices by interpolating within a given slice and between adjacent slices (Figure 1a). The idea is to harness the DDPM's ability to learn the underlying data distribution through gradual transformations from noise to data, thus effectively navigating a learned latent space. We trained *stDiffusion* on a subset of ST slices and then tested if non-border held-out slices could be predicted. We used three different datasets from various tissues and technologies.

We extract gene expression data and spatial coordinates for each slice based on layer and Bregma values, normalized for consistent distance calculations. Stochastic gene expression encoding prepares data for interpolation in the model's latent space, capturing biological variability. For each target slice spot, Euclidean distance identifies the closest points in adjacent reference slices, guiding feature selection (Methods). Interpolation blends noised feature representations using a tunable factor (λ), enabling controlled transitions between slices (Supp Fig 3–5). After interpolation in latent space, reverse diffusion reconstructs gene expression profiles. As a baseline, we compare against linear interpolation by averaging two reference slices (Methods).

2.1 stDIFFUSION INTERPOLATES ACROSS LAYERS OF HUMAN DORSOLATERAL PREFRONTAL CORTEX SLICE

We first challenged *stDiffusion* to interpolate within a slice of the human dorsolateral prefrontal cortex (DLPFC (Maynard et al., 2021), Methods) between layers. Using sample profiling layers $Layer_1$ through $Layer_6$ and white matter (WM), we aimed to predict the $Layer_4$ ST between $Layer_3$ and $Layer_5$ (Figure 2 a-c). stDiffusion performs remarkably well (Figure 2b,c) in interpolating layers that were not included during the initial and validation phases while preserving spatial neighborhoods compared to linear interpolation. Large neighborhood enrichment values along the diagonal (Figure 2b) indicate a robust within-cluster relationship, demonstrating that cells of the same type are located near each other. The pattern is very similar in the reference and interpolated data, suggesting that *stDiffusion* has effectively recapitulated actual spatial relationships in the real tissue. Additionally, it produced a strong positive association between interpolated and actual expression values (Supp Fig 2a), confirming that gene expression spatial arrangements closely resemble those in ground truth layers.

2.2 stDIFFUSION INTERPOLATES ACROSS LAYERS OF MOUSE VISUAL CORTEX STARMAP DATA SLICE

Next, we asked whether *stDiffusion* could navigate across the layers within a single slice of the mouse visual cortex. Here we used the Starmap (Wang et al., 2018) encompassing distinct cortical layers L1, L2/L3, L4, L5, and L6. The task is to interpolate the intermediate layer L2/L3 by interpolating between layers L1 and L4 (Figure 2 d-f), thus recapitulating the organization of clusters as observed in the real data. To this end, stDiffusion accurately reconstructs the omitted layer L2/L3 (Figure 2e). Notably, stDiffusion outperforms a linear interpolation method (Figure 2e,f), which falls short in capturing the spatial distribution and cluster neighborhood structure (Figure 2e, Supp Fig 2d). stDiffusion effectively (Supp Fig 2c) interpolates and generates gene expression profiles of the target layer.

2.3 stDIFFUSION INTERPOLATES BETWEEN THE SLICES IN MOUSE MERFISH DATA

We used mouse MERFISH(Moffitt et al., 2018) data from 12 consecutive hypothalamus preoptic region slices (Figure 3a-d). The results (Figure 3b,c,d) show that *stDiffusion* effectively interpolates between slices, preserving cluster neighborhood structures seen in the original data. For example, in the MERFISH data slice, the spatial structure of clusters like cluster 8 and cluster 11 is maintained in the interpolated slice, aligning with the ground truth slice (Figure 3d). This outperforms linear interpolation, which fails to preserve cluster structure (Figure 3d). Using Leiden clustering, we compare the spatial distribution of clusters in the ground truth and generated data. Visualizing clusterings in spatial coordinates (Figure 3b) demonstrates the model's ability to replicate the reference data's structure (Figure 3a). We also conducted a normalized Spearman correlation analysis between the expression levels of all genes in the ground truth and interpolated slices for each spot. The findings (Supp Fig 2e) indicate that stDiffusion maintains the spatial distribution of gene expression levels in the interpolated slice, consistent with the original data.

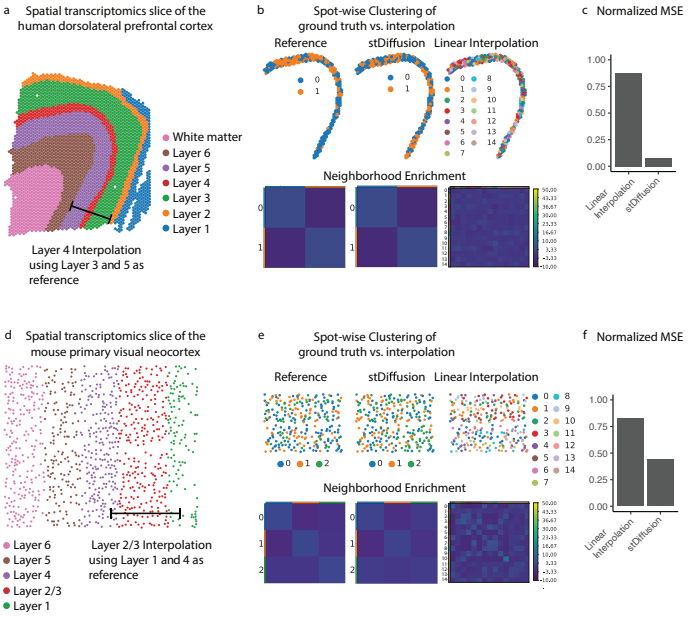

Figure 2: **Interpolation across human dorsolateral prefrontal cortex (DLPFC) with stDiffusion:** a) DLPFC slice visualization in spatial coordinates, b) Leiden clustering and neighborhood enrichment heatmap for ground truth $Layer_4$, showing two distinct clusters. *stDiffusion* replicates this pattern, unlike linear interpolation. Strong diagonal heatmap colors indicate within-cluster relationships, while off-diagonal elements show proximity between clusters. c) Normalized mean square errors between ground truth and interpolated slice, d) Starmap data visualization, e) Leiden clustering and neighborhood enrichment heatmap for ground truth (L2/3), where stDiffusion preserves spatial structure better than linear interpolation, f) Normalized mean square errors showing stDiffusion's superior performance.

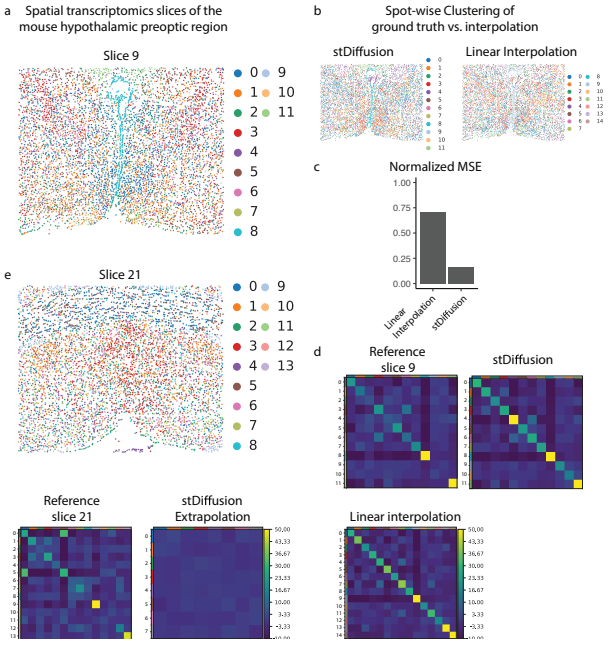

Figure 3: **Interpolation and extrapolation - MERFISH slices with stDiffusion:** a) Visualization of ground truth slices (Bregma 9, Bregma 21) in spatial coordinates; b) Leiden clustering of interpolated slices; c) normalized mean square errors showing stDiffusion outperforming linear interpolation; d) neighborhood enrichment for ground truth (Bregma 9), highlighting stDiffusion's ability to capture spatial structure (e.g., cluster 8 reconstruction); e) neighborhood enrichment heatmap for extrapolation, illustrating stDiffusion's limitations in out-of-distribution predictions.

In summary, *stDiffusion* learns ST data from a single slide and predicts held-out slices, effectively interpolating between a finite set of ST slices. We next tested whether it could extrapolate beyond the collected spatial region. Since out-of-distribution regions may differ statistically, we expected lower performance. Training stDiffusion on posterior brain slices while excluding anterior slice 21 (Figure 3e upper) confirmed this: extrapolation performance (Figure 3e lower) was weaker than interpolation. This controlled experiment highlights DDPM's predictive limits, emphasizing that extrapolation depends on distributional similarity. Our results suggest an experimental strategy that ensures diverse sampling for more effective predictions.

3 DISCUSSION

We used the *stDiffusion* model based on DDPM (Ho et al., 2020b) to tackle key challenges in spatial transcriptomics: in silico data generation and gene expression interpolation across tissue slices and layers. While extensive work has improved 2D spatial transcriptomics in areas like data quality, normalization, and clustering (Zeng et al., 2022) (Fang et al., 2023) (Palla et al., 2022a), aligning 2D slices in a 3D space remains limited by finite samples (Zeira et al., 2022) (Wang et al., 2023a) (Ortiz et al., 2020). Our approach extends generative modeling to predict ST data, paving the way for continuous 3D tissue reconstruction.A generative 3D spatial transcriptomics AI (Wang et al., 2023b) integrating multimodal cellular analysis could impact fundamental biology, organ modeling (Tegnér et al., 2009), and precision medicine (Zhang et al., 2022).

REFERENCES

K. Dong and S. Zhang. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nature Communications*, 13:1–12, 2022.

S. Fang et al. Computational approaches and challenges in spatial transcriptomics. *Genomics, Proteomics and Bioinformatics*, 21:24–47, 2023. doi: 10.1016/j.gpb.2022.10.001.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 2020-December. Neural Information Processing Systems Foundation, 2020a.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. Available at `https://github.com/hojonathanho/diffusion`, 2020b.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations (ICLR)*, 2019.

K. R. Maynard et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature Neuroscience*, 24:425–436, 2021.

J. R. Moffitt et al. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science*, 362, 2018.

C. Ortiz et al. Molecular atlas of the adult mouse brain. Available at `https://www.science.org`, 2020.

G. Palla, D. S. Fischer, A. Regev, and F. J. Theis. Spatial components of molecular tissue biology. *Nature Biotechnology*, 40:308–318, 2022a.

G. Palla et al. Squidpy: a scalable framework for spatial omics analysis. *Nature Methods*, 19: 171–178, 2022b.

J. N. Tegnér et al. Computational disease modeling - fact or fiction? *BMC Systems Biology*, 3, 2009.

K. Vandereyken, A. Sifrim, B. Thienpont, and T. Voet. Methods and applications for single-cell and spatial multi-omics. *Nature Reviews Genetics*, pp. 1–22, 2023. doi: 10.1038/s41576-023-00580-2.

G. Wang et al. Construction of a 3d whole organism spatial atlas by joint modelling of multiple slices with deep neural networks. *Nature Machine Intelligence*, 2023a. doi: 10.1038/s42256-023-00734-1.

H. Wang et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620:47–60, 2023b.

X. Wang et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*, 361, 2018.

F. A. Wolf, P. Angerer, and F. J. Theis. Scanpy: Large-scale single-cell gene expression data analysis. *Genome Biology*, 19:15, 2018.

Z. Xiao, K. Kreis, and A. Vahdat. Tackling the generative learning trilemma with denoising diffusion gans. In *International Conference on Learning Representations (ICLR)*, 2022.

L. Yang et al. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56:1–39, 2024.

R. Zeira, M. Land, A. Strzalkowski, and B. J. Raphael. Alignment and integration of spatial transcriptomics data. *Nature Methods*, 19:567–575, 2022.

Z. Zeng, Y. Li, Y. Li, and Y. Luo. Statistical and machine learning methods for spatially resolved transcriptomics data analysis. *Genome Biology*, 23, 2022. doi: 10.1186/s13059-022-02653-7.

J. Zhang et al. Spatiotemporal omics-refining the landscape of precision medicine. *Life Medicine*, 1:84–102, 2022.

A Appendix

A.1 Data Preprocessing

We developed and evaluated SpatialDiffusion (stDiffusion) using three different datasets. The MERFISH (Moffitt et al., 2018) dataset from (Palla et al., 2022b) was used to assess the in-silico generation and interpolation performance of stDiffusion across distinct slices. MERFISH consists of 12 consecutive slices from the mouse hypothalamic preoptic region. The dataset is segmented based on the Bregma reference point, allowing for the exclusion or inclusion of specific slices to tailor our training, validation, and test sets for focused analyses. Each instance comprises gene expression values, classified cell types, and two-dimensional spatial coordinates, facilitating a multifaceted approach to understanding spatial gene expression patterns. The dataset has 73655 spots and 161 genes per spot. The second dataset we used for interpolation across layers is the mouse visual cortex Starmap (Wang et al., 2018) data obtained from (Dong & Zhang, 2022) in the preprocessed format with 984 spots and 1020 genes per spot. This dataset has spots of gene expression profiles for L1, L2/L3, L4, L5, and L6 layers. The third dataset is the human dorsolateral prefrontal cortex from (DLPFC) (Maynard et al., 2021) (Dong & Zhang, 2022). Specifically, the DLPFC dataset includes 12 human DLPFC sections sampled from three individual experiments (Maynard et al., 2021). The number of spots ranges from 3498 to 4789 for each section, and the original authors have manually annotated the area of DLPFC layers and white matter (WM). We used one sample from this dataset, and we preprocessed with the SCANPY (Wolf et al., 2018) package, and used 3000 highly variable genes with 3431 spots as an input to our stDiffusion model. Additional information, such as spatial coordinates and layers, is included in the dataset. For DLPFC dataset, we applied our stDiffusion model to interpolate across layers, i.e., $Layer_1$, $Layer_2$, $Layer_3$, $Layer_4$, $Layer_5$, $Layer_6$, and white matter (WM).

A.2 The SpatialDiffusion Model

In spatial transcriptomics, stDiffusion (Figure 1b) adapts Denoising Diffusion Probabilistic Models (Ho et al., 2020a) (Ho et al., 2020b) (DDPM) principles to address the unique challenges posed by spatially resolved gene expression data. The model begins by learning the distribution of gene expression profiles across spatial coordinates, effectively capturing the spatial transcriptional landscape of the tissue. Through forward and reverse diffusion steps, stDiffusion interpolates missing

gene expression data within this landscape, providing a continuous view of transcriptional activity across the tissue. In spatial transcriptomics, stDiffusion (Figure 1b) adapts Denoising Diffusion Probabilistic Models (Ho et al., 2020a) (Ho et al., 2020b) (DDPM) principles to address the unique challenges posed by spatially resolved gene expression data. The model begins by learning the distribution of gene expression profiles across spatial coordinates, effectively capturing the spatial transcriptional landscape of the tissue. Through forward and reverse diffusion steps, stDiffusion interpolates missing gene expression data within this landscape, providing a continuous view of transcriptional activity across the tissue.

A.3 IN SILICO GENERATION

The in-silico generation of spatial transcriptomics data aims at augmenting existing datasets with synthetic yet biologically plausible data points. This approach enhances the dataset's density, providing a more continuous spatial representation of gene expression patterns. We trained the stDiffusion model on the preprocessed spatial transcriptomics data, allowing it to learn the underlying distribution of gene expression across spatial coordinates.

A.4 NETWORK ARCHITECTURE: DIFFUSION MODEL

The core of our approach is a denoising diffusion probabilistic model (DDPM) (Ho et al., 2020a) specifically designed to handle the complexities of spatial transcriptomics data. This model incorporates an embedding layer for cell types and a linear transformation for spatial coordinates, ensuring both are integral to the learning process (Figure 1b). An embedding layer for cell type classification allows the model to interpret cell types as dense vectors of a specified dimension. A linear transformation is applied to spatial coordinates, mapping them to a similarly dimensioned space as the cell-type embeddings. A concatenation of gene expression data with the transformed cell type and spatial information is followed by a sequential network comprising linear layers and activation functions.

A.5 NETWORK ARCHITECTURE: ENHANCEMENT OF MODEL INPUTS

Let x denote the gene expression data for a given sample, c denotes the cell type, and s represents the spatial coordinates. The cell type c is transformed through an embedding layer, and the spatial coordinates s are processed through a linear transformation to ensure they are in a compatible format for concatenation:

$$e_c = \mathrm{Embed}(c), \tag{1}$$

$$e_s = \mathrm{Lin}(s), \tag{2}$$

The enhanced input $x^{'}$ to the model is then a concatenation of the original gene expression data with the transformed cell type and spatial information:

$$x' = \left[x;\, e_c;\, e_s\right] \tag{3}$$

A.6 NETWORK ARCHITECTURE: FORWARD PROCESS (DIFFUSION)

The forward diffusion process gradually adds Gaussian noise to the data across several discrete time steps. The process at a specific time step t can be mathematically represented as follows:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\big(x_t;\, \sqrt{\bar{\alpha}_t}\, x_{t-1},\, (1-\bar{\alpha}_t)\, I\big), \quad \bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s. \tag{4}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ is the cumulative product of $\alpha_t = 1 - \beta_t$, with β_t being the variance of the noise added at step *t*. *I* is the identity matrix, ensuring the noise is independently added to each dimension.

A.7 NETWORK ARCHITECTURE: REVERSE PROCESS (DIFFUSION)

During the reverse process, the enhanced input x' is used to predict the original data from its noised version. The model utilizes the concatenated features to predict the mean μ of the reverse process distribution:

$$p(x_{t-1} \mid x_t, c, s, \theta) = \mathcal{N}(x_{t-1}; \mu_\theta(x', t), \sigma_t^2 I) \tag{5}$$

Where $\mu_\theta(x', t)$ depends on the enhanced input x', incorporating embeddings for cell type and spatial coordinates, along with the noised data x_t. Incorporating cell type and spatial coordinates into the stDiffusion allows the model to make more informed predictions during the reverse diffusion process. By embedding these contextual features and concatenating them with the gene expression data, the model can leverage the full spectrum of biological information available in the dataset. This procedure enhances the model's predictive accuracy and the biological relevance of the reconstructed data, facilitating a deeper understanding of the spatial patterns of gene expression within tissue samples.

A.8 NETWORK ARCHITECTURE: LOSS FUNCTION

The loss function aims to minimize the difference between the actual noise introduced during the forward diffusion process and the noise predicted by the model during reverse diffusion, incorporating cell type (c) and spatial information (s) as conditioning variables. To this end we use the Mean Squared Error (MSE) loss between the actual and predicted noise:

$$\mathcal{L}(\theta) = \mathbb{E}_{x_0 \epsilon, t, c, s} \left[\|\epsilon - \hat{\epsilon}_\theta(x_t, t, c, s)\|^2 \right] \tag{6}$$

Where ϵ is the actual noise added to the original data $x_{(0)}$ to obtain the noised version x_t, $\hat{\epsilon}_\theta(x_t, t, c, s)$ is the noise predicted by the model, and $x_{(0)}$ is the original data. t is a randomly chosen diffusion time step, and θ denotes the model parameters. This objective encourages the model to accurately predict the noise at any given time step, thereby learning the reverse of the diffusion process to reconstruct the original data from noised observations.

A.9 INTERPOLATION WITH STDIFFUSION

The stochastic encoding process incrementally introduces Gaussian noise into the original gene expression data x_0 across a series of discrete time steps. This stochastic encoding process transforms the data into a latent space where the original high-dimensional structure is preserved amidst the added noise. The process at a specific time step t can be mathematically represented as follows:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t} \cdot x_{t-1}, (1 - \bar{\alpha}_t) I\right) \tag{7}$$

where: $\bar{\alpha}_t$ is the cumulative product of $1 - \beta_t$ over time, indicating the proportion of the original signal preserved. I is the identity matrix, ensuring that noise is added isotropically.

A.10 SPATIAL PROXIMITY CALCULATION

The interpolation between two slices involves calculating the spatial proximity to determine the blending factor λ. This is achieved by normalizing the coordinates and calculating distances to the target coordinates to ensure comparability:

$$\text{coords}_{\text{normalized}} = \frac{\text{coords} - \min(\text{coords})}{\max(\text{coords}) - \min(\text{coords})} \tag{8}$$

A.11 DISTANCE CALCULATION

The Euclidean distance from each point in the slices to the target coordinates is calculated. The blending factor λ for each target point is then determined based on the proximity of points in slice

1 and slice 2 to the target coordinates. Our experimentation with λ values ranging from 0.1 to 0.9 has revealed its significant impact on interpolation quality, indicating that the optimized λ can vary across different datasets (Supp Fig 3, Fig 4, Fig 5).

A.12 Latent Space Blending for Interpolation

Given noised data $x_{(t,1)}$ from slice 1 and $x_{(t,2)}$ from slice 2 at time t, the interpolation for a target spatial coordinate is computed by blending these representations based on spatial proximity. The interpolated data $x_{(t,\text{interp})}$ at time t is calculated as:

$$x_{t,\text{interp}} = (1 - \lambda)\, x_{t,1} + \lambda\, x_{t,2}. \tag{9}$$

λ is adjusted based on the spatial proximity of the points in the slices to the target coordinate, influencing their contribution to the interpolated output.

A.13 Reverse Diffusion Process (Reconstruction)

The reverse diffusion process reconstructs the original data from the noised state to generate the interpolated gene expression data.

$$p(x_{t-1} \mid x_t) = \mathcal{N}\big(x_{t-1};\, \mu_\theta(x_t, t),\, \sigma_t^2(x_t, t)\, I\big) \tag{10}$$

where: $x_{(t-1)}$ and x_t represent the gene expression data at two consecutive time steps, with $(t-1)$ being closer to the original data. $\mu_\theta(x_t, t)$ and $\sigma_t^2(x_t, t)$ are the functions modeled by the neural network parameterized by θ. These functions predict the mean and variance of the Gaussian distribution from which $x_{(t-1)}$ is sampled, given x_t.

A.14 Loss function for interpolation

The loss function aims to minimize the difference between the actual noise added to the data during the forward process and the noise predicted by the model during the reverse process. For interpolation, we used mean squared error (MSE) between the actual noise ϵ and the estimated noise $\bar{\epsilon}_\theta$ given the noised data x_t and model parameters θ:

$$L(\theta) = \mathbb{E}_{x_0 \epsilon, t}\left[\|\epsilon - \hat{\epsilon}_\theta(x_t, t)\|^2\right] \tag{11}$$

where: ϵ is the actual noise vector sampled from a Gaussian distribution during the forward process, $\hat{\epsilon}_\theta(x_t, t)$ is the noise estimated by the model, given the noised data x_t at time t, and $\mathbb{E}$ denotes the expectation over the distribution of original data x_0, noise ϵ, and time steps t.

For interpolation, the focus is on blending noised representations from different slices based on spatial proximity and then reconstructing the interpolated data through the reverse process. Although the loss function remains centered on noise prediction accuracy, the application in interpolation emphasizes the model's capability to handle blended data from multiple sources and generate coherent interpolated outputs.

A.15 Linear Interpolation

The linear interpolation method calculates the gene expression profile for a target slice by averaging the expression profiles of two reference slices. For each spatial coordinate in the target slice S_2, the closest points in reference slices S_1 and S_3 are identified based on Euclidean distance. The gene expression profile for a point in target slice S_2 is computed as the average of the gene expression profiles of the closest point in slices S_1 and S_3.

A.16 Training Parameters

For tasks such as in silico generation and interpolation of spatial transcriptomics data, training parameters play an essential role in the performance of stDiffusion. The key parameters optimized

for stDiffusion for the task of in silico generation and interpolation across slices and layers are the noise schedule (β_t) - linearly spaced from 1×10^{-4} to 0.02. The noise schedule directly impacts the model's ability to learn the data distribution through diffusion, requiring tuning to match the complexity of spatial transcriptomics data. We used AdamW (Loshchilov & Hutter, 2019) as an optimizer and learning rate (lr $= 1 \times 10^{-3}$). In addition to this, stDiffusion is trained for 300 epochs with an early stopping criterion. To adapt the learning rate, using OneCycleLR promotes faster convergence and mitigates the risk of getting stuck in local minima. For interpolation tasks, most hyperparameters remain the same as those for in silico generation, reflecting the shared underlying model architecture and training strategy. However, specific considerations for interpolation include the interpolation factor (λ), which ranges from 0.1 to 1.0 in increments of 0.1 in experiments. This factor controls the blend between the gene expression profiles of two slices.

A.17 Evaluation Metrics: Neighborhood Enrichment Test

We used Squidpy's (Palla et al., 2022b) Neighborhood Enrichment Test. This test evaluates the spatial relationships between different clusters or types of cells within a tissue or dataset. It helps to understand how certain groups of cells are distributed to one another and whether there are significant patterns of co-localization or segregation. The test first identifies pairs of nodes (cells) belonging to specific classes or clusters (i and j). The next step is to count the sum of nodes that are proximal to each other, represented as $x_{i,j}$. This count reflects the direct interactions or the degree of proximity between cells of two different clusters. To determine whether the observed proximity of cluster pairs is significant, the test compares the cluster distance against what would be expected by chance. The randomized background is computed by scrambling the cluster labels while keeping the spatial connectivity unchanged, effectively randomizing the distribution of clusters. This process is repeated multiple times (default: 1,000 iterations) for a robust statistical comparison. From these iterations, the test calculates the expected means ($\mu_{i,j}$) and standard deviations ($\sigma_{i,j}$) for the proximity counts of each cluster pair in the randomized configurations. A z-score is then computed for each cluster pair using the formula:

$$z_{(i,j)} = \frac{x_{(i,j)} - \mu_{(i,j)}}{\sigma_{(i,j)}} \tag{12}$$

This z-score provides insights into the spatial organization of the tissue or dataset and helps to identify which cell types or clusters are more likely to be found near each other, potentially indicating functional relationships, shared microenvironments, or developmental pathways. To assess our model's performance, we applied the Leiden clustering method to group spots/cells into clusters and used these clusters as the basis for our neighborhood enrichment test. This approach allowed us to directly compare the spatial relationships observed in both the original (ground truth) data and the data generated by our model (both in-silico and interpolated slices), providing a clear metric for evaluating how well our model can replicate the complex spatial organization found in actual tissue samples. The second evaluation metric used is the Spearman correlation, which assesses the similarity between ground truth gene expression data and interpolated gene expression data across spatial spots in a dataset. This approach is advantageous in spatial transcriptomics to understand how well the interpolation process preserves the original data's spatial gene expression patterns. For each spot i, we calculated the ρ Spearman correlation coefficient between the corresponding rows (gene expression profiles) in $X_{\text{ground truth}}$ and $X_{\text{interpolated}}$ as:

$$\rho_i = \text{spearmanr}\left(X_{\text{ground truth}_i}, :, X_{\text{interpolated}_i}, :\right) \tag{13}$$

where: $X_{\text{ground truth}}$ and $X_{\text{interpolated}}$ denote the gene expression profiles for the spot i in the ground truth and interpolated data, respectively, and `spearmanr()` is the function that computes the Spearman correlation coefficient. After calculating the Spearman correlation for each spot and comparing the original and interpolated expression matrices, these correlations are normalized. This normalization step adjusts the correlation values so they fit within a specific range, making it easier to visualize differences.

A.18 EVALUATION METRICS: SPATIAL DISTRIBUTION OF NORMALIZED SPEARMAN CORRELATION

The second evaluation metric used is the Spearman correlation, which assesses the similarity between ground truth gene expression data and interpolated gene expression data across spatial spots in a dataset. This approach is advantageous in spatial transcriptomics to understand how well the interpolation process preserves the original data's spatial gene expression patterns. For each spot i, we calculated the ρ Spearman correlation coefficient between the corresponding rows (gene expression profiles) in the $X_{\text{ground truth}}$ and $X_{\text{interpolated}}$ as:

$$\rho_i = \text{spearmanr}\left(X_{\text{ground truth}_i}, :, X_{\text{interpolated}_i}, :\right) \tag{14}$$

where: $X_{\text{ground truth}}$ and $X_{\text{interpolated}}$ denote the gene expression profiles for the spot i in the ground truth and interpolated data, respectively, and `spearmanr()` is the function that computes the Spearman correlation coefficient. After calculating the Spearman correlation for each spot and comparing the original and interpolated expression matrices, these correlations are normalized. This normalization step adjusts the correlation values so they fit within a specific range, making it easier to visualize differences.

A.19 SUPPLEMENTARY FIGURES

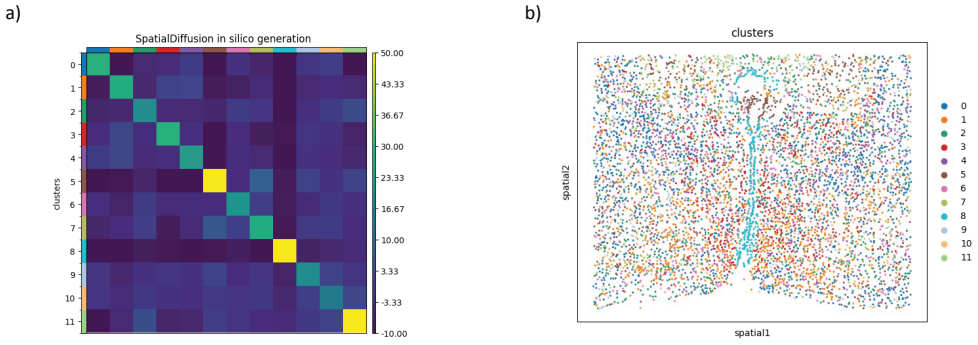

Figure S1: in silico generation of MERFISH data, a) showcases a heatmap of neighborhood enrichment, highlighting how stDiffusion effectively simulates spatial transcriptomics data in silico, ensuring the preservation of cluster neighborhoods as seen in actual data, b) presents a scatter plot of clusters identified through Leiden clustering and plotted in spatial coordinates, effectively reflecting the clustering and spatial distributions found in the actual data.

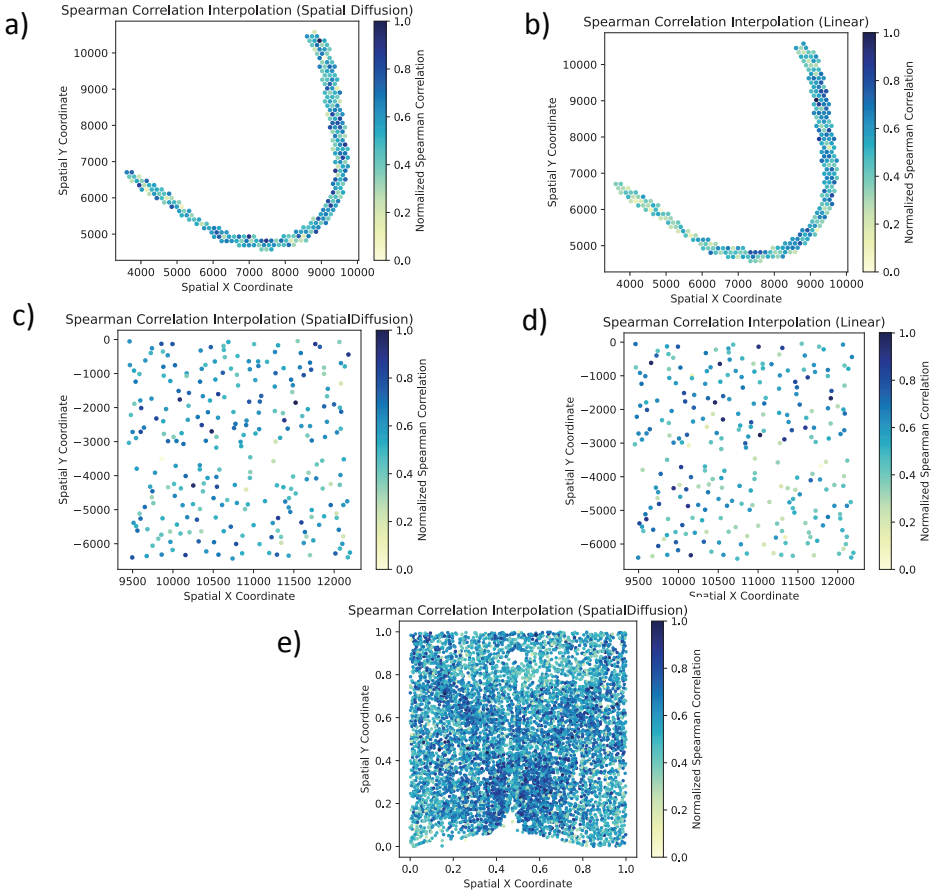

Figure S2: Spatial distribution of normalized Spearman correlation for DLPFC and Starmap data a, b) spatial distribution of normalized Spearman correlation between the ground truth Layer 4 from a DLPFC slice and interpolated Layer 4 with stDiffusion and linear interpolation, respectively, c, d) spatial distribution of normalized spearman correlation between the ground truth layer 2/3 from a Starmap slice and interpolated layer 2/3 with stDiffusion and linear interpolation. Results show that stDiffusion captures the spatial structure as it is distributed more evenly than linear interpolation, e) spatial distribution of normalized spearman correlation between the ground truth slice 9 from a mouse MERFISH data and interpolated slice 9 with stDiffusion, showing its efficacy in maintaining the spatial distribution of gene expression levels in the interpolated slice, consistent with the original data.

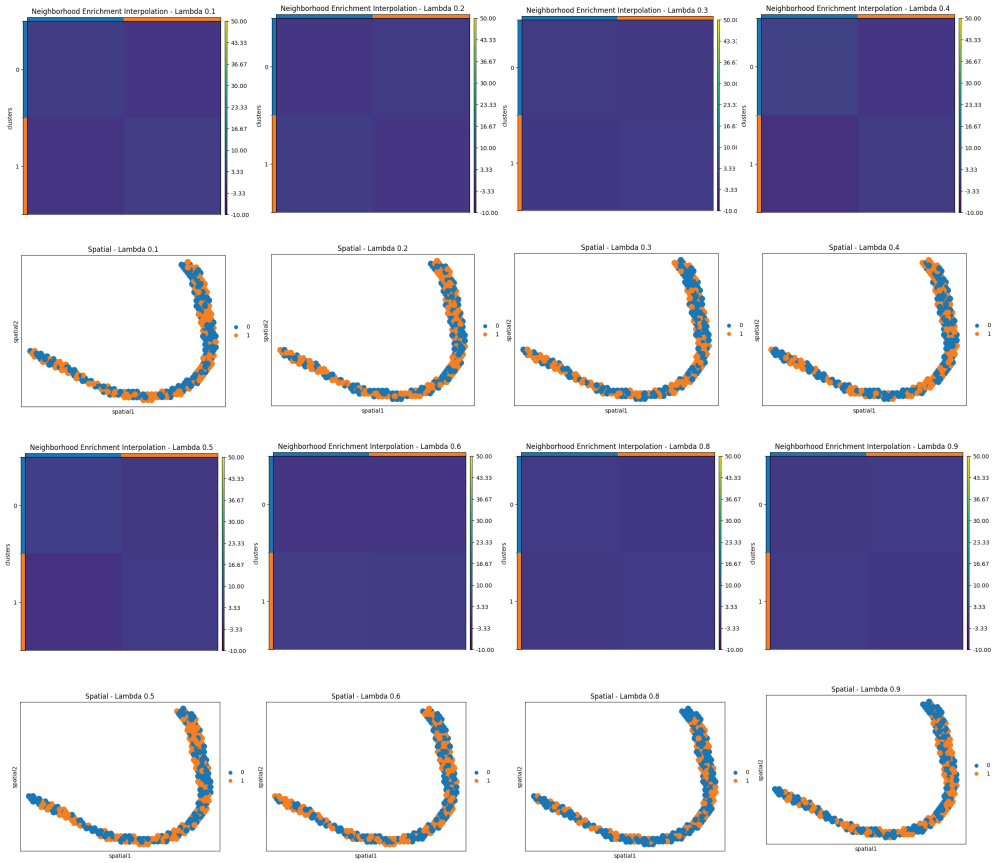

Figure S3: Effect of lambda on interpolation within and across slices: Figures S3 (DLPFC), S4 (Starmap), and S5 (MERFISH) illustrate the interpolation results for varying lambda (λ) values from 0.1 to 0.9, demonstrating how different blending ratios between reference slices and layers within a slice affect the quality and accuracy of the interpolated gene expression profiles. Each subplot represents the interpolated data generated using a specific λ value, highlighting the gradual transition in gene expression patterns as the blending ratio changes. The comparison underscores the significance of selecting an optimal λ value to achieve the most biologically plausible interpolation between slices.