Predicting Gene Expression in Spatially Resolved Transcriptomics Across Samples Through Probabilistic Fusion of Hierarchical Histology and Spatial Information

Supplementary Material

7. Mathematical Derivations

In this section, we provide detailed mathematical derivations for the key theoretical components of the STevs framework.

7.1. Theoretical Justification for PoE Fusion

Here, we prove that the PoE fusion mechanism yields a posterior distribution with higher certainty (i.e., lower variance) than any of the individual expert distributions [27]. This provides a strong theoretical motivation for its use over simpler fusion methods like feature concatenation or averaging.

Theorem 1. *Given two independent Gaussian distributed experts, the variance of the fused distribution obtained via PoE is less than or equal to the variance of each individual expert.*

Proof. Let the latent distributions from the image encoder and the spatial encoder be two independent Gaussian "experts":

$$p(z|I) = \mathcal{N}(z|\mu_{\text{img}}, \sigma^2_{\text{img}}) \tag{4}$$

$$p(z|s) = \mathcal{N}(z|\mu_{\text{spatial}}, \sigma^2_{\text{spatial}}) \tag{5}$$

The Product of Experts framework defines the fused distribution $p_{\text{PoE}}(z)$ by multiplying the probability density functions of the individual experts:

$$p_{\text{PoE}}(z) \propto p(z|I) \cdot p(z|s) \tag{6}$$

The product of two Gaussian distributions is an unnormalized Gaussian. By completing the square, we find that the resulting fused distribution $p_{\text{PoE}}(z) = \mathcal{N}(z|\mu_{\text{fused}}, \sigma^2_{\text{fused}})$ has parameters defined by its precision (inverse variance). The precision of the fused distribution is the sum of the individual precisions:

$$\frac{1}{\sigma^2_{\text{fused}}} = \frac{1}{\sigma^2_{\text{img}}} + \frac{1}{\sigma^2_{\text{spatial}}} \tag{7}$$

From Equation 7, since variances are strictly positive ($\sigma^2 > 0$), it directly follows that:

$$\frac{1}{\sigma^2_{\text{fused}}} \geq \frac{1}{\sigma^2_{\text{img}}} \implies \sigma^2_{\text{fused}} \leq \sigma^2_{\text{img}} \tag{8}$$

$$\frac{1}{\sigma^2_{\text{fused}}} \geq \frac{1}{\sigma^2_{\text{spatial}}} \implies \sigma^2_{\text{fused}} \leq \sigma^2_{\text{spatial}} \tag{9}$$

This proves that the fused posterior distribution is always more certain (has a smaller variance) than any of the individual expert distributions. This property makes PoE a robust mechanism for integrating multimodal information, as it naturally produces a more confident estimate by combining evidence from different sources.

7.2. Information-Theoretic Perspective on Latent Space Alignment

Here, we provide a theoretical justification for our latent space alignment loss, $\mathcal{L}_{\text{align}}$, from an information-theoretic perspective. We argue that minimizing the MSE between **stochastic samples** from the unimodal latent distributions serves as a practical and powerful method for aligning these distributions and maximizing their shared information.

Let $q(z_{\text{img}})$ and $q(z_{\text{spatial}})$ be the latent distributions produced by the image and spatial encoders, respectively, where $q(z_{\text{img}}) = \mathcal{N}(\mu_{\text{img}}, \Sigma_{\text{img}})$ and $q(z_{\text{spatial}}) = \mathcal{N}(\mu_{\text{spatial}}, \Sigma_{\text{spatial}})$. A principled way to enforce consistency between these two distributions is to minimize their KL divergence, $D_{KL}(q(z_{\text{img}})||q(z_{\text{spatial}}))$.

For two multivariate Gaussian distributions with diagonal covariance matrices, the KL divergence has a closed-form solution [19]:

$$\begin{aligned} D_{KL}(q_{\text{img}}\|q_{\text{spat}}) = \frac{1}{2}\Bigg[& \log \frac{|\Sigma_{\text{spat}}|}{|\Sigma_{\text{img}}|} - d \\ & + \text{tr}\left(\Sigma^{-1}_{\text{spat}}\Sigma_{\text{img}}\right) \\ & + (\mu_{\text{spat}} - \mu_{\text{img}})^T \Sigma^{-1}_{\text{spat}} \\ & (\mu_{\text{spat}} - \mu_{\text{img}}) \Bigg] \end{aligned} \tag{10}$$

where d is the dimensionality of the latent space.

While directly minimizing Equation 10 is a valid approach, it can introduce training instability due to the log-determinant and matrix inversion terms. We therefore adopt a more direct, sampling-based strategy. Our proposed alignment loss, $\mathcal{L}_{\text{align}}$, minimizes the squared Euclidean distance between latent variables z_{img} and z_{spatial} that are sampled from their respective distributions using the reparameterization trick:

$$\mathcal{L}_{\text{align}} = \mathbb{E}_{z_{\text{img}} \sim q_{\text{img}}, z_{\text{spatial}} \sim q_{\text{spatial}}}[||z_{\text{img}} - z_{\text{spatial}}||_2^2] \tag{11}$$

In practice, this expectation is approximated with a single sample per training instance.

This objective provides a powerful implicit regularization. By minimizing the distance between the **samples** (z), we are not only aligning the **means** (μ) but also encouraging consistency in the **variances** (σ^2). If the variances were

significantly different, the expected distance between samples would remain large even if the means were perfectly aligned. Therefore, this loss term forces the **entire distributions to overlap**, not just their central points. This approach is computationally efficient, stable to train, and has been empirically shown to be highly effective. It directly enforces that the informational content from both modalities maps to a coherent and shared region in the latent space, which is a crucial step towards learning a slide-invariant representation.

7.3. Derivation of the Evidence Lower Bound (ELBO) for STevs

Here, we show that the composite loss function used to train STevs is a principled objective derived from maximizing the ELBO on the marginal log-likelihood of the observed data [64].

Let our observed data be $X = \{I, R\}$, representing the histology image and the corresponding RNA expression profile. Our goal is to maximize the marginal log-likelihood $\log p(X)$. We introduce a latent variable z that captures the underlying biological state from which the observations are generated. The marginal log-likelihood is given by:

$$\log p(X) = \log \int p(X, z) dz \tag{12}$$

Directly optimizing this integral is intractable. Therefore, we introduce an approximate posterior distribution $q_\phi(z|I, s)$, parameterized by an encoder with parameters ϕ, which takes both the image I and spatial coordinates s as input to approximate the true posterior $p(z|X)$. In STevs, $q_\phi(z|I, s)$ is the PoE-fused distribution.

We can rewrite the marginal log-likelihood as:

$$\log p(X) = \log \int p(X, z) \frac{q_\phi(z|I, s)}{q_\phi(z|I, s)} dz \tag{13}$$

$$= \log \mathbb{E}_{q_\phi(z|I,s)} \left[\frac{p(X, z)}{q_\phi(z|I, s)} \right] \tag{14}$$

By applying Jensen's inequality, we obtain the ELBO, denoted as $\mathcal{L}$:

$$\log p(X) \geq \mathbb{E}_{q_\phi(z|I,s)} \left[\log \frac{p(X, z)}{q_\phi(z|I, s)} \right] \tag{15}$$

$$\mathcal{L}(\phi, \theta; X, s) = \mathbb{E}_{q_\phi(z|I,s)} \Big[\log p_\theta(X|z) + \log p(z) - \log q_\phi(z|I, s) \Big] \tag{16}$$

$$= \mathbb{E}_{q_\phi(z|I,s)} [\log p_\theta(X|z)] - D_{\text{KL}}\big(q_\phi(z|I, s) \| p(z)\big) \tag{17}$$

where $p_\theta(X|z)$ is the decoder parameterized by θ, and $p(z)$ is the prior distribution over the latent space, which we set to a standard normal distribution $\mathcal{N}(0, I)$.

Assuming conditional independence between the image and RNA data given the latent variable z, the reconstruction term can be decomposed:

$$p_\theta(X|z) = p_{\theta_I}(I|z) \cdot p_{\theta_R}(R|z) \tag{18}$$

Substituting this back into Equation 17, we get:

$$\begin{aligned} \mathcal{L} = & \underbrace{\mathbb{E}_{q_\phi}[\log p_{\theta_I}(I|z)]}_{\text{Image Recon.}} \\ & + \underbrace{\mathbb{E}_{q_\phi}[\log p_{\theta_R}(R|z)]}_{\text{Gene Recon.}} \\ & - \underbrace{D_{\text{KL}}(q_\phi(z|I, s) \| p(z))}_{\text{KL Regularization}} \end{aligned} \tag{19}$$

Maximizing this ELBO is equivalent to minimizing its negative. Each term corresponds to a component of our loss function:

- $-\mathbb{E}_{q_\phi}[\log p_{\theta_I}(I|z)]$ corresponds to the image reconstruction loss L_{img}. It is implemented as an MSE loss under a Gaussian likelihood assumption.
- $-\mathbb{E}_{q_\phi}[\log p_{\theta_R}(R|z)]$ corresponds to the gene expression reconstruction loss L_{rna}. It uses the Negative Log-Likelihood (NLL) of the Negative Binomial distribution.
- $D_{\text{KL}}(q_\phi(z|I, s) \| p(z))$ is the KL divergence loss L_{KLD}.

Finally, we introduce the latent space alignment loss L_{align} as an additional regularization term to enforce consistency between the modality-specific encoders. This term is not derived from the ELBO itself but is added to the objective to improve generalization, a common practice in representation learning. Thus, our final objective is to minimize the total loss L_{total}, which is equivalent to maximizing a regularized ELBO:

$$\min_{\phi,\theta} L_{\text{total}} \quad \Longleftrightarrow \quad \max_{\phi,\theta} (\mathcal{L} - \gamma L_{\text{align}}) \tag{20}$$

8. STevs Model Architecture

This section provides the comprehensive architectural details of the proposed STevs model. Table 6 outlines the specific layer configurations, parameter specifications, and output dimensions for each core component, including the hierarchical Image Encoder, the coordinate-based Spatial Encoder, the Product of Experts (PoE) Fusion module, and the multi-task decoders. The source code and detailed implementation are publicly available at https://github.com/yinboliu-git/STevs.

9. Detailed Explanation of the STevs Model Architecture

9.1. Model Overview

STevs is a deep generative model based on a MM-VAE[60], designed to robustly predict spatially resolved transcrip-

Table 6. Detailed architecture of the STevs model. The table outlines the layers, specifications, and output dimensions for each component of the network, from the parallel encoders to the multi-task decoders.

Component	Module / Layer	Specification	Output Dimension
Image Encoder	Input Image Patches	3-channel RGB image	(3, 160, 160)
	Patch Embedding (Conv2d)	kernel=(4,4), stride=(4,4)	(96, 40, 40)
	Swin Stage 1 (2 blocks)	Window Attention, MLP	(96, 40, 40)
	Patch Merging + Swin Stage 2 (2 blocks)	Downsamples feature map	(192, 20, 20)
	Patch Merging + Swin Stage 3 (6 blocks)	Downsamples feature map	(384, 10, 10)
	Patch Merging + Swin Stage 4 (2 blocks)	Downsamples feature map	(768, 5, 5)
	Global Average Pooling	-	768
	Latent Head ($\mu_{\text{img}}, \log \sigma^2_{\text{img}}$)	Two Linear Layers	2×128
Spatial Encoder	Input Coordinates	Normalized (x, y) coordinates	2
	MLP (2 hidden layers)	Linear(2, 64) $\rightarrow$ ReLU $\rightarrow$ Linear(64, 128) $\rightarrow$ ReLU	128
	Latent Head ($\mu_{\text{spatial}}, \log \sigma^2_{\text{spatial}}$)	Two Linear Layers	2×128
Fusion	PoE	Fuses image and spatial latent distributions	Fused Latent ($z \in \mathbb{R}^{128}$)
Image Decoder	Input Linear Layer	Projects z and reshapes	(256, 20, 20)
	Upsampling Stage 1 (ConvT + ConvBlock)	ConvTranspose2d(256, 128), stride=2	(128, 40, 40)
	Upsampling Stage 2 (ConvT + ConvBlock)	ConvTranspose2d(128, 64), stride=2	(64, 80, 80)
	Upsampling Stage 3 (ConvT + ConvBlock)	ConvTranspose2d(64, 32), stride=2	(32, 160, 160)
	Final Layer (Conv2d + Tanh)	kernel=(3,3), padding=1	(3, 160, 160)
Gene Decoder	Base MLP (2 hidden layers)	Linear(128, 256) $\rightarrow$ BN/ReLU/Dropout $\rightarrow$ Linear(256, 512) BN/ReLU/Dropout	512
	Mean (μ) Head	Linear(512, 2350) $\rightarrow$ Softplus	2350
	Dispersion (θ) Head	Linear(512, 2350) $\rightarrow$ Softplus	2350

tomics by fusing visual histological information from histological images with their spatial positional context. The core architecture of the model consists of three main components:

- Parallel dual-path encoders that process images and spatial coordinates, respectively. The image encoder adopts a hierarchical vision Transformer [33], while the spatial encoder uses a MLP [31].
- A probabilistic fusion module based on the PoE [21], used to integrate the latent distributions generated by the dual-path encoders.
- A multi-task decoder that is simultaneously responsible for image reconstruction and the generation of gene expression profiles based on the Negative Binomial Distribution [35].

9.2. Detailed Network Structure

9.2.1. Image histology Encoder

This encoder is responsible for extracting rich, hierarchical visual features from the input histological image patches.

- **Input**: An image patch $I_i \in \mathbb{R}^{N \times H \times W \times C}$, where H and W are the height and width of a single spot's image patch, C is the number of channels (for RGB images, $C = 3$), and N is the total number of neighborhood patches.
- **Backbone Network**: We use a Swin Transformer [33] as the feature extraction backbone. This network, through its hierarchical structure and shifted window self-attention mechanism [33], can effectively capture long-range dependencies within the image, which is crucial for understanding complex tissue structures [15].
- **Feature Extraction**: The Swin Transformer maps the input image I_i to a fixed-dimensional feature vector $f_{\text{img}} \in \mathbb{R}^{D_{\text{feat}}}$.

$$f_{\text{img}} = \text{SwinTransformer}(I_i) \tag{21}$$

- **Latent Space Mapping**: This feature vector f_{img} is then passed through two independent fully connected (FC) layers [35] to generate the mean vector $\mu_{\text{img}} \in \mathbb{R}^{D_{\text{latent}}}$ and the log-variance vector $\log \sigma^2_{\text{img}} \in \mathbb{R}^{D_{\text{latent}}}$ of the image modality latent space, respectively.

$$\mu_{\text{img}} = \text{FC}_{\mu,\text{img}}(f_{\text{img}}) \tag{22}$$

$$\log \sigma^2_{\text{img}} = \text{FC}_{\sigma,\text{img}}(f_{\text{img}}) \tag{23}$$

To leverage the prior knowledge from large-scale natural image datasets, our Swin Transformer backbone loads weights pre-trained on the ImageNet dataset [14].

9.2.2. Spatial Context Encoder

This encoder is used to capture the global positional information of each image patch within the tissue slice.

- **Input**: A two-dimensional spatial coordinate vector $s_i = (x_i, y_i) \in \mathbb{R}^2$, representing the center coordinates of the image patch I_i.
- **Network Structure**: We use a Multilayer Perceptron [3] with two hidden layers to perform a non-linear transformation on the coordinate information. The activation function in the network is the Rectified Linear Unit (ReLU).

- **Feature Extraction**: The MLP maps the input coordinates s_i to a high-dimensional feature vector $f_{\text{spatial}} \in \mathbb{R}^{D_{\text{hidden}}}$.

$$f_{\text{spatial}} = \text{MLP}_{\text{spatial}}(s_i) \tag{24}$$

- **Latent Space Mapping**: Similar to the image encoder, f_{spatial} is passed through two independent fully connected layers to generate the mean vector $\mu_{\text{spatial}} \in \mathbb{R}^{D_{\text{latent}}}$ and the log-variance vector $\log \sigma^2_{\text{spatial}} \in \mathbb{R}^{D_{\text{latent}}}$ of the spatial modality latent space.

$$\mu_{\text{spatial}} = \text{FC}_{\mu,\text{sp}}(f_{\text{spatial}}) \tag{25}$$

$$\log \sigma^2_{\text{spatial}} = \text{FC}_{\sigma,\text{sp}}(f_{\text{spatial}}) \tag{26}$$

9.2.3. Multimodal Fusion and Latent Space Sampling

- **PoE Fusion** To integrate information from the two modalities and their respective uncertainties, we adopt the PoE framework [21]. We sum the precisions (the inverse of the variance) of the latent distributions output by the two encoders (both are Gaussian distributions) to calculate the precision of the fused distribution, and then derive the fused mean and variance.
  - **Precision Calculation**:

$$T_{\text{img}} = (\sigma^2_{\text{img}})^{-1} = \exp(-\log \sigma^2_{\text{img}}) \tag{27}$$

$$T_{\text{spatial}} = (\sigma^2_{\text{spatial}})^{-1} = \exp(-\log \sigma^2_{\text{spatial}}) \tag{28}$$

  - **Fused Distribution Parameters**:

$$\sigma^2_{\text{fused}} = (T_{\text{img}} + T_{\text{spatial}})^{-1} \tag{29}$$

$$\mu_{\text{fused}} = (\mu_{\text{img}} T_{\text{img}} + \mu_{\text{spatial}} T_{\text{spatial}}) \sigma^2_{\text{fused}} \tag{30}$$

- **Reparameterization Sampling** To enable backpropagation of gradients through the sampling process, we use the reparameterization trick [29]. We sample a random noise vector $\epsilon \sim \mathcal{N}(0, I)$ from a standard normal distribution and then generate the final latent vector $z \in \mathbb{R}^{D_{\text{latent}}}$.

$$z = \mu_{\text{fused}} + \sigma_{\text{fused}} \odot \epsilon \tag{31}$$

where $\odot$ denotes element-wise multiplication.

9.3. Multi-task Decoder

9.3.1. Image Reconstruction Decoder

This decoder reconstructs the original histological image from the latent vector z.

- **Initial Transformation**: First, a fully connected layer projects z into a high-dimensional space and reshapes it into a small 3D feature map $h_{\text{img}} \in \mathbb{R}^{C' \times H' \times W'}$ to serve as the starting point for subsequent convolutional operations.
- **Upsampling**: Next, a series of transposed convolution modules (including ConvTranspose2d [50], BatchNorm2d [52], and LeakyReLU[37]) are used to progressively increase the spatial dimensions of the feature map while reducing its number of channels.
- **Final Output**: The final layer is a 3×3 convolutional layer that maps the feature map's channel count back to the number of channels of the input image, C. A Tanh activation function [9] is then used to normalize the output pixel values to the range $[-1, 1]$, yielding the reconstructed image I_{recon}.

9.3.2. Gene Expression Decoder

This decoder generates the distribution parameters for the gene expression profile from the latent vector z.

- **Feature Transformation**: The latent vector z is first passed through an MLP network, which includes Batch Normalization (BatchNorm1d) [25], ReLU [1], and Dropout [57], to extract high-level features h_{rna} for gene expression prediction.
- **Parameter Prediction**: The feature vector h_{rna} is then fed into two independent linear output layers, which are used to predict the mean parameter $\mu_{\text{rna}} \in \mathbb{R}^M$ and the dispersion parameter $\theta_{\text{rna}} \in \mathbb{R}^M$ of the NB distribution, where M is the number of target genes.

$$\mu_{\text{rna}} = \text{Softplus}(\text{Linear}_{\mu,\text{rna}}(h_{\text{rna}})) \tag{32}$$

$$\theta_{\text{rna}} = \text{Softplus}(\text{Linear}_{\theta,\text{rna}}(h_{\text{rna}})) \tag{33}$$

We use the Softplus activation function [5] to ensure that the values of μ_{rna} and θ_{rna} are positive, which is consistent with the parameter definition of the Negative Binomial distribution.

9.4. Loss Function and Optimization

9.4.1. Composite Loss Function

The training objective of STevs is optimized through a carefully designed composite loss function L_{total}, which consists of four components:

$$L_{\text{total}} = \lambda_{\text{img}} L_{\text{img}} + \lambda_{\text{rna}} L_{\text{rna}} + \beta L_{\text{KLD}} + \gamma L_{\text{align}} \tag{34}$$

- **Image Reconstruction Loss (L_{img})** We use the MSE to measure the difference between the reconstructed image I_{recon} and the original image I_{true}:

$$L_{\text{img}} = \frac{1}{N} \sum_{i=1}^{N} \|I_{\text{recon}}(i) - I_{\text{true}}(i)\|_2^2 \tag{35}$$

- **Gene Expression Reconstruction Loss (L_{rna})** We use the Negative Log-Likelihood (NLL) [35]of the Negative Binomial distribution. The probability mass function (PMF) of the Negative Binomial distribution is defined as:

$$P(Y = k|\mu, \theta) = \frac{\Gamma(k+\theta)}{\Gamma(k+1)\Gamma(\theta)} \left(\frac{\theta}{\theta+\mu}\right)^{\theta} \left(\frac{\mu}{\theta+\mu}\right)^{k} \tag{36}$$

where $\Gamma(\cdot)$ is the Gamma function.

- **KL Divergence Loss (L_{KLD})** As a standard component of the VAE framework [29], we use the KL divergence to regularize the fused latent space. For a single sample, its analytic form is [11]:

$$L_{\text{KLD}} = \frac{1}{2} \sum_{j=1}^{D_{\text{latent}}} (\sigma_{\text{fused},j}^2 + \mu_{\text{fused},j}^2 - 1 - \log \sigma_{\text{fused},j}^2) \quad (37)$$

- **Latent Space Alignment Loss (L_{align})** We introduce an additional MSE loss to encourage different modality encoders to learn semantically aligned representations [50, 68]. This loss directly pulls the mean vectors of the two modalities closer in the latent space.

9.4.2. Optimization Strategy

- **KL Annealing** To stabilize the training of the VAE, we adopt a KL annealing strategy [6]. The weight β in the loss function is dynamically adjusted during training, its value increasing linearly from 0 to a preset maximum value β_{max} with training epoch e, and then remaining constant.

$$\beta_e = \beta_{\text{max}} \cdot \min\left(1.0, \frac{e}{E_{\text{anneal}}}\right) \quad (38)$$

  where E_{anneal} is the total number of annealing epochs.
- **Optimizer** We use the AdamW optimizer [36] to update all learnable parameters of the model. Compared to the traditional Adam, AdamW typically provides better generalization performance by decoupling weight decay from the gradient update.

10. Data Processing and Experimental Design

10.1. Datasets and Preprocessing

10.1.1. Data Sources

Our study utilized a total of five dataset groups from two sources. The first three groups are from the public human dorsolateral prefrontal cortex (DLPFC) dataset [39], comprising 12 tissue sections from 3 different individuals. The latter two groups are public Visium mouse brain datasets from the 10x Genomics platform [58], containing 4 tissue sections from the same mice but different region. An overview of these datasets is provided in Figure 6 and Table 7. Additionally, to further validate the model's generalization ability, we conducted extended experiments on the Human Breast Cancer (HBC) [69] and HSC [27] datasets. These datasets is provided in Figure 22.

10.1.2. Image Preprocessing

For each spot, we extracted image patches from the corresponding high-resolution H&E stained whole-slide image (Figure 7). Specifically, we defined a perceptual field of a 5×5 grid of base patches centered on each spot's coordinates. With each base patch having a resolution of 32×32 pixels, this resulted in a final input image tensor of $160 \times 160 \times 3$ for the model, capturing both the fine-grained histology of the target spot and its adjacent microenvironment. Prior to being fed into the model, all image patches were normalized to the range $[-1, 1]$ using min-max normalization to stabilize training:

$$I_{\text{norm}} = \frac{2(I - I_{\text{min}})}{I_{\text{max}} - I_{\text{min}}} - 1 \quad (39)$$

where I is the original pixel value, and I_{min} and I_{max} are the minimum and maximum pixel values within the patch, respectively. The normalized patches also serve as the ground truth target for the image reconstruction task.

10.1.3. Spatial Coordinate Preprocessing

For each image patch, we extract the corresponding 2D coordinate vector (x_i, y_i) from the tissue position file. These coordinates, representing the relative center position of the spot within the whole-slide image, are normalized and directly fed into the spatial encoder to preserve the global positional context of each patch. Generally, for relatively well-aligned slices, no operation on the spatial coordinates is necessary. However, for slices with discrepancies such as rotation or displacement, we recommend flattening the image patch corresponding to the coordinates to serve as features, and then using `Spateo` for coordinate alignment [49].

10.1.4. Gene Expression Preprocessing

Given the high dimensionality and sparsity of SRT data, we performed a gene filtering step to identify SVGs. Using `scanpy` [67] and `squidpy` [46], we calculated spatial autocorrelation (Moran's I) for each gene and retained those with a p-value less than 0.05. To ensure robustness, a gene was only included in the final set if it was identified as an SVG in at least two samples within the same dataset group. After filtering, each group contained over 2,000 SVGs. These filtered gene expression profiles serve as the ground truth target for the gene expression decoder, which models them using a Negative Binomial distribution to account for their count-based and over-dispersed nature.

10.2. Experimental Design and Evaluation

10.2.1. Experimental Settings

We evaluated model performance under two distinct settings:

- **Intra-slice Prediction:** For each slice, we randomly and independently split the spots into training, validation, and test sets with a 7:1:2 ratio to perform standard cross-validation within a single tissue slice.
- **Cross-slice Prediction:** To assess generalization, we employed a more challenging leave-one-out approach within each dataset group. One slice was used for training and

all other slices in the group were used for testing (as illustrated in Fig. 2c). This setup mimics the real-world scenario of applying a pre-trained model to new, unseen patient samples.

10.2.2. Evaluation Metrics

To quantitatively evaluate the model's prediction accuracy for gene expression, we employed three standard statistical metrics [3, 4]:

- **Mean Squared Error (MSE)** [29]: To measure the average squared difference between predicted and actual gene expression values.
- **Pearson Correlation Coefficient (PCC)** [48]: To assess the linear relationship between predicted and ground-truth expression profiles.
- **Spearman's Rank Correlation Coefficient (SCC)** [56]: To evaluate the monotonic relationship, which is robust to outliers.

10.2.3. Experimental Setup

Hardware Configuration Our experiments were conducted on a high-performance server equipped with four NVIDIA A100 GPUs (80GB of VRAM each), dual Intel(R) Xeon(R) Gold 6267C CPUs, and 1.5TB of system memory. The runtimes reported in the main paper are for model inference on a single GPU.

Training Parameters We trained all models for 100 epochs using a learning rate of 1×10^{-4} on four NVIDIA A100 (80GB) GPUs. A comprehensive list of all training hyperparameters is provided in Appendix 17.2.

Loss Function Weights In our experiments, the default weights for the composite loss function were set to $\lambda_{\text{img}} = 1.0$, $\lambda_{\text{rna}} = 10.0$, $\beta = 0.5$, and $\gamma = 0.5$. A detailed sensitivity analysis of these weights on model performance is discussed in Appendix 15.

10.2.4. Baseline set

For the baseline models, we strictly adhered to their officially provided pipelines for training and evaluation, making minor adaptive modifications to some for compatibility. Notably, since iStar's methodology involves predicting all spots at once, we employed a masking strategy during the intra-slice training phase to conceal the test set samples and prevent data leakage. For the STAGE model, we deviated from its provided pipeline, which calculates metrics on the combined training and test sets. To maintain consistency and ensure a fair comparison with all other methods, we adopted a stricter approach, evaluating its performance exclusively on the test set.

11. MAIN PERFORMANCE Details

11.1. DLPFC on Intra-Slice Experiments

As shown in Table 9, in the intra-slice experiments conducted on the DLPFC datasets, we systematically evaluated the performance of our STevs model against various existing mainstream methods. This evaluation was performed on three public datasets: Human 1, Human 2, and Human 3. The evaluation metrics include MSE, PCC, and SCC. The experimental results clearly indicate that our STevs model achieves optimal performance across all three datasets and on all three evaluation metrics. This data provides strong evidence for the superiority and robustness of the STevs model in the task of intra-slice spatial gene expression prediction.

11.2. DLPFC on Cross-Slice Experiments

To further evaluate the model's generalization ability, this subsection presents the results from the more challenging cross-slice experiments. In this setting, the model must utilize information from the training slices to predict the gene expression profile of a completely unseen slice from the same tissue, posing a stringent test of its knowledge transfer capabilities. As detailed in Table 9, the performance comparison on the DLPFC datasets (Human 1, 2, and 3) shows that most baseline models suffered a significant performance drop due to their inability to generalize effectively. The predictions of some models were even indistinguishable from random noise (with PCC/SCC values close to zero or negative). However, the STevs model still performed exceptionally well in this rigorous test, with its performance significantly surpassing all competing methods across all metrics. Specifically, compared to the best-performing baseline model, iStar, STevs demonstrated a substantial advantage in correlation metrics (PCC and SCC). For example, on the Human 3 dataset, the PCC of STevs reached 0.256, far exceeding iStar's 0.122. This result provides strong evidence for the powerful cross-slice generalization ability of the STevs model, showcasing its capacity to effectively transfer spatial pattern knowledge learned from training slices to new, unseen target slices.

11.3. 10x Mouse Brain on Intra-slice Experiments

To further validate our model's broad applicability and cross-species generalization ability, we also conducted a series of rigorous intra-slice performance evaluations on the 10x Mouse Brain dataset. As shown in Table 10, we performed a comprehensive performance comparison between STevs and various mainstream baseline models on two different brain region slices: Sagittal-Anterior and Sagittal-Posterior. The experimental results clearly show that the STevs model consistently outperformed all competing methods on both mouse brain datasets. This suc-

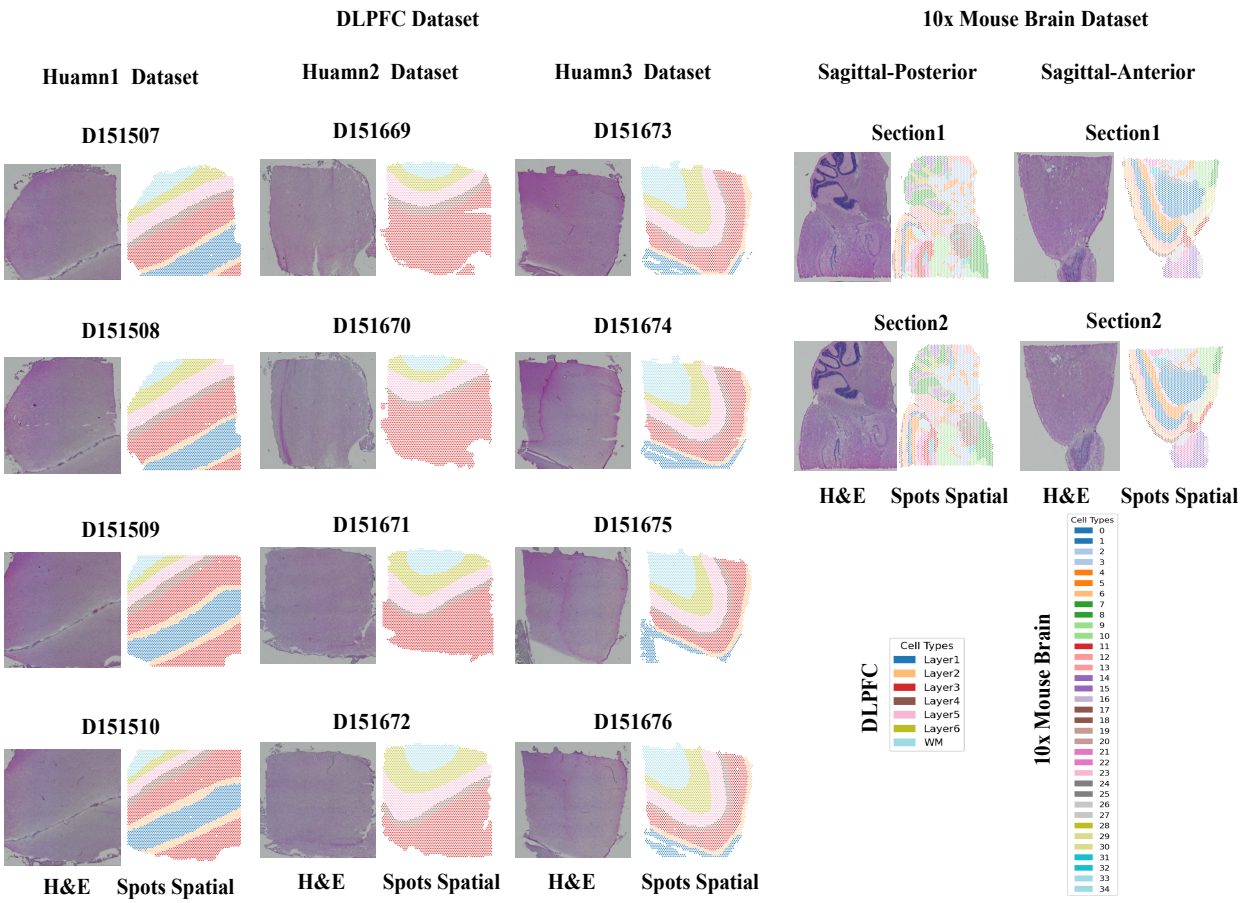

Figure 6. Overview of the primary spatial transcriptomics datasets used in this study. The figure shows the H&E images and their corresponding annotated cell type/tissue layer maps for a total of 16 datasets from five groups.

cessful validation on datasets from a different species provides strong evidence for the STevs model's powerful generalization ability and its great potential for application as a general-purpose framework in diverse biological scenarios.

11.4. 10x Mouse Brain on Cross-Slice Experiments

In this section, we subject our model to the most rigorous test: a cross-slice generalization performance evaluation on the 10x Mouse Brain dataset. This task requires the model to transfer and apply knowledge learned from one brain region slice to another, completely unseen one, posing the ultimate challenge to its generalization and robustness. The experimental results in Table 11 once again unequivocally demonstrate the superior performance of STevs. In this highly challenging scenario, STevs not only far surpassed most baseline models, whose predictions were close to random, but also achieved a comprehensive and significant victory over the strongest competitor, iStar. This is especially evident in the prediction for the posterior (Sagittal-Posterior) slice, where STevs's PCC reached as high as 0.442, a substantial lead compared to iStar's 0.363, fully reflecting its powerful predictive capability and generalization stability. Synthesizing the performance on both human and mouse datasets, these cross-slice experimental results ultimately establish the status of STevs as a high-performance, cross-species applicable, and general-purpose framework for spatial gene expression prediction.

12. Ablation Study Details

To systematically validate the necessity of each core component within the STevs model and to demonstrate the superiority of our design choices, we conducted a series of comprehensive ablation studies. We evaluated the impact on performance by removing or replacing the model's key modules, with the results summarized in Table 12 and 14 (intra-slice) and Table 13 and 15 (cross-slice).

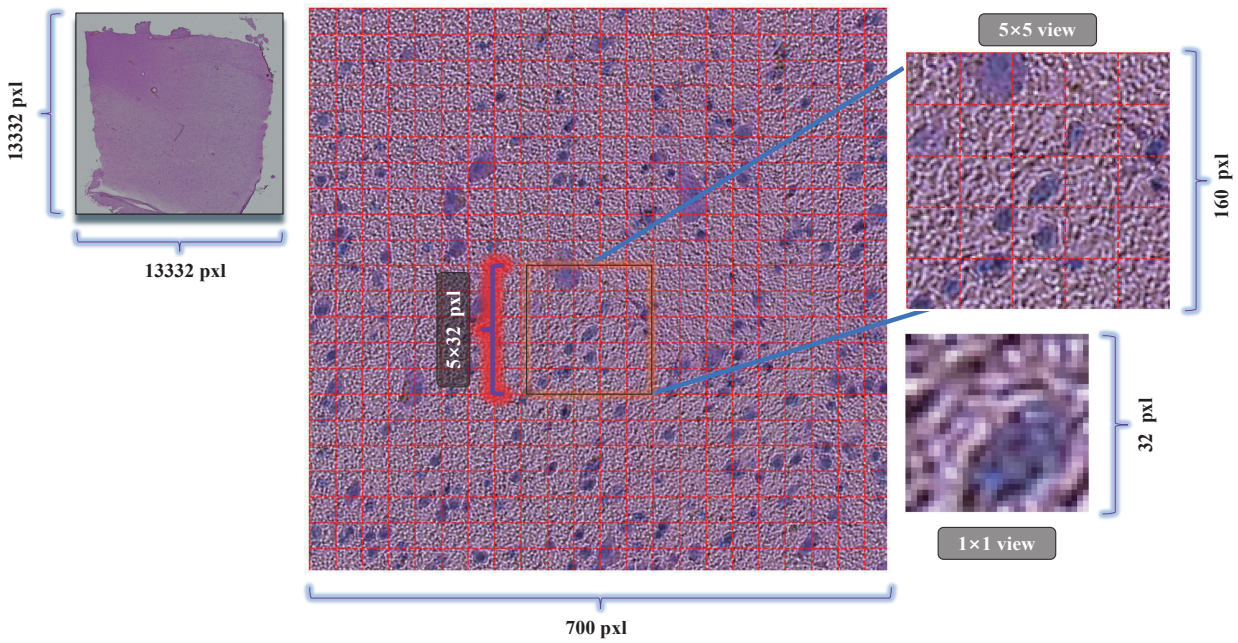

Figure 7. Schematic diagram of the image patch extraction process from a whole-slide H&E image

Table 7. Spots and Gene Counts per Sample Group

Group	Sample	Spots Number	Gene Number
Human 1	DLPFC 151507	4221	2239
	DLPFC 151508	4381	2239
	DLPFC 151509	4788	2239
	DLPFC 151510	4595	2239
Human 2	DLPFC 151669	3636	2253
	DLPFC 151670	3484	2253
	DLPFC 151671	4093	2253
	DLPFC 151672	3888	2253
Human 3	DLPFC 151673	3611	3271
	DLPFC 151674	3635	3271
	DLPFC 151675	3566	3271
	DLPFC 151676	3431	3271
Sagittal-Anterior	Sagittal-Anterior section1 (SA-1)	2695	3310
	Sagittal-Anterior section2 (SA-2)	2825	3310
Sagittal-Posterior	Sagittal-Posterior section1 (SP-1)	3355	3310
	Sagittal-Posterior section2 (SP-2)	3289	3310

12.1. Contribution of Core Components

The experimental results clearly reveal the contribution of each core component. Removing the **Spatial Encoder** (`STevs w/o Spatial Encoder`) leads to a significant performance drop in the cross-slice task, which demonstrates that relying solely on a powerful visual feature extractor is insufficient; spatial coordinate information is crucial for capturing the macroscopic patterns of gene expression. Removing the **Latent Space Alignment Loss** (`STevs w/o Alignment Loss`) has a minor impact on the simpler intra-slice task but leads to performance degradation in the cross-slice setting. This indicates that the alignment loss is effective for learning a slice-invariant spatial representation. Finally, removing the **Image Decoder** (`STevs w/o Image Decoder`) causes a slight but consistent decrease in performance, proving that image reconstruction serves as an important auxiliary task and reg-

Table 8. Model Performance on Human 1, Human 2, and Human 3 Datasets

Model Category	Human 1			Human 2			Human 3		
	MSE ↓	PCC ↑	SCC ↑	MSE ↓	PCC ↑	SCC ↑	MSE ↓	PCC ↑	SCC ↑
Local Image-based									
ST-Net (Nat. B.E. [19])	1.494 ± 0.048	0.033 ± 0.022	0.070 ± 0.011	1.348 ± 0.028	0.053 ± 0.023	0.086 ± 0.020	0.365 ± 0.068	0.123 ± 0.019	0.134 ± 0.045
BLEEP (NeurIPS [70])	1.551 ± 0.058	0.036 ± 0.003	0.037 ± 0.005	1.365 ± 0.067	0.058 ± 0.011	0.057 ± 0.013	1.067 ± 0.085	0.086 ± 0.007	0.077 ± 0.008
Graph-based Context									
EGN (PR [73])	0.995 ± 0.053	0.051 ± 0.006	0.054 ± 0.005	1.008 ± 0.035	0.053 ± 0.006	0.066 ± 0.011	0.997 ± 0.074	0.103 ± 0.016	0.109 ± 0.014
IGI-DL (Cell R.M. [17])	0.205 ± 0.009	0.115 ± 0.008	0.117 ± 0.008	0.297 ± 0.023	0.155 ± 0.048	0.152 ± 0.048	0.284 ± 0.048	0.138 ± 0.036	0.124 ± 0.026
Transformer-based Context									
iStar (Nat. Biot. [76])	0.149 ± 0.050	0.191 ± 0.029	0.189 ± 0.016	0.194 ± 0.058	0.204 ± 0.016	0.229 ± 0.008	0.171 ± 0.036	0.236 ± 0.020	0.230 ± 0.018
TRIPLEX (CVPR [10])	0.181 ± 0.009	0.131 ± 0.007	0.125 ± 0.007	0.211 ± 0.006	0.194 ± 0.012	0.186 ± 0.013	0.179 ± 0.014	0.211 ± 0.020	0.199 ± 0.020
M2ORT (AAAI [63])	1.000 ± 0.003	-0.001 ± 0.001	0.000 ± 0.002	1.006 ± 0.006	-0.001 ± 0.001	-0.000 ± 0.001	1.019 ± 0.010	-0.000 ± 0.001	-0.000 ± 0.002
Coordinate-based Generative									
STAGE (NAR [32])	0.259 ± 0.007	0.108 ± 0.013	0.105 ± 0.017	0.307 ± 0.018	0.139 ± 0.034	0.130 ± 0.029	0.339 ± 0.046	0.150 ± 0.016	0.149 ± 0.013
STevs (Ours)	**0.142 ± 0.008**	**0.215 ± 0.011**	**0.202 ± 0.008**	**0.188 ± 0.005**	**0.281 ± 0.005**	**0.271 ± 0.004**	**0.166 ± 0.014**	**0.296 ± 0.019**	**0.263 ± 0.020**

Table 9. Model Performance on Human 1, Human 2, and Human 3 Datasets

Model Category	Human 1			Human 2			Human 3		
	MSE ↓	PCC ↑	SCC ↑	MSE ↓	PCC ↑	SCC ↑	MSE ↓	PCC ↑	SCC ↑
Local Image-based									
ST-Net (Nat. B.E. [19])	1.471 ± 0.060	0.009 ± 0.019	0.062 ± 0.009	1.571 ± 0.054	0.008 ± 0.026	0.063 ± 0.019	1.283 ± 0.110	0.040 ± 0.069	0.043 ± 0.090
BLEEP (NeurIPS [70])	1.758 ± 0.084	0.029 ± 0.009	0.030 ± 0.011	1.274 ± 0.028	0.039 ± 0.018	0.036 ± 0.019	1.574 ± 0.105	0.039 ± 0.046	0.034 ± 0.039
Graph-based Context									
EGN (PR [73])	0.905 ± 0.033	0.049 ± 0.005	0.056 ± 0.006	0.896 ± 0.022	0.052 ± 0.023	0.045 ± 0.033	0.937 ± 0.078	0.052 ± 0.050	0.052 ± 0.058
IGI-DL (Cell R.M. [17])	0.717 ± 0.070	0.059 ± 0.011	0.059 ± 0.010	1.859 ± 0.038	0.029 ± 0.038	0.030 ± 0.034	1.908 ± 0.041	0.008 ± 0.026	0.001 ± 0.030
Transformer-based Context									
iStar (Nat. Biot. [76])	0.262 ± 0.046	0.126 ± 0.018	0.136 ± 0.012	0.215 ± 0.053	0.105 ± 0.073	0.109 ± 0.088	0.319 ± 0.162	0.122 ± 0.095	0.118 ± 0.096
TRIPLEX (CVPR [10])	0.487 ± 0.009	0.097 ± 0.018	0.092 ± 0.016	0.566 ± 0.038	0.083 ± 0.086	0.083 ± 0.084	0.814 ± 0.037	0.071 ± 0.125	0.069 ± 0.117
M2ORT (AAAI [63])	1.205 ± 0.078	0.005 ± 0.007	0.005 ± 0.006	1.188 ± 0.035	-0.004 ± 0.006	-0.002 ± 0.004	1.106 ± 0.151	-0.001 ± 0.003	0.001 ± 0.002
Coordinate-based Generative									
STAGE (NAR [32])	1.186 ± 0.018	0.044 ± 0.038	0.042 ± 0.036	0.921 ± 0.055	0.046 ± 0.045	0.047 ± 0.043	0.615 ± 0.518	0.074 ± 0.045	0.077 ± 0.048
STevs (Ours)	**0.145 ± 0.008**	**0.153 ± 0.018**	**0.152 ± 0.015**	**0.202 ± 0.036**	**0.167 ± 0.075**	**0.166 ± 0.071**	**0.174 ± 0.024**	**0.256 ± 0.032**	**0.231 ± 0.029**

ularizer, compelling the encoder to learn more fine-grained visual representations.

12.2. Effectiveness of the Multimodal Fusion Mechanism

Our probabilistic PoE fusion mechanism is significantly superior to other common fusion methods. Simple **Feature Concatenation** (`STevs (Concat)`) and **Deterministic Mean Fusion** (`STevs (Deterministic)`) perform far below our model, especially in the cross-slice task, as they cannot effectively handle the uncertainty and relative importance of different modalities. Interestingly, the more advanced **Cross-Attention** mechanism (`STevs (Cross-Attention)`), while showing competitive performance on some intra-slice tasks, exhibits insufficient generalization ability with a noticeable performance drop in the cross-slice setting. This, in turn, highlights the superiority of our PoE-based probabilistic fusion method in modeling uncertainty and enhancing generalization.

12.3. Choice of Encoder Architectures

Architectural comparisons validate the rationale of our choices. For the image encoder, the **Swin Transformer**, with its hierarchical structure, outperforms both a standard `ViT` and a traditional `CNN`. Furthermore, removing the **ImageNet pre-trained weights** (`STevs w/o Pretrained`) leads to a substantial drop in performance, demonstrating the effectiveness of transfer learning. For the spatial encoder, we also explored more complex alternatives, including a **Gaussian Process** (`STevs (Gaussian Process)`) and an **MLP with Fourier features** (`STevs (MLP w/ Fourier)`). Although these variants perform adequately on the intra-slice task, their performance degrades severely in the cross-slice setting. This indicates that our simple MLP, when combined with our fusion and alignment strategy, provides a more robust and generalizable foundation.

In summary, this series of exhaustive ablation studies, spanning different species and task difficulties, systematically validates the necessity and advanced nature of each design element in the STevs model, collectively forming the solid foundation for its accurate and robust predictions in diverse biological scenarios. Additionally, the decoder could be modified into a Transformer decoder to consider gene co-expression and potentially improve prediction performance

Table 10. Intra-slice Performance Comparison on 10x Mouse Brain Datasets

Model Category	Sagittal-Anterior			Sagittal-Posterior		
	MSE ↓	PCC ↑	SCC ↑	MSE ↓	PCC ↑	SCC ↑
Local Image-based						
ST-Net (Nat. B.E. [19])	0.896 ± 0.024	0.051 ± 0.005	0.066 ± 0.004	0.915 ± 0.006	0.043 ± 0.006	0.121 ± 0.018
BLEEP (NeurIPS [70])	0.772 ± 0.005	0.086 ± 0.013	0.087 ± 0.019	0.967 ± 0.057	0.123 ± 0.010	0.117 ± 0.009
Graph-based Context						
EGN (PR [73])	0.739 ± 0.043	0.084 ± 0.013	0.076 ± 0.015	1.087 ± 0.011	0.099 ± 0.008	0.107 ± 0.010
IGI-DL (Cell R.M. [17])	0.324 ± 0.068	0.239 ± 0.046	0.242 ± 0.051	0.584 ± 0.033	0.292 ± 0.004	0.264 ± 0.011
Transformer-based Context						
iStar (Nat. Biot. [76])	0.254 ± 0.054	0.384 ± 0.025	0.375 ± 0.055	0.264 ± 0.103	0.459 ± 0.012	0.397 ± 0.011
TRIPLEX (CVPR [10])	0.372 ± 0.021	0.232 ± 0.018	0.216 ± 0.018	0.345 ± 0.006	0.315 ± 0.011	0.297 ± 0.006
M2ORT (AAAI [63])	1.008 ± 0.004	0.001 ± 0.001	0.001 ± 0.000	1.020 ± 0.023	0.001 ± 0.001	0.001 ± 0.000
Coordinate-based Generative						
STAGE (NAR [32])	0.462 ± 0.067	0.104 ± 0.035	0.094 ± 0.039	0.502 ± 0.037	0.120 ± 0.048	0.123 ± 0.051
STevs (Ours)	**0.239 ± 0.015**	**0.413 ± 0.028**	**0.396 ± 0.040**	**0.208 ± 0.008**	**0.486 ± 0.008**	**0.423 ± 0.008**

Table 11. Cross-slice Performance Comparison on 10x Mouse Brain Datasets

Model Category	Sagittal-Anterior			Sagittal-Posterior		
	MSE ↓	PCC ↑	SCC ↑	MSE ↓	PCC ↑	SCC ↑
Local Image-based						
ST-Net (Nat. B.E. [19])	1.861 ± 0.084	0.010 ± 0.008	0.052 ± 0.010	1.502 ± 0.020	0.071 ± 0.022	0.131 ± 0.002
BLEEP (NeurIPS [70])	1.436 ± 0.027	0.069 ± 0.003	0.067 ± 0.002	1.229 ± 0.009	0.118 ± 0.024	0.111 ± 0.024
Graph-based Context						
EGN (PR [73])	1.159 ± 0.055	0.084 ± 0.006	0.075 ± 0.004	0.825 ± 0.014	0.117 ± 0.000	0.130 ± 0.002
IGI-DL (Cell R.M. [17])	0.918 ± 0.071	0.089 ± 0.005	0.087 ± 0.007	0.924 ± 0.024	0.118 ± 0.037	0.126 ± 0.030
Transformer-based Context						
iStar (Nat. Biot. [76])	0.273 ± 0.070	0.301 ± 0.014	0.300 ± 0.017	0.269 ± 0.105	0.363 ± 0.038	0.325 ± 0.030
TRIPLEX (CVPR [10])	0.438 ± 0.022	0.197 ± 0.010	0.180 ± 0.013	0.450 ± 0.008	0.256 ± 0.006	0.247 ± 0.006
M2ORT (AAAI [63])	1.133 ± 0.000	0.006 ± 0.002	0.007 ± 0.003	1.253 ± 0.145	0.001 ± 0.006	0.001 ± 0.005
Coordinate-based Generative						
STAGE (NAR [32])	0.624 ± 0.028	0.125 ± 0.021	0.118 ± 0.019	0.631 ± 0.021	0.156 ± 0.028	0.158 ± 0.026
STevs (Ours)	**0.261 ± 0.011**	**0.362 ± 0.014**	**0.351 ± 0.018**	**0.223 ± 0.003**	**0.442 ± 0.000**	**0.392 ± 0.006**

[13]; however, this is beyond the scope of this paper's focus on representation fusion.

13. Gene Expression Visualization for Each Dataset

This appendix section provides supplementary visualizations for the gene expression prediction performance of STevs and all baseline models, corresponding to the results discussed in the main manuscript. The following figures are organized by the two core validation strategies.

Intra-Slice Validation Figures 8-12 display the qualitative results for the intra-slice validation task. For these experiments, models were evaluated on a 20% held-out test set from within each individual slice. Visualizations are shown for representative spatially variable genes: OLFM1 [55] for the `human1`, `human2`, and `human3` datasets, and Dgkz [26] for the `anterior` and `posterior` datasets.

Cross-Slice Validation Figures 13 to 19 present the results for the more challenging cross-slice validation task. A model is trained on a single slice from a group (e.g., D151507) and evaluated on all other unseen slices from the same group. These figures visually demonstrate the robust generalization capability of our model in contrast to the baseline methods.

Table 12. Intra-slice cross-validation Performance Comparison of STevs Variants on DLPFC Datasets

Model Variant	Human 1			Human 2			Human 3		
	MSE ↓	PCC ↑	SCC ↑	MSE ↓	PCC ↑	SCC ↑	MSE ↓	PCC ↑	SCC ↑
Component Ablation									
STevs w/o Image Decoder	0.151 ± 0.010	0.203 ± 0.015	0.195 ± 0.016	0.192 ± 0.012	0.274 ± 0.018	0.260 ± 0.019	0.176 ± 0.018	0.283 ± 0.021	0.253 ± 0.023
STevs w/o Spatial Encoder	0.171 ± 0.025	0.172 ± 0.021	0.162 ± 0.022	0.226 ± 0.031	0.225 ± 0.025	0.217 ± 0.026	0.199 ± 0.032	0.237 ± 0.028	0.210 ± 0.030
STevs w/o Alignment Loss	0.147 ± 0.009	0.209 ± 0.012	0.200 ± 0.013	0.188 ± 0.010	0.280 ± 0.015	0.266 ± 0.016	0.172 ± 0.015	0.289 ± 0.019	0.261 ± 0.020
Fusion Mechanism Ablation									
STevs (Concat)	0.171 ± 0.018	0.191 ± 0.017	0.179 ± 0.019	0.226 ± 0.021	0.243 ± 0.024	0.239 ± 0.025	0.201 ± 0.022	0.260 ± 0.027	0.222 ± 0.028
STevs (Deterministic)	0.184 ± 0.015	0.179 ± 0.014	0.177 ± 0.015	0.241 ± 0.018	0.234 ± 0.020	0.231 ± 0.021	0.213 ± 0.019	0.251 ± 0.023	0.215 ± 0.024
STevs (Cross-Attention)	0.143 ± 0.014	0.213 ± 0.019	0.201 ± 0.018	0.187 ± 0.015	0.283 ± 0.021	0.273 ± 0.020	0.167 ± 0.021	0.294 ± 0.026	0.261 ± 0.027
Spatial Encoder Variants									
STevs (Gaussian Process)	0.144 ± 0.010	0.212 ± 0.013	0.200 ± 0.014	0.189 ± 0.011	0.279 ± 0.016	0.269 ± 0.017	0.168 ± 0.016	0.293 ± 0.020	0.260 ± 0.021
STevs (MLP w/ Fourier)	0.145 ± 0.011	0.210 ± 0.014	0.198 ± 0.015	0.191 ± 0.012	0.276 ± 0.017	0.265 ± 0.018	0.170 ± 0.017	0.290 ± 0.021	0.258 ± 0.022
Architecture Variants									
STevs (Convolutional)	0.217 ± 0.022	0.163 ± 0.018	0.162 ± 0.019	0.259 ± 0.026	0.215 ± 0.028	0.203 ± 0.029	0.246 ± 0.028	0.224 ± 0.031	0.203 ± 0.033
STevs (ViT)	0.149 ± 0.012	0.210 ± 0.015	0.194 ± 0.014	0.199 ± 0.015	0.271 ± 0.018	0.266 ± 0.017	0.176 ± 0.018	0.290 ± 0.023	0.252 ± 0.025
STevs w/o Pretrained	0.191 ± 0.026	0.176 ± 0.025	0.173 ± 0.024	0.243 ± 0.030	0.231 ± 0.033	0.218 ± 0.032	0.223 ± 0.033	0.239 ± 0.037	0.213 ± 0.036
STevs (Ours)	$\mathbf{0.142 \pm 0.013}$	$\mathbf{0.215 \pm 0.018}$	$\mathbf{0.202 \pm 0.017}$	$\mathbf{0.188 \pm 0.014}$	$\mathbf{0.281 \pm 0.020}$	$\mathbf{0.271 \pm 0.019}$	$\mathbf{0.166 \pm 0.020}$	$\mathbf{0.296 \pm 0.025}$	$\mathbf{0.263 \pm 0.026}$

Table 13. Cross-slice cross-validation Performance Comparison of STevs Variants on DLPFC Datasets

Model Variant	Human 1			Human 2			Human 3		
	MSE ↓	PCC ↑	SCC ↑	MSE ↓	PCC ↑	SCC ↑	MSE ↓	PCC ↑	SCC ↑
Component Ablation									
STevs w/o Image Decoder	0.160 ± 0.018	0.143 ± 0.025	0.138 ± 0.024	0.227 ± 0.035	0.148 ± 0.041	0.150 ± 0.040	0.196 ± 0.033	0.233 ± 0.040	0.203 ± 0.038
STevs w/o Spatial Encoder	0.350 ± 0.031	0.103 ± 0.015	0.113 ± 0.016	0.324 ± 0.033	0.144 ± 0.020	0.137 ± 0.021	0.357 ± 0.038	0.156 ± 0.023	0.153 ± 0.024
STevs w/o Alignment Loss	0.158 ± 0.017	0.145 ± 0.024	0.142 ± 0.023	0.225 ± 0.036	0.150 ± 0.042	0.148 ± 0.041	0.195 ± 0.032	0.235 ± 0.039	0.205 ± 0.037
Fusion Mechanism Ablation									
STevs (Concat)	0.211 ± 0.022	0.117 ± 0.019	0.110 ± 0.018	0.290 ± 0.029	0.122 ± 0.025	0.122 ± 0.026	0.252 ± 0.031	0.190 ± 0.028	0.165 ± 0.027
STevs (Deterministic)	0.233 ± 0.019	0.107 ± 0.016	0.108 ± 0.017	0.303 ± 0.024	0.120 ± 0.022	0.111 ± 0.023	0.278 ± 0.026	0.173 ± 0.025	0.157 ± 0.024
STevs (Cross-Attention)	0.155 ± 0.020	0.148 ± 0.027	0.145 ± 0.026	0.242 ± 0.038	0.134 ± 0.044	0.133 ± 0.043	0.255 ± 0.041	0.128 ± 0.045	0.120 ± 0.046
Spatial Encoder Variants									
STevs (Gaussian Process)	0.345 ± 0.030	0.108 ± 0.016	0.117 ± 0.017	0.318 ± 0.032	0.149 ± 0.021	0.141 ± 0.022	0.352 ± 0.037	0.160 ± 0.024	0.156 ± 0.025
STevs (MLP w/ Fourier)	0.341 ± 0.029	0.112 ± 0.017	0.120 ± 0.018	0.315 ± 0.031	0.152 ± 0.022	0.145 ± 0.023	0.349 ± 0.036	0.163 ± 0.025	0.159 ± 0.026
Architecture Variants									
STevs (Convolutional)	0.298 ± 0.032	0.089 ± 0.014	0.090 ± 0.013	0.362 ± 0.038	0.104 ± 0.018	0.095 ± 0.019	0.343 ± 0.040	0.132 ± 0.021	0.124 ± 0.022
STevs (ViT)	0.166 ± 0.019	0.140 ± 0.026	0.134 ± 0.025	0.237 ± 0.037	0.145 ± 0.043	0.147 ± 0.042	0.201 ± 0.034	0.229 ± 0.041	0.200 ± 0.039
STevs w/o Pretrained	0.254 ± 0.036	0.099 ± 0.035	0.101 ± 0.034	0.319 ± 0.041	0.112 ± 0.048	0.104 ± 0.047	0.300 ± 0.046	0.159 ± 0.053	0.147 ± 0.052
STevs (Ours)	$\mathbf{0.145 \pm 0.021}$	$\mathbf{0.153 \pm 0.028}$	$\mathbf{0.152 \pm 0.027}$	$\mathbf{0.202 \pm 0.040}$	$\mathbf{0.167 \pm 0.045}$	$\mathbf{0.166 \pm 0.044}$	$\mathbf{0.174 \pm 0.038}$	$\mathbf{0.256 \pm 0.042}$	$\mathbf{0.231 \pm 0.040}$

Table 14. Intra-slice cross-validation Performance Comparison of STevs Variants on 10x Mouse Brain Datasets

Model Variant	Sagittal-Anterior			Sagittal-Posterior		
	MSE ↓	PCC ↑	SCC ↑	MSE ↓	PCC ↑	SCC ↑
Component Ablation						
STevs w/o Image Decoder	0.249 ± 0.018	0.403 ± 0.028	0.387 ± 0.030	0.215 ± 0.013	0.473 ± 0.016	0.415 ± 0.015
STevs w/o Spatial Encoder	0.330 ± 0.024	0.317 ± 0.021	0.287 ± 0.022	0.250 ± 0.020	0.389 ± 0.018	0.338 ± 0.019
STevs w/o Alignment Loss	0.241 ± 0.017	0.411 ± 0.027	0.390 ± 0.029	0.213 ± 0.012	0.479 ± 0.015	0.421 ± 0.014
Fusion Mechanism Ablation						
STevs (Concat)	0.353 ± 0.025	0.300 ± 0.022	0.362 ± 0.026	0.262 ± 0.021	0.428 ± 0.023	0.373 ± 0.022
STevs (Deterministic)	0.353 ± 0.020	0.336 ± 0.019	0.310 ± 0.020	0.281 ± 0.017	0.411 ± 0.019	0.363 ± 0.018
STevs (Cross-Attention)	0.240 ± 0.022	0.411 ± 0.033	0.394 ± 0.034	0.206 ± 0.017	0.488 ± 0.021	0.425 ± 0.020
Spatial Encoder Variants						
STevs (Gaussian Process)	0.240 ± 0.016	0.410 ± 0.026	0.393 ± 0.028	0.209 ± 0.011	0.483 ± 0.014	0.420 ± 0.013
STevs (MLP w/ Fourier)	0.238 ± 0.016	0.415 ± 0.027	0.399 ± 0.029	0.211 ± 0.012	0.480 ± 0.015	0.417 ± 0.014
Architecture Variants						
STevs (Convolutional)	0.351 ± 0.028	0.307 ± 0.025	0.322 ± 0.027	0.332 ± 0.025	0.382 ± 0.026	0.313 ± 0.024
STevs (ViT)	0.251 ± 0.019	0.403 ± 0.029	0.391 ± 0.031	0.217 ± 0.014	0.476 ± 0.017	0.414 ± 0.016
STevs w/o Pretrained	0.329 ± 0.034	0.347 ± 0.036	0.318 ± 0.035	0.290 ± 0.031	0.403 ± 0.033	0.357 ± 0.032
STevs (Ours)	$\mathbf{0.239 \pm 0.021}$	$\mathbf{0.413 \pm 0.032}$	$\mathbf{0.396 \pm 0.033}$	$\mathbf{0.208 \pm 0.016}$	$\mathbf{0.486 \pm 0.020}$	$\mathbf{0.423 \pm 0.019}$

14. ACCURATE RECOVERY OF SPATIAL DOMAINS

To evaluate whether the gene expression profiles predicted by our model can accurately recover the spatial domains

Table 15. Cross-slice cross-validation Performance Comparison of STevs Variants on 10x Mouse Brain Datasets

Model Variant	Sagittal-Anterior			Sagittal-Posterior		
	MSE ↓	PCC ↑	SCC ↑	MSE ↓	PCC ↑	SCC ↑
Component Ablation						
STevs w/o Image Decoder	0.291 ± 0.026	0.327 ± 0.029	0.309 ± 0.030	0.252 ± 0.020	0.403 ± 0.021	0.356 ± 0.022
STevs w/o Spatial Encoder	0.433 ± 0.035	0.212 ± 0.024	0.209 ± 0.025	0.351 ± 0.031	0.280 ± 0.026	0.263 ± 0.027
STevs w/o Alignment Loss	0.285 ± 0.027	0.331 ± 0.030	0.315 ± 0.031	0.249 ± 0.021	0.408 ± 0.022	0.360 ± 0.023
Fusion Mechanism Ablation						
STevs (Concat)	0.372 ± 0.031	0.266 ± 0.028	0.261 ± 0.029	0.319 ± 0.026	0.356 ± 0.027	0.308 ± 0.028
STevs (Deterministic)	0.353 ± 0.027	0.251 ± 0.025	0.230 ± 0.026	0.347 ± 0.024	0.337 ± 0.025	0.400 ± 0.026
STevs (Cross-Attention)	0.313 ± 0.030	0.290 ± 0.028	0.280 ± 0.029	0.268 ± 0.024	0.354 ± 0.026	0.314 ± 0.027
Spatial Encoder Variants						
STevs (Gaussian Process)	0.425 ± 0.034	0.218 ± 0.023	0.214 ± 0.024	0.345 ± 0.030	0.287 ± 0.025	0.270 ± 0.026
STevs (MLP w/ Fourier)	0.421 ± 0.033	0.223 ± 0.024	0.219 ± 0.025	0.340 ± 0.029	0.291 ± 0.026	0.275 ± 0.027
Architecture Variants						
STevs (Convolutional)	0.450 ± 0.041	0.207 ± 0.031	0.196 ± 0.033	0.410 ± 0.038	0.279 ± 0.029	0.257 ± 0.030
STevs (ViT)	0.318 ± 0.029	0.301 ± 0.027	0.309 ± 0.028	0.261 ± 0.023	0.394 ± 0.020	0.356 ± 0.022
STevs w/o Pretrained	0.413 ± 0.052	0.237 ± 0.043	0.226 ± 0.044	0.370 ± 0.040	0.319 ± 0.036	0.286 ± 0.037
STevs (Ours)	$\mathbf{0.261 \pm 0.033}$	$\mathbf{0.362 \pm 0.035}$	$\mathbf{0.350 \pm 0.036}$	$\mathbf{0.223 \pm 0.028}$	$\mathbf{0.442 \pm 0.030}$	$\mathbf{0.392 \pm 0.029}$

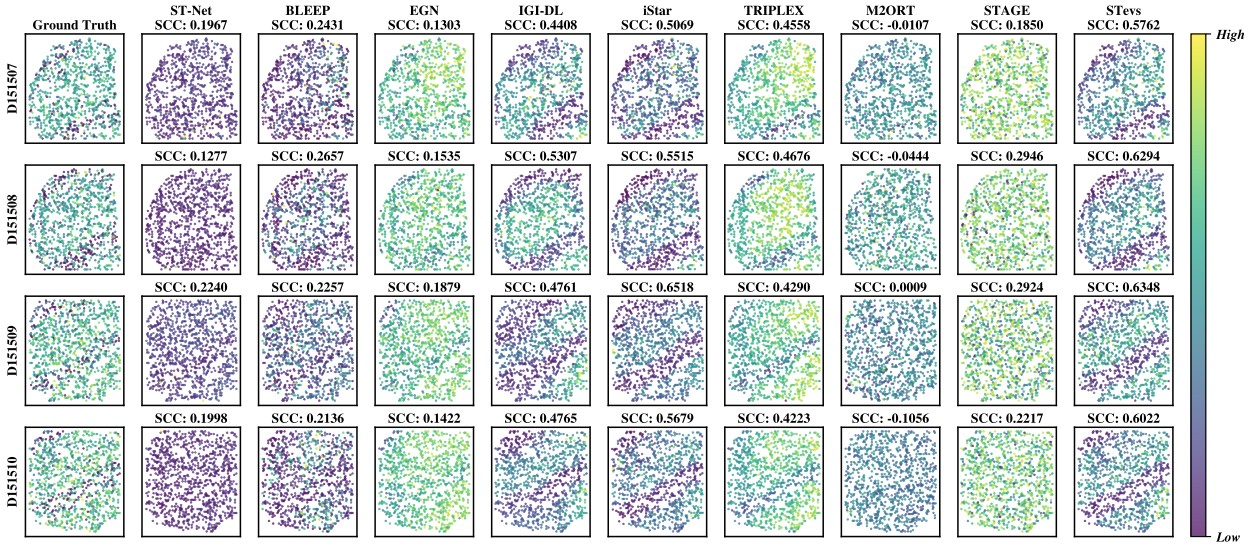

Figure 8. The results of intra-slice validation for the OLFM1 gene on 4 datasets of human1 (with a 20% test set)

of the tissue, we performed clustering analyses[38] on the expression profiles generated by each method after cross-slice training, and calculated the ARI against the manually annotated ground-truth domains. As summarized in Table 16, STevs demonstrates superior performance across all five dataset groups. Notably, the ARI score from clustering on STevs's predictions not only significantly surpasses that of other advanced predictive models like iStar, but also consistently outperforms the baseline results from clustering on the original ground-truth RNA counts. This suggests that the predictions from STevs may serve a denoising function[16], capturing the essential biological structures more clearly than the potentially noisy raw data, thereby enabling a more accurate recovery of the tissue's spatial domains.

15. Parameter Sensitivity Analysis

As shown in the Figure 20, we conducted a sensitivity analysis on three key hyperparameters: latent dimension, learning rate, and KLD loss weight. The model's performance is relatively sensitive to the choice of learning rate, while showing some, but not particularly high, sensitivity to the latent dimension and KLD loss weight. The parameters achieve optimal performance within a specific range.

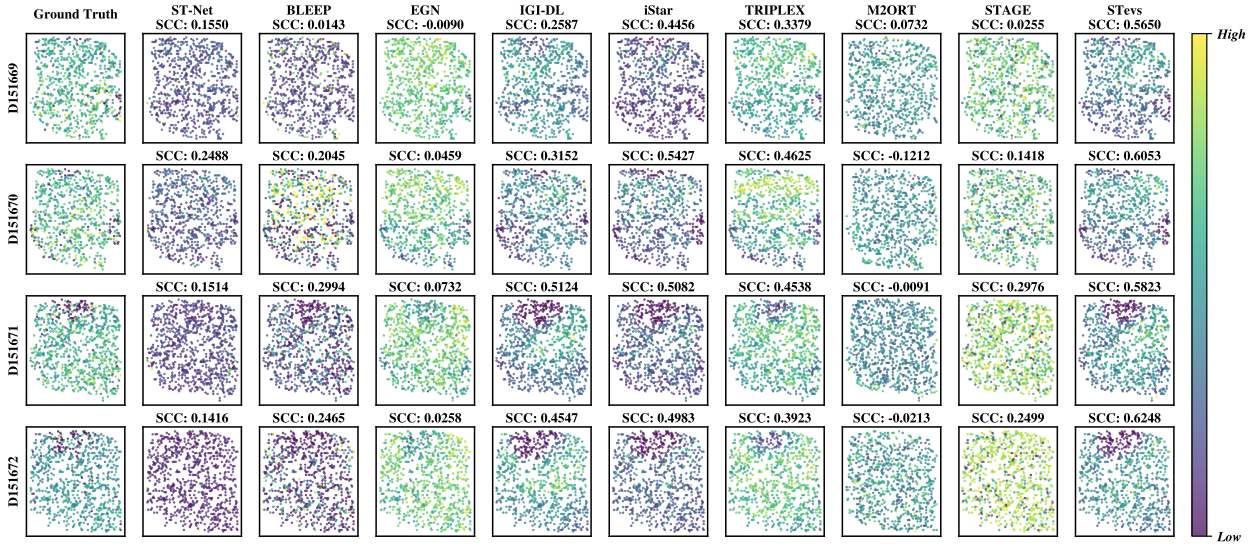

Figure 9. The results of intra-slice validation for the OLFM1 gene on 4 datasets of human2 (with a 20% test set)

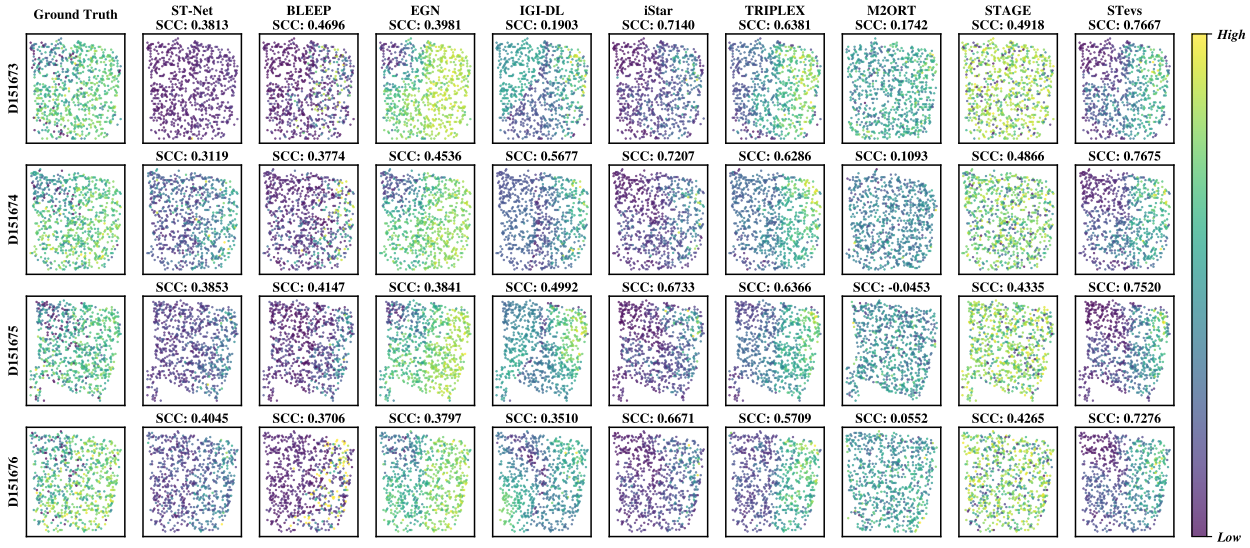

Figure 10. The results of intra-slice validation for the OLFM1 gene on 4 datasets of human3 (with a 20% test set)

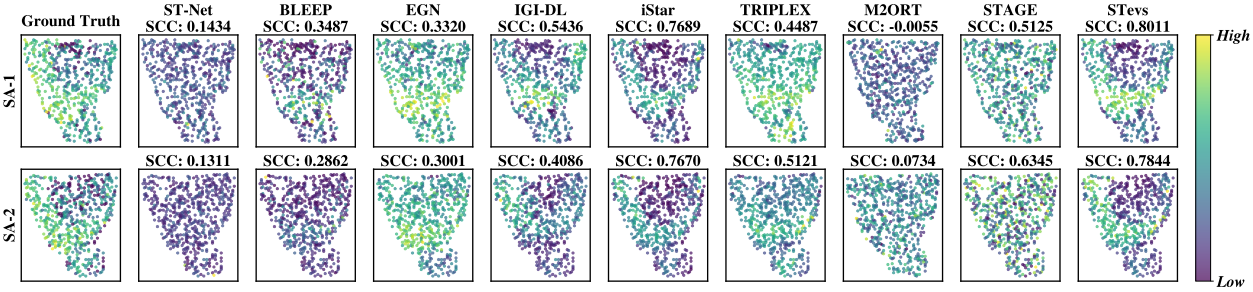

Figure 11. The results of intra-slice validation for the Dgkz gene on 2 datasets of anterior (with a 20% test set)

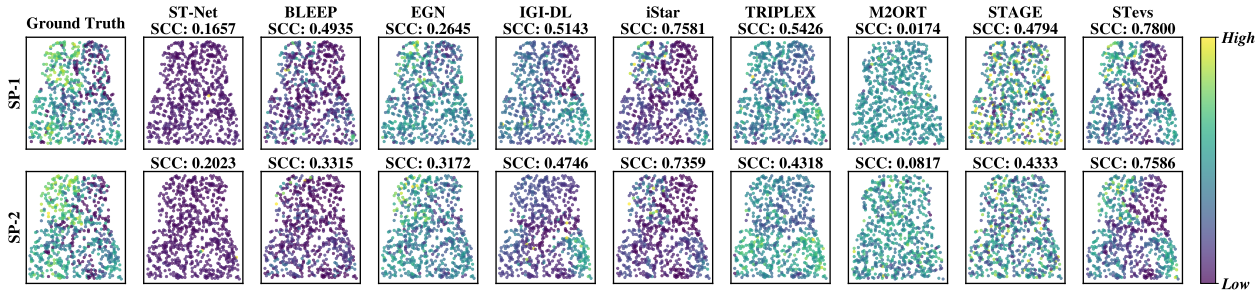

Figure 12. The results of intra-slice validation for the Dgkz gene on 2 datasets of posterior (with a 20% test set)

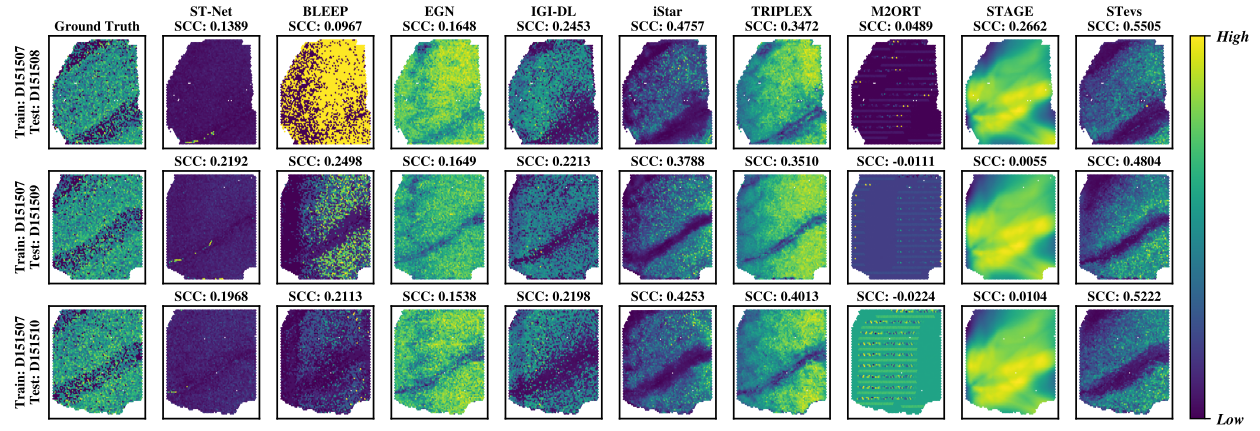

Figure 13. The cross-slice validation results of the OLFM1 gene on the other 3 slices of human1, with D151507 used as the training set

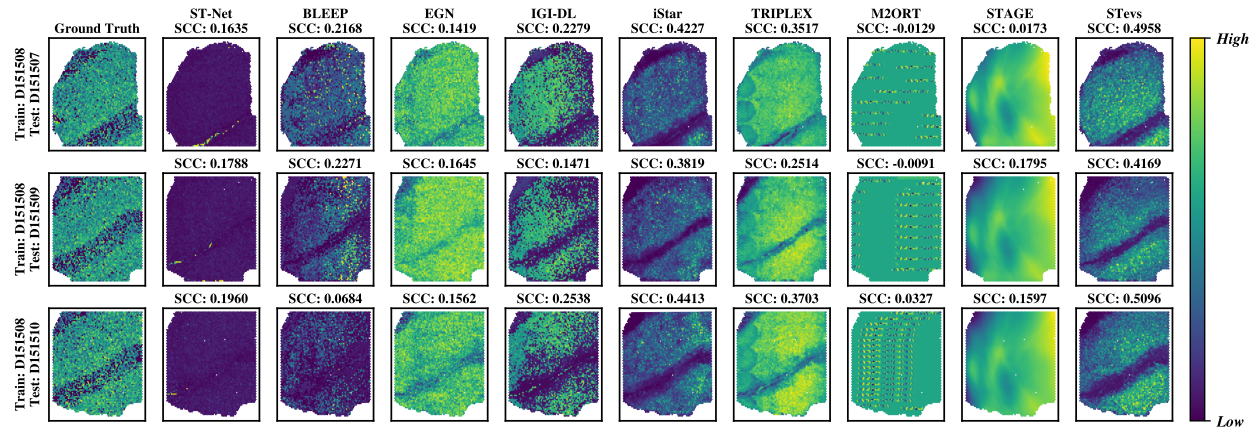

Figure 14. The cross-slice validation results of the OLFM1 gene on the other 3 slices of human1, with D151508 used as the training set

16. Extended Experiment

16.1. Robustness Experiments

16.1.1. Methodology

To comprehensively evaluate the robustness of our model, particularly its ability to handle color variations arising from different experimental batches or staining procedures,

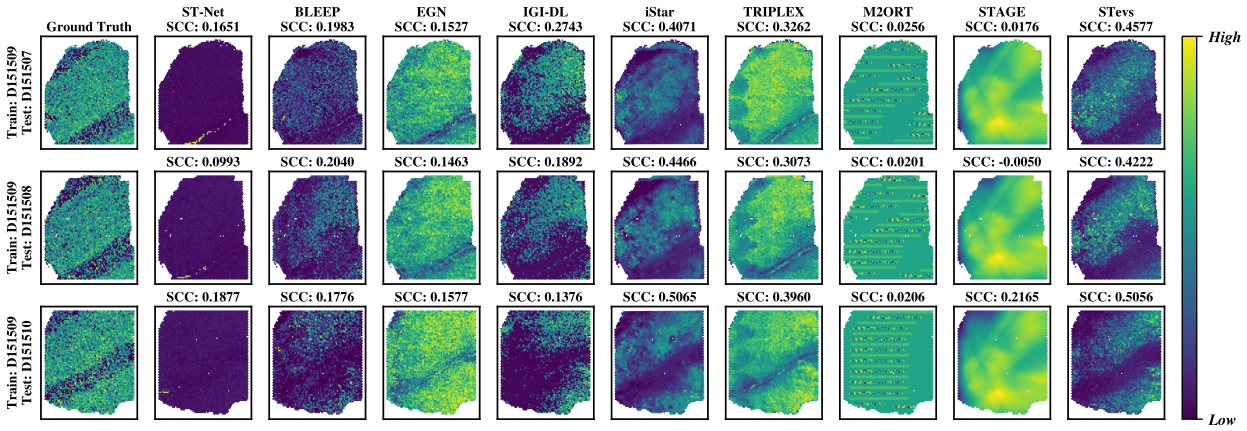

Figure 15. The cross-slice validation results of the OLFM1 gene on the other 3 slices of human1, with D151509 used as the training set

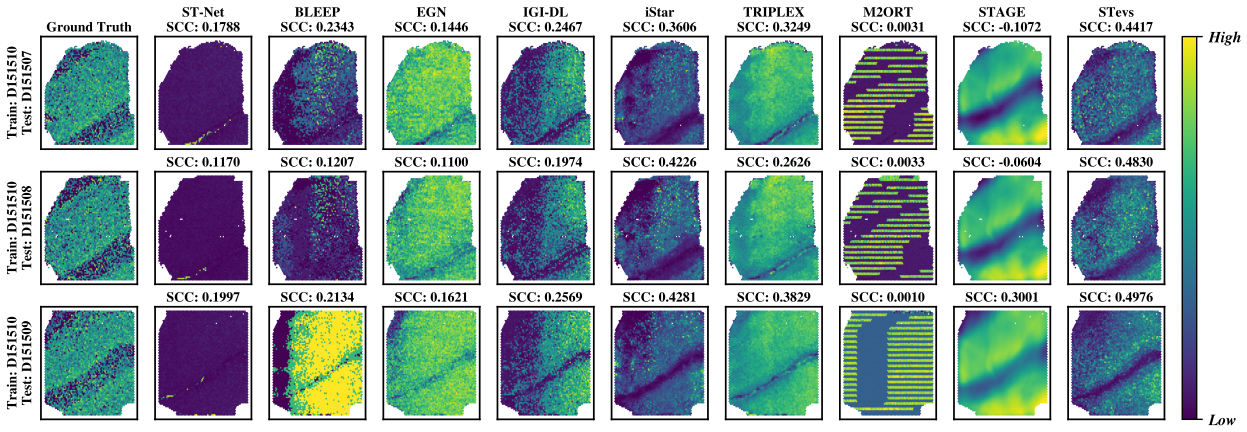

Figure 16. The cross-slice validation results of the OLFM1 gene on the other 3 slices of human1, with D151510 used as the training set

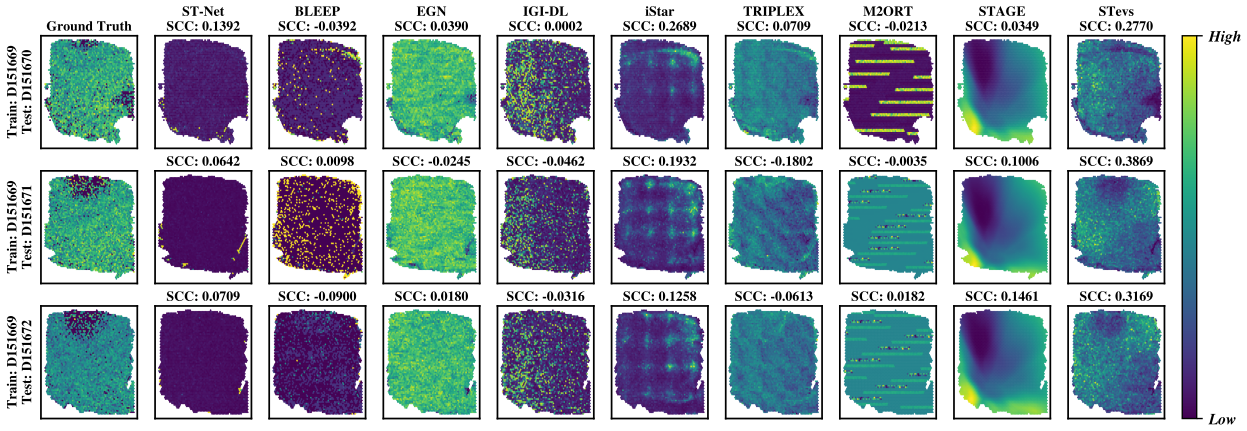

Figure 17. The cross-slice validation results of the OLFM1 gene on the other 3 slices of human2, with D151669 used as the training set

we designed and conducted a color augmentation simulation. We selected the H&E images from the D151673 and D151674 datasets and applied a **Spectral Blue Shift** transformation. This transformation is controlled by an intensity

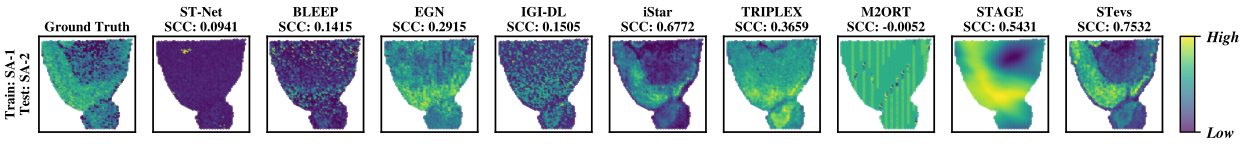

Figure 18. The cross-slice validation results of the Dgkz gene on the other 2 slices of anterior, with SA-1 used as the training set

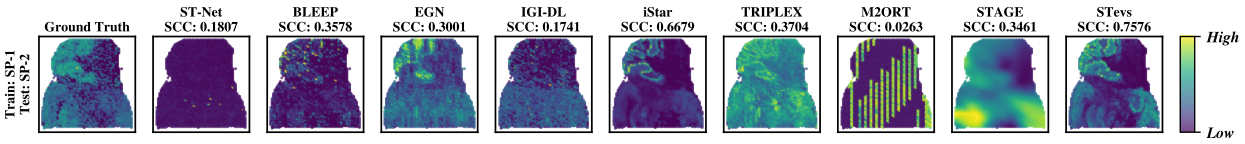

Figure 19. The cross-slice validation results of the Dgkz gene on the other 2 slices of posterior, with SP-1 used as the training set

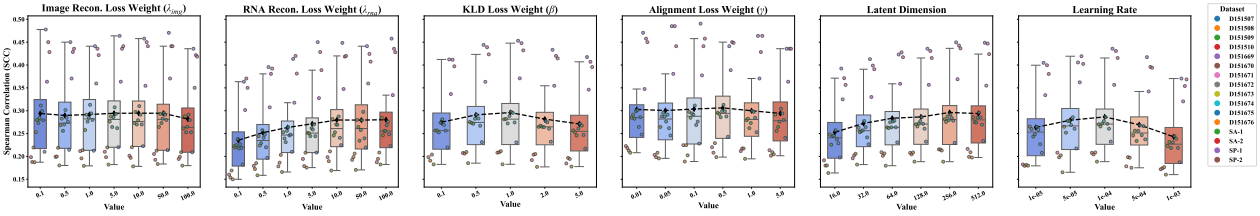

Figure 20. Parameter Sensitivity Analysis

parameter, α, which we varied from 0.1 to 1.0 in increments of 0.1, thereby generating a series of images with a progressively blueish hue. The key advantage of this method is its ability to induce a global spectral shift across the image without altering the microscopic cellular histology or macroscopic tissue structure, thus effectively simulating inter-slice color variations (i.e., batch effects).

The specific operation of this transformation on any given pixel color value, represented as $P = [R, G, B]^T$ (normalized to the range $[0, 1]$), is defined by the following mathematical formula:

$$P_{\text{shifted}} = \text{clip}_{[0,255]} \left(\begin{bmatrix} R \\ G \\ B \end{bmatrix} + 255 * \begin{bmatrix} -\alpha \\ -\alpha \\ +\alpha \end{bmatrix} \right) \quad (40)$$

where P_{shifted} is the transformed pixel value and α is the parameter controlling the intensity of the color shift.

16.1.2. Results

We tested the model on the images processed with varying intensities of the Spectral Blue Shift and recorded its performance. The experimental results, as illustrated in Figure 21, clearly show the trend of the model's performance as a function of the color shift intensity (α). As observed in the figure, the performance of all models exhibited a decline with increasing spectral distortion. However, our proposed model demonstrated superior robustness. Compared to the baseline methods, our model's performance degradation curve was considerably flatter, with its advantages becoming more pronounced at higher intensity levels (e.g., $\alpha > 0.5$). This result strongly demonstrates that our model is insensitive to color variations in H&E images.

16.2. Extended Dataset Experiment

16.2.1. Dataset

The first group of datasets consists of three Human Breast Cancer (HBC) [69] samples, sourced from a publicly available Visium dataset. Breast cancer serves as a classic model for studying the heterogeneity of the Tumor Microenvironment (TME), as its tissue sections contain multiple cell types—including tumor, stromal, and infiltrating immune cells—that present a complex spatial architecture, making it highly suitable for evaluating the foundational generalization performance of a model. The second group is from a study on Human Squamous Cell Carcinoma (HSC) published by [27], which is renowned for its high-quality multimodal data. From this, we selected three technical replicate sections from Patient 10. This provides an ideal validation scenario to rigorously test our model's stability and consistency when processing technical replicates from the same source tissue. The third group is derived from a MISAR-seq (Microfluidic Indexing-based Spatial Assay for Transposase-Accessible Chromatin and RNA-sequencing) [28] dataset, chosen to evaluate the model's capability in handling complex spatiotemporal and multi-

Table 16. ARI Comparison across different methods and datasets

width=center

	DLPFC Dataset			10x Mouse Brain Dataset	
Method	**Human 1**	**Human 2**	**Human 3**	**Sagittal-Anterior**	**Sagittal-Posterior**
RNA Counts	0.139 ± 0.032	0.113 ± 0.014	0.176 ± 0.009	0.067 ± 0.001	0.053 ± 0.001
STevs(Ours)	$\mathbf{0.246 \pm 0.040}$	$\mathbf{0.238 \pm 0.069}$	$\mathbf{0.238 \pm 0.051}$	$\mathbf{0.280 \pm 0.002}$	$\mathbf{0.295 \pm 0.018}$
iStar	0.214 ± 0.037	0.071 ± 0.063	0.172 ± 0.075	0.241 ± 0.016	0.221 ± 0.006
TRIPLEX	0.080 ± 0.024	0.018 ± 0.016	0.123 ± 0.036	0.098 ± 0.004	0.107 ± 0.009

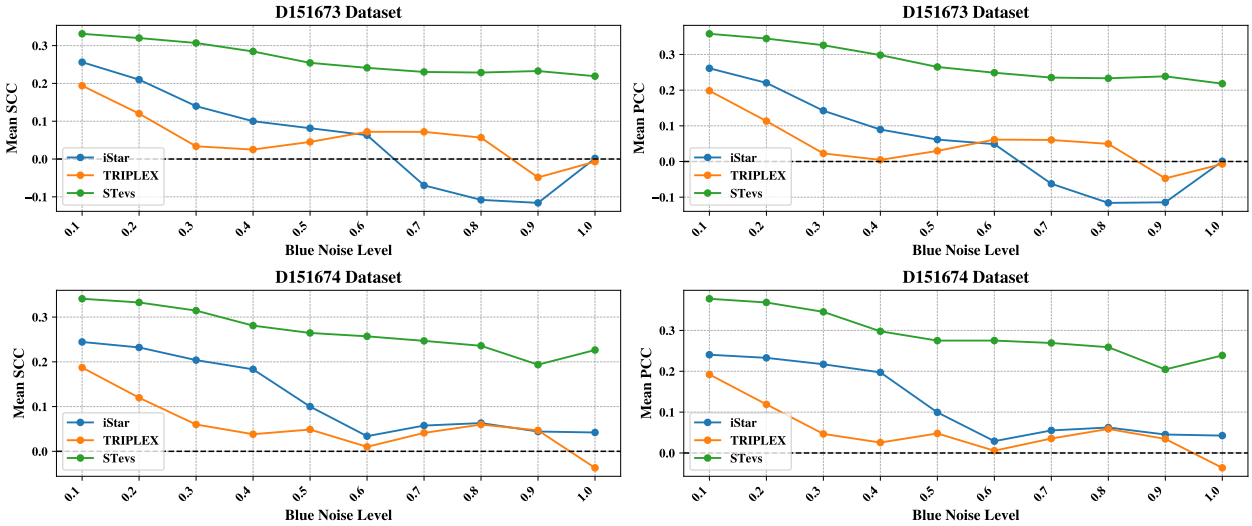

Figure 21. Model performance evaluation under varying levels of Spectral Blue Shift (Blue Noise Level) across the D151673 and D151674 datasets.

omics data. This advanced dataset simultaneously provides spatial transcriptomics and spatial chromatin accessibility information from the same tissue section. We utilized data from different individuals at distinct developmental time points (E15 and E18) to challenge the model's robustness against biological variability. During the data preprocessing stage, we performed Spatially Variable Gene (SVG) filtering on each dataset. For the HBC dataset, a final set of 851 high-confidence SVGs was retained for subsequent experiments. Similarly, for the HSC dataset, we obtained 1483 SVGs. For the MISAR-seq dataset, a total of 678 SVGs were selected for subsequent experiments. The style of the sections for all datasets is illustrated in Figure 22.

16.2.2. Results

The detailed experimental results are presented in Table 17 (for the intra-slice task) and Table 18 (for the cross-slice task), demonstrating the superior performance of our model. In both prediction settings, STevs consistently outperforms all baseline methods across the HBC, HSC, and MISAR datasets, achieving the lowest MSE and the highest PCC and SCC correlations.

17. Additional Experimental Details

17.1. Running Time

We evaluated the computational efficiency of STevs against all baseline models. As detailed in the training time tables in the Appendix, STevs demonstrates excellent computational performance. In both intra-slice and cross-slice settings, the training time for STevs is significantly lower than that of other high-performance Transformer-based models, such as iStar and M2ORT, and is one to two orders of magnitude faster than STAGE. This indicates that STevs maintains low computational overhead while achieving state-of-the-art predictive performance, showcasing an excellent balance between efficiency and accuracy.

17.2. Parameter Settings for Other Methods

To ensure a fair comparison with baseline models and to facilitate the reproducibility of our results, we conducted a comprehensive hyperparameter search for all baselines. As detailed in the parameter search table in the Appendix, we defined a search space for the key hyperparameters of each

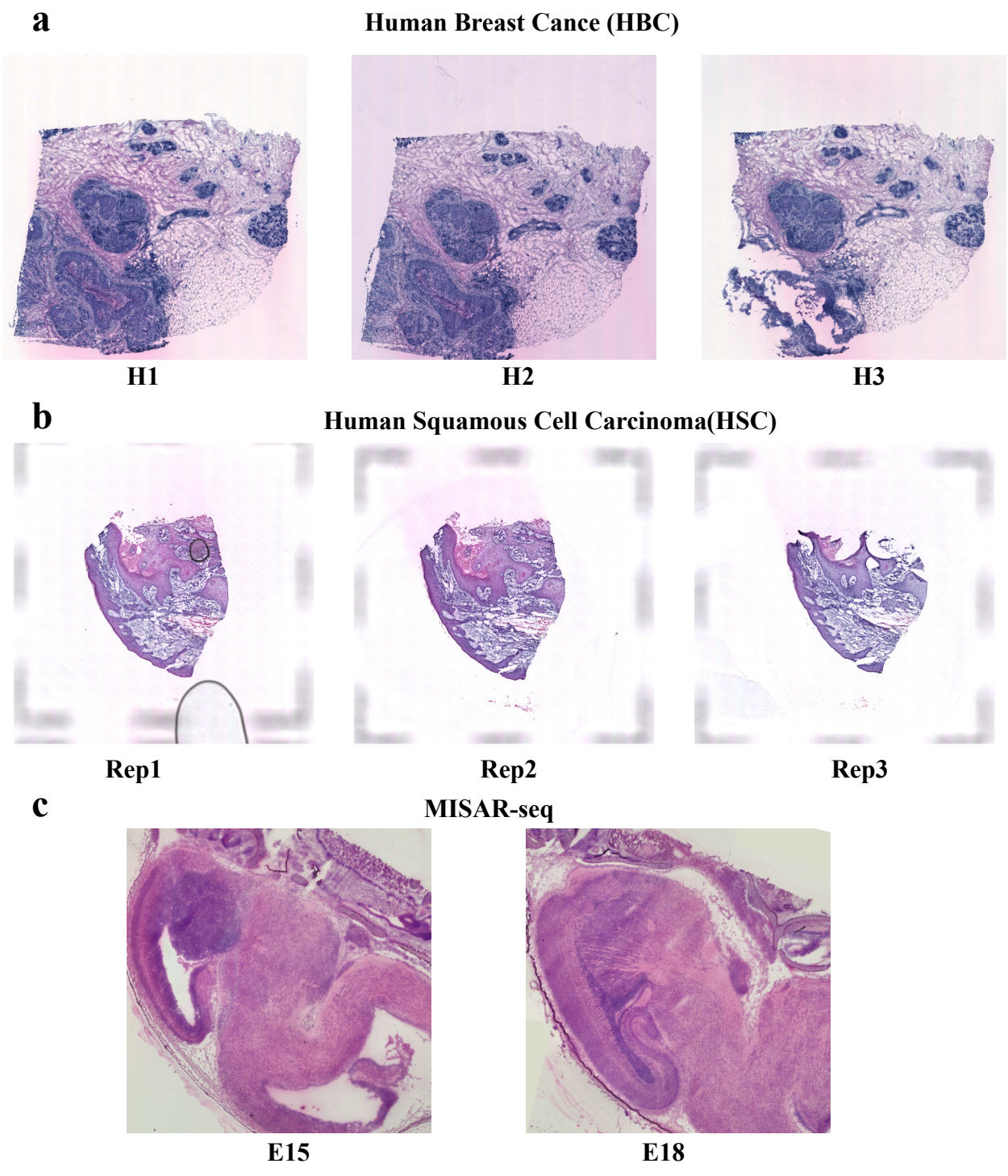

Figure 22. H&E stained images of the datasets used in the additional experiments. (Top row) Three HBC samples from a public dataset. (Bottom row) Three technical replicate samples of HSC from [27].

model and determined the optimal parameter combination for each dataset independently. This targeted tuning strategy ensures that every baseline model was performing at or near its optimal state for comparison against STevs, thereby validating the rigor of our experimental evaluation.

Table 17. Intra-slice Performance comparison of different models on the HBC, HSC, and MISAR datasets.

Method	HBC			HSC			MISAR		
	MSE ↓	PCC ↑	SCC ↑	MSE ↓	PCC ↑	SCC ↑	MSE ↓	PCC ↑	SCC ↑
IGI-DL	0.8055 ± 0.4531	0.1269 ± 0.0957	0.1284 ± 0.0761	0.7310 ± 0.8815	0.1724 ± 0.0674	0.1349 ± 0.0675	0.9542 ± 0.2588	0.1150 ± 0.0880	0.1090 ± 0.0813
iStar	0.3804 ± 0.1322	0.3001 ± 0.0652	0.2441 ± 0.0680	0.8981 ± 0.4588	0.5789 ± 0.0420	0.4058 ± 0.0354	0.4958 ± 0.5120	0.3787 ± 0.0450	0.3595 ± 0.0411
TRIPLEX	0.2100 ± 0.0216	0.2390 ± 0.0720	0.2290 ± 0.0701	0.3070 ± 0.0901	0.3597 ± 0.0526	0.2884 ± 0.0251	0.5982 ± 0.1105	0.2591 ± 0.0615	0.2513 ± 0.0588
STevs (Ours)	**0.1559** ± 0.0162	**0.3802** ± 0.0672	**0.3125** ± 0.0598	**0.1855** ± 0.0487	**0.5951** ± 0.0313	**0.4350** ± 0.0374	**0.3988** ± 0.0415	**0.3986** ± 0.0391	**0.3866** ± 0.0352

Table 18. Cross-slice Performance comparison of different models on the HBC, HSC, and MISAR datasets.

Method	HBC			HSC			MISAR		
	MSE ↓	PCC ↑	SCC ↑	MSE ↓	PCC ↑	SCC ↑	MSE ↓	PCC ↑	SCC ↑
IGI-DL	1.3441 ± 0.1693	0.0899 ± 0.0871	0.0901 ± 0.0908	1.4990 ± 0.1542	0.1280 ± 0.0629	0.1147 ± 0.0543	1.6567 ± 0.2289	0.1287 ± 0.0347	0.1061 ± 0.0321
iStar	0.3977 ± 0.0836	0.2731 ± 0.0221	0.2013 ± 0.0304	0.9479 ± 0.3782	0.4849 ± 0.0347	0.3700 ± 0.0176	0.6946 ± 0.0807	0.1602 ± 0.0145	0.1521 ± 0.0135
TRIPLEX	0.1999 ± 0.0272	0.2577 ± 0.0222	0.2393 ± 0.0252	0.3256 ± 0.0809	0.3141 ± 0.0243	0.2626 ± 0.0166	0.9548 ± 0.1671	0.1096 ± 0.0225	0.1063 ± 0.0188
STevs (Ours)	**0.1619** ± 0.0092	**0.3640** ± 0.0333	**0.2939** ± 0.0260	**0.2719** ± 0.0195	**0.5443** ± 0.0338	**0.4039** ± 0.0237	**0.5699** ± 0.0553	**0.2087** ± 0.0132	**0.2136** ± 0.0110

17.3. Quantitative Evaluation of Latent Space Integration

To quantitatively support the qualitative visualization of our representation spaces (as shown in Figure 4 of the main text), we computed the Average Silhouette Width (ASW) across different latent modalities. Evaluating spatial transcriptomics integration requires balancing two competing objectives: effectively removing batch effects across different tissue slices while preserving the intrinsic biological structures (e.g., cell types).

We utilized two specific ASW metrics to evaluate the representation spaces:

- **Slice ASW:** Measures the separation of cells originating from different slices. In the context of batch integration, an ASW value close to 0 is highly desirable, as it indicates that cells from different slices are thoroughly mixed and the boundaries between slices are effectively eliminated.
- **Cell Type ASW:** Measures the cohesion and separation of true biological cell types. A higher value indicates better preservation of the biological identity despite the cross-slice fusion process.

Table 19. Quantitative evaluation of cross-sample integration and biological structure preservation across different representation spaces. For Slice ASW, values closer to 0 indicate better cross-slice mixing. For Cell Type ASW, higher values indicate better preservation of biological variance.

Representation Space	Slice ASW	Cell Type ASW ↑
Raw Image	-0.0295	-0.0111
Image Latent	-0.0127	-0.0127
Spatial Latent	-0.0627	-0.2882
Fused Latent (STevs)	**-0.0138**	**-0.0074**

As shown in Table 19, both the Image Latent and Fused Latent representations achieve ideal near-zero Slice ASW scores, demonstrating that slice-specific spatial-concomitant heterogeneity is effectively mitigated. Furthermore, among all compared spaces, the **Fused Latent** representation achieves the highest Cell Type ASW (-0.0074). This quantitative result firmly demonstrates that our Product of Experts (PoE) fusion mechanism successfully aligns the diverse sample distributions across slices without compromising the underlying biological structures.

18. The Use of Large Language Models (LLMs)

During the preparation of this manuscript, we utilized Large Language Models (LLMs) as writing assistants. Specifically, we used Gemini Pro and DeepSeek to improve the grammar, clarity, and readability of the text. The models' role was strictly limited to rephrasing sentences for better flow and correcting typographical errors. The core scientific ideas, experimental design, analysis, and conclusions presented in this paper were conceived and developed entirely by the human authors. We have carefully reviewed and edited all model-generated text and take full responsibility for the final content of this paper, ensuring its scientific accuracy and originality.

Table 20. Intra-slice Training Time(s). Our experiments were conducted on a high-performance server equipped with four NVIDIA A100 GPUs (80GB of VRAM each), dual Intel(R) Xeon(R) Gold 6267C CPUs, and 1.5TB of system memory. The runtimes reported in the table are for running the model on a single GPU.

width=center

Dataset	**ST-Net**	**BLEEP**	**EGN**	**IGI-DL**	**iStar**	**TRIPLEX**	**M2ORT**	**STAGE**	**STevs**
D151507	816.66	710.66	723.82	46.42	4263.32	329.58	3273.99	19840.76	315.07
D151508	632.69	670.33	776.64	45.91	5057.34	311.76	3854.78	18280.31	349.79
D151509	677.69	629.27	973.14	44.98	4784.25	368.39	3470.88	23201.72	354.20
D151510	694.55	632.45	917.32	38.12	5266.73	328.05	3575.34	17053.08	332.63
D151669	588.12	694.86	925.39	38.45	3457.32	269.61	3583.80	16024.78	288.73
D151670	576.13	530.80	796.37	44.07	3211.23	249.29	3950.09	14621.26	238.13
D151671	676.64	641.43	799.86	41.83	3415.35	298.14	3870.76	19408.45	272.04
D151672	691.75	499.68	813.08	42.16	3454.13	280.87	3661.85	24743.11	277.08
D151673	680.40	602.45	904.72	49.36	4236.42	283.91	3270.12	17625.90	268.60
D151674	886.31	656.57	902.83	52.55	5283.45	279.64	3584.43	21101.03	258.41
D151675	690.43	481.23	946.43	47.80	3436.56	421.84	3289.74	20454.75	246.32
D151676	647.71	597.35	925.94	56.38	4203.35	274.28	3611.79	17363.65	302.48
SA-1	639.54	514.74	878.00	53.71	4201.45	236.85	3467.52	19567.36	196.62
SA-2	775.66	687.63	816.49	54.96	2376.21	249.96	3359.48	21511.63	198.71
SP-1	851.26	444.32	961.98	56.53	3201.29	266.62	3690.96	24367.12	260.48
SP-2	661.83	442.03	920.32	53.27	3815.53	266.39	3879.65	26614.65	272.23

Table 21. Cross-slice Training Time(s). Our experiments were conducted on a high-performance server equipped with four NVIDIA A100 GPUs (80GB of VRAM each), dual Intel(R) Xeon(R) Gold 6267C CPUs, and 1.5TB of system memory. The runtimes reported in the table are for running the model on a single GPU.

width=center

Dataset	**ST-Net**	**BLEEP**	**EGN**	**IGI-DL**	**iStar**	**TRIPLEX**	**M2ORT**	**STAGE**	**STevs**
D151507	720.07	302.15	870.61	51.40	5813.85	715.52	4123.56	18426.30	413.85
D151508	870.90	360.68	978.36	51.77	6344.25	672.18	4087.91	27341.57	494.83
D151509	799.45	354.93	1385.73	46.03	6427.45	681.49	4210.34	21036.94	597.39
D151510	862.53	403.01	1112.65	42.23	5132.24	795.83	3989.45	19508.30	508.69
D151669	748.22	219.61	863.06	48.53	5383.45	655.76	4056.22	23193.66	598.03
D151670	1024.44	282.19	857.38	57.46	6642.34	622.20	4188.76	21346.31	569.06
D151671	792.82	313.80	968.52	57.28	5632.64	743.64	3998.11	22938.56	470.93
D151672	981.05	310.15	928.61	56.01	4330.84	895.64	4065.99	17814.49	500.48
D151673	899.40	309.10	1047.13	48.23	5245.45	839.44	4176.54	26316.95	425.79
D151674	777.04	509.31	1126.61	47.37	5246.56	787.96	3954.88	24910.94	454.83
D151675	841.35	406.07	1220.79	47.41	4256.57	786.02	4022.65	24798.90	556.87
D151676	788.38	367.95	1095.60	48.63	6423.63	746.98	4101.99	28050.22	544.18
SA-1	730.78	309.12	914.99	54.55	5356.29	472.68	4005.43	18317.55	454.80
SA-2	915.74	359.40	1016.93	56.32	4356.43	493.94	3978.22	22109.28	284.63
SP-1	901.73	371.56	1228.21	95.84	5485.25	651.50	4011.67	27389.33	518.49
SP-2	935.85	338.43	1224.59	61.14	4352.24	653.60	4155.88	28165.83	527.08

Table 22. Baseline Parameter Search Results

Model	Hyperparameter	Search Range	D151507	D151508	D151509	D151510	D151669	D151670	D151671	D151672	D151673	D151674	D151675	D151676	SA-1	SA-2	SP-1	SP-2
ST-Net	learning_rate	[1e-4, 5e-4, 1e-3]	5e-04	1e-03	5e-04	5e-04	1e-04	5e-04	5e-04	1e-03	5e-04	5e-04	5e-04	5e-04	1e-03	5e-04	1e-03	1e-03
	l2_reg	[0.001, 0.005, 0.01]	0.005	0.005	0.001	0.005	0.005	0.01	0.005	0.01	0.005	0.005	0.005	0.005	0.001	0.001	0.005	0.001
BLEEP	lr	[1e-4, 5e-4, 1e-3]	1e-03	1e-03	1e-03	1e-03	1e-03	1e-03	1e-03	1e-03	1e-03	1e-03	1e-03	1e-03	1e-03	1e-03	1e-03	1e-03
	hidden_dim	[128, 256, 512]	256	256	256	512	128	256	256	256	512	256	256	256	512	256	512	512
	lambda	[0.5, 1, 2]	1	1	1	0.5	1	2	1	1	1	2	1	1	0.5	1	0.5	0.5
EGN	lr	[1e-4, 5e-4, 1e-3]	5e-04	5e-04	5e-04	5e-04	5e-04	5e-04	5e-04	5e-04	5e-04	5e-04	5e-04	5e-04	1e-03	1e-03	1e-03	1e-03
	hidden_dim	[64, 128, 256]	128	128	128	64	128	128	128	128	128	128	256	128	256	256	256	256
	num_layers	[2, 3, 4]	3	2	3	3	4	3	3	3	3	2	3	3	4	3	4	4
	dropout	[0.3, 0.5, 0.7]	0.5	0.5	0.5	0.5	0.3	0.5	0.7	0.5	0.5	0.5	0.5	0.5	0.3	0.5	0.3	0.3
	lam	[0.1, 0.5, 1.0]	0.5	0.5	0.5	1.0	0.5	0.5	0.5	0.1	0.5	0.5	0.5	0.5	1	1	1	1
IGI-DL	lr	[1e-4, 5e-4, 1e-3]	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04	5e-04	5e-04	5e-04	5e-04
	gat_hidden_dim	[128, 256, 512]	256	256	256	256	256	256	256	256	256	256	256	256	128	128	128	128
	gat_layer_num	[2, 3, 4]	4	4	4	4	3	3	3	3	3	3	3	3	2	2	2	2
	gat_dropout	[0.1, 0.2, 0.3]	0.2	0.1	0.2	0.2	0.2	0.3	0.2	0.2	0.1	0.2	0.2	0.2	0.1	0.2	0.1	0.1
iStar	lr	[1e-4, 5e-4, 1e-3]	5e-04	5e-04	1e-03	5e-04	5e-04	5e-04	5e-04	5e-04	5e-04	1e-04	5e-04	5e-04	1e-03	1e-03	1e-03	1e-03
	weight_decay	[1e-5, 1e-4, 1e-3]	1e-04	1e-04	1e-04	1e-05	1e-04	1e-04	1e-04	1e-04	1e-03	1e-04	1e-04	1e-04	1e-05	1e-04	1e-05	1e-05
TRIPLEX	learning_rate	[1e-4, 5e-4, 1e-3]	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04	5e-04	5e-04	5e-04	5e-04	1e-04	1e-04	1e-04	1e-04
	n_hidden	[64, 128, 256]	128	128	128	128	128	128	128	128	128	128	128	128	128	128	128	128
	n_layers	[2, 3, 4]	3	3	3	3	3	4	3	3	4	3	3	3	4	4	4	4
	dropout	[0.1, 0.3, 0.5]	0.3	0.3	0.1	0.3	0.3	0.3	0.3	0.5	0.3	0.3	0.3	0.1	0.5	0.3	0.5	0.5
	weight_decay	[1e-5, 1e-4, 1e-3]	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04
	alpha	[0.1, 0.5, 1.0]	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	1	1	1	1
M2ORT	lr	[1e-4, 5e-4, 1e-3]	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04
	weight_decay	[1e-5, 1e-4, 1e-3]	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04	1e-05	1e-05	1e-05	1e-05
STAGE	lr	[0.01, 0.005, 0.001]	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.001	0.001	0.001	0.001
	hidden_dims	[512, 256, 128]	256	256	256	256	256	256	256	256	256	256	256	256	256	256	256	256
	lambda_recon	[0.1, 1, 10]	1	10	1	1	1	1	10	1	1	1	1	1	10	1	10	10
	lambda_kl	[0.1, 1, 10]	1	1	1	0.1	1	1	1	1	1	1	1	0.1	0.1	0.1	1	0.1
	lambda_graph	[0.1, 1, 10]	1	1	0.1	1	1	1	1	1	1	0.1	1	1	0.1	1	0.1	0.1

References

[1] Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018. 4

[2] Nicole M Anderson and M Celeste Simon. The tumor microenvironment. *Current Biology*, 30(16):R921–R925, 2020. 2

[3] Alma Andersson, Ludvig Larsson, Linnea Stenbeck, Fredrik Salmén, Anna Ehinger, Sunny Z. Wu, Ghamdan Al-Eryani, Daniel Roden, Alex Swarbrick, Åke Borg, Jonas Frisén, Camilla Engblom, and Joakim Lundeberg. Spatial deconvolution of her2-positive breast cancer delineates tumor-associated cell type interactions. *Nature Communications*, 12(1):6012, 2021. 1, 2, 4, 6

[4] Michaela Asp, Stefania Giacomello, Ludvig Larsson, Chenglin Wu, Daniel Fürth, Xiaoyan Qian, Eva Wärdell, Joaquin Custodio, Johan Reimegård, Fredrik Salmén, Cecilia Österholm, Patrik L. Ståhl, Erik Sundström, Elisabet Åkesson, Olaf Bergmann, Magda Bienko, Agneta Månsson-Broberg, Mats Nilsson, Christer Sylvén, and Joakim Lundeberg. A spatiotemporal organ-wide gene expression and cell atlas of the developing human heart. *Cell*, 179(7):1647–1660, 2019. 1, 6

[5] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine intelligence*, 41(2):423–443, 2018. 2, 5, 4

[6] Samuel Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL conference on computational natural language learning*, pages 10–21, 2016. 5

[7] Darren J Burgess. Spatial transcriptomics coming of age. *Nature Reviews Genetics*, 20(6):317–317, 2019. 1

[8] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature medicine*, 30(3):850–862, 2024. 8

[9] Wei-Ting Chen, Ashley Lu, Katleen Craessaerts, Benjamin Pavie, Carlo Sala Frigerio, Nikky Corthout, Xiaoyan Qian, Jana Laláková, Malte Kühnemund, Iryna Voytyuk, Leen Wolfs, Renzo Mancuso, Evgenia Salta, Sriram Balusu, An Snellinx, Sebastian Munck, Aleksandra Jurek, Jose Fernandez Navarro, Takaomi C. Saido, Inge Huitinga, Joakim Lundeberg, Mark Fiers, and Bart De Strooper. Spatial transcriptomics and in situ sequencing to study alzheimer's disease. *Cell*, 182(4):976–991, 2020. 1, 4

[10] Youngmin Chung, Ji Hun Ha, Kyeong Chan Im, and Joo Sang Lee. Accurate spatial gene expression prediction by integrating multi-resolution features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11591–11600, 2024. 2, 3, 6, 9, 10

[11] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999. 5

[12] Guanshen Cui, Kangning Dong, Jia-Yi Zhou, Shang Li, Ying Wu, Qinghua Han, Bofei Yao, Qunlun Shen, Yong-Liang Zhao, Ying Yang, Jun Cai, Shihua Zhang, and Yun-Gui Yang. Spatiotemporal transcriptomic atlas reveals the dynamic characteristics and key regulators of planarian regeneration. *Nature Communications*, 14(1):3205, 2023. 1

[13] H. Cui, C. Wang, H. Maan, K. Pang, F. Luo, N. Duan, and B. Wang. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8):1470–1480, 2024. 10

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, and Sylvain Gelly. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[16] Gökcen Eraslan, Lukas M Simon, Maria Mircea, Nikola S Mueller, and Fabian J Theis. Single-cell rna-seq denoising using a deep count autoencoder. *Nature Communications*, 10(1):390, 2019. 12

[17] Ruitian Gao, Xin Yuan, Yanran Ma, Ting Wei, Luke Johnston, Yanfei Shao, Wenwen Lv, Tengteng Zhu, Yue Zhang, and Junke Zheng. Harnessing tme depicted by histological images to improve cancer prognosis through a deep learning system. *Cell Reports Medicine*, 5(5), 2024. 2, 3, 4, 6, 9, 10

[18] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, Zhaohui Yang, Yiman Zhang, and Dacheng Tao. A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):87–110, 2022. 2, 5

[19] Bryan He, Ludvig Bergenstråhle, Linnea Stenbeck, Abubakar Abid, Alma Andersson, Åke Borg, Jonas Maaskola, Joakim Lundeberg, and James Zou. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature Biomedical Engineering*, 4(8):827–834, 2020. 2, 3, 6, 1, 9, 10

[20] Zhen He, Shuofeng Hu, Yaowen Chen, Sijing An, Jiahao Zhou, Runyan Liu, Junfeng Shi, Jing Wang, Guohua Dong, and Jinhui Shi. Mosaic integration and knowledge transfer of single-cell multimodal data with midas. *Nature Biotechnology*, 42(10):1594–1605, 2024. 3

[21] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002. 2, 3, 4

[22] 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. 2

[23] Tinglin Huang, Tianyu Liu, Mehrtash Babadi, Wengong Jin, and Rex Ying. Scalable generation of spatial transcriptomics from histology images via whole-slide flow matching, 2025. 2

[24] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985. 6

[25] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015. 4

[26] Mitsue Ishisaka and Hideaki Hara. The roles of diacylglycerol kinases in the central nervous system: review of genetic studies in mice. *Journal of Pharmacological Sciences*, 124 (3):336–343, 2014. 10

[27] Andrew L. Ji, Adam J. Rubin, Kim Thrane, Sizun Jiang, David L. Reynolds, Robin M. Meyers, Margaret G. Guo, Benson M. George, Annelie Mollbrink, Joseph Bergenstråhle, Ludvig Larsson, Yunhao Bai, Bokai Zhu, Aparna Bhaduri, Jordan M. Meyers, Xavier Rovira-Clavé, S. Tyler Hollmig, Sumaira Z. Aasi, Garry P. Nolan, Joakim Lundeberg, and Paul A. Khavari. Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *Cell*, 182(2):497–514, 2020. 1, 2, 4, 5, 16, 18

[28] Fuqing Jiang, Xin Zhou, Yingying Qian, Miao Zhu, Li Wang, Zhuxia Li, Qingmei Shen, Minhan Wang, Fangfang Qu, Guizhong Cui, Kai Chen, and Guangdun Peng. Simultaneous profiling of spatial gene expression and chromatin accessibility during mouse brain development. *Nature Methods*, 20:1048–1057, 2023. 5, 16

[29] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2, 4, 5, 6

[30] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 2012. 5

[31] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. 3

[32] 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. 2, 6, 9, 10

[33] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 3

[34] Yahui Long, Kok Siong Ang, Mengwei Li, Kian Long Kelvin Chong, Raman Sethi, Chengwei Zhong, Hang Xu, Zhiwei Ong, Karishma Sachaphibulkij, and Ao Chen. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with graphst. *Nature Communications*, 14 (1):1155, 2023. 2, 3

[35] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018. 2, 3, 4

[36] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[37] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, page 3. Atlanta, GA, 2013. 4

[38] J MacQueen. Multivariate observations. In *Proceedings ofthe 5th Berkeley Symposium on Mathematical Statisticsand Probability*, pages 281–297, 1967. 12

[39] Kristen R Maynard, Leonardo Collado-Torres, Lukas M Weber, Cedric Uytingco, Brianna K Barry, Stephen R Williams, Joseph L Catallini, Matthew N Tran, Zachary Besich, and Madhavi Tippani. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature Neuroscience*, 24(3):425–436, 2021. 5, 6

[40] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 1

[41] Zhen Miao, Tian Tian, Wei Chen, Qianwen Wang, Liang Ma, Dan Zhang, Min Xie, Zijin Yu, Xiya Guo, Genxiang Bai, Shaoli Zhao, Xi Chen, Wenyi Wang, Yizhou Gao, Shicheng Guo, Ming Luo, Ling Yuan, Caihuan Tian, Liang Wu, Guangchuang Yu, Dake Zhang, and Shiquan Sun. Spatial resolved transcriptomics: Computational insights into gene transcription across tissue and organ architecture in diverse applications. *The Innovation Life*, 2(4):100097, 2024. 2

[42] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 5

[43] Reuben Moncada, Dalia Barkley, Florian Wagner, Marta Chiodin, Joseph C Devlin, Maayan Baron, Cristina H Hajdu, Diane M Simeone, and Itai Yanai. Integrating microarray-based spatial transcriptomics and single-cell rna-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nature Biotechnology*, 38(3):333–342, 2020. 1

[44] Taku Monjo, Masaru Koido, Satoi Nagasawa, Yutaka Suzuki, and Yoichiro Kamatani. Efficient prediction of a spatial transcriptomics profile better characterizes breast cancer tissue sections without costly experimentation. *Scientific Reports*, 12(1):4133, 2022. 2

[45] Nikhil Naik, Ali Madani, Andre Esteva, Nitish Shirish Keskar, Michael F Press, Daniel Ruderman, David B Agus, and Richard Socher. Deep learning-enabled breast cancer hormonal receptor status determination from base-level h&e stains. *Nature Communications*, 11(1):5727, 2020. 2, 4

[46] Giovanni Palla, Hannah Spitzer, Michal Klein, David Fischer, Anna Christina Schaar, Louis Benedikt Kuemmerle, Sergei Rybakov, Ignacio L Ibarra, Olle Holmberg, and Isaac Virshup. Squidpy: a scalable framework for spatial omics analysis. *Nature Methods*, 19(2):171–178, 2022. 5

[47] Minxing Pang, Kenong Su, and Mingyao Li. Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. *BioRxiv*, pages 2021–11, 2021. 2

[48] Karl Pearson. Vii. mathematical contributions to the theory of evolution.—iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, (187):253–318, 1896. 5, 6

[49] Xiaojie Qiu, Daniel Y Zhu, Yifan Lu, Jiajun Yao, Zehua Jing, Kyung Hoi Min, Mengnan Cheng, Hailin Pan, Lulu

Zuo, Samuel King, Qi Fang, Huiwen Zheng, Mingyue Wang, Shuai Wang, Qingquan Zhang, Sichao Yu, Sha Liao, Chao Liu, Xinchao Wu, Yiwei Lai, Shijie Hao, Zhewei Zhang, Liang Wu, Yong Zhang, Mei Li, Zhencheng Tu, Jinpei Lin, Zhuoxuan Yang, Yuxiang Li, Ying Gu, David Ellison, Yuancheng Ryan Lu, Qinan Hu, Yuhui Hu, Ao Chen, Longqi Liu, Jonathan S. Weissman, Jiayi Ma, Xun Xu, Shiping Liu, and Yinqi Bai. Spatiotemporal modeling of molecular holograms. *Cell*, 187(26):7351–7373, 2024. 5

[50] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 4, 5

[51] Anjali Rao, Dalia Barkley, Gustavo S França, and Itai Yanai. Exploring tissue architecture using spatial transcriptomics. *Nature*, 596(7871):211–220, 2021. 1

[52] Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization? *Advances in neural information processing systems*, 31, 2018. 4

[53] Benoît Schmauch, Alberto Romagnoni, Elodie Pronier, Charlie Saillard, Pascale Maillé, Julien Calderaro, Aurélie Kamoun, Meriem Sefta, Sylvain Toldo, and Mikhail Zaslavskiy. A deep learning model to predict rna-seq expression of tumours from whole slide images. *Nature Communications*, 11(1):3877, 2020. 2

[54] Amelia Schroeder, Melanie L Loth, Chunyu Luo, Sicong Yao, Hanying Yan, Daiwei Zhang, Sarbottam Piya, Edward Plowey, Wenxing Hu, Jean R Clemenceau, Inyeop Jang, Minji Kim, Isabel Barnfather, Su Jing Chan, Taylor L. Reynolds, Thomas Carlile, Patrick Cullen, Ji-Youn Sung, Hui-Hsin Tsai, Jeong Hwan Park, Tae Hyun Hwang, Baohong Zhang, and Mingyao Li. Scaling up spatial transcriptomics for large-sized tissues: uncovering cellular-level tissue architecture beyond conventional platforms with iscale. *Nature Methods*, pages 1–12, 2025. 2

[55] Yiqi Shen, Yao Shen, Menglei Wang, Kaiyu Jin, Penghui Yang, Zuozhen Cao, Qinfeng Zhu, Zhiyong Zhao, Haotian Li, Lei Han, Shiping Liu, Jie Liao, Jing Zhang, Xiaohui Fan, and Dan Wu. A spatial imaging-transcriptomics paradigm for deciphering the molecular basis of microscopic mri in the normal brain and alzheimer's disease. *Cell Reports*, 44 (8), 2025. 6, 10

[56] Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 100(3/4):441–471, 1987. 5, 6

[57] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014. 4

[58] Patrik L Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O Westholm, and Mikael Huss. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294): 78–82, 2016. 2, 5

[59] Robert R Stickels, Evan Murray, Pawan Kumar, Jilong Li, Jamie L Marshall, Daniela J Di Bella, Paola Arlotta, Evan Z Macosko, and Fei Chen. Highly sensitive spatial transcriptomics at near-cellular resolution with slide-seqv2. *Nature Biotechnology*, 39(3):313–319, 2021. 1, 4

[60] Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891*, 2016. 2, 3

[61] 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. 4

[62] Sophia J Wagner, Daniel Reisenbüchler, Nicholas P West, Jan Moritz Niehues, Jiefu Zhu, Sebastian Foersch, Gregory Patrick Veldhuizen, Philip Quirke, Heike I Grabsch, and Piet A van den Brandt. Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study. *Cancer Cell*, 41(9):1650–1661, 2023. 2, 4

[63] Hongyi Wang, Xiuju Du, Jing Liu, Shuyi Ouyang, Yen-Wei Chen, and Lanfen Lin. M2ost: Many-to-one regression for predicting spatial transcriptomics from digital pathology images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7709–7717, 2025. 3, 6, 9, 10

[64] Mingyue Wang, Qinan Hu, Tianhang Lv, Yuhang Wang, Qing Lan, Rong Xiang, Zhencheng Tu, Yanrong Wei, Kai Han, Chang Shi, Junfu Guo, Chao Liu, Tao Yang, Wensi Du, Yanru An, Mengnan Cheng, Jiangshan Xu, Haorong Lu, Wangsheng Li, Shaofang Zhang, Ao Chen, Wei Chen, Yuxiang Li, Xiaoshan Wang, Xun Xu, Yuhui Hu, and Longqi Liu. High-resolution 3d spatiotemporal transcriptomic maps of developing drosophila embryos and larvae. *Developmental Cell*, 57(10):1271–1283, 2022. 1, 2

[65] Cameron G. Williams, Hyun Jae Lee, Takahiro Asatsuma, Roser Vento-Tormo, and Ashraful Haque. An introduction to spatial transcriptomics for biomedical research. *Genome Medicine*, 14(1):68, 2022. 2

[66] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006. 5, 8

[67] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, 2018. 5

[68] Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. *Advances in Neural Information Processing Systems*, 31, 2018. 5

[69] Sunny Z Wu, Ghamdan Al-Eryani, Daniel Lee Roden, Simon Junankar, Kate Harvey, Alma Andersson, Aatish Thennavan, Chenfei Wang, James R Torpy, and Nenad Bartonicek. A single-cell and spatially resolved atlas of human breast cancers. *Nature Genetics*, 53(9):1334–1347, 2021. 5, 16

[70] Ronald Xie, Kuan Pang, Sai Chung, Catia Perciani, Sonya MacParland, Bo Wang, and Gary Bader. Spatially resolved gene expression prediction from histology images via bimodal contrastive learning. *Advances in Neural Information Processing Systems*, 36:70626–70637, 2023. 2, 6, 9, 10

[71] Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero,

Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, 2024. 8

[72] Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12113–12132, 2023. 3, 4, 5

[73] Yan Yang, Md Zakir Hossain, Eric Stone, and Shafin Rahman. Spatial transcriptomics analysis of gene expression prediction using exemplar guided graph neural network. *Pattern Recognition*, 145:109966, 2024. 2, 3, 6, 9, 10

[74] Kun-Hsing Yu, Ce Zhang, Gerald J Berry, Russ B Altman, Christopher Ré, Daniel L Rubin, and Michael Snyder. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nature Communications*, 7(1):12474, 2016. 2

[75] Yuansong Zeng, Zhuoyi Wei, Weijiang Yu, Rui Yin, Yuchen Yuan, Bingling Li, Zhonghui Tang, Yutong Lu, and Yuedong Yang. Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. *Briefings in Bioinformatics*, 23(5), 2022. 2, 3

[76] Daiwei Zhang, Amelia Schroeder, Hanying Yan, Haochen Yang, Jian Hu, Michelle YY Lee, Kyung S Cho, Katalin Susztak, George X Xu, and Michael D Feldman. Inferring super-resolution tissue architecture by integrating spatial transcriptomics with histology. *Nature Biotechnology*, 42(9):1372–1377, 2024. 2, 6, 9, 10

[77] Sichen Zhu, Yuchen Zhu, Molei Tao, and Peng Qiu. Diffusion generative modeling for spatially resolved gene expression inference from histology images, 2025. 2