HiFusion: Hierarchical Intra-Spot Alignment and Regional Context Fusion for Spatial Gene Expression Prediction from Histopathology

Ziqiao Weng[1,2*], Yaoyu Fang[1], Jiahe Qian[1,3], Xinkun Wang[1], Lee AD Cooper[1], Weidong Cai[2], Bo Zhou[1*]

[1]Northwestern University, Chicago, 60611, IL, USA.
[2]The University of Sydney, Sydney, NSW, 2006, Australia
[3]Chinese Academy of Sciences, Beijing, 100190, China.
alexziqiaoweng@gmail.com, bo.zhou@northwestern.edu

Abstract

Spatial transcriptomics (ST) bridges gene expression and tissue morphology but faces clinical adoption barriers due to technical complexity and prohibitive costs. While computational methods predict gene expression from H&E-stained whole-slide images (WSIs), existing approaches often fail to capture the intricate biological heterogeneity within spots and are susceptible to morphological noise when integrating contextual information from surrounding tissue. To overcome these limitations, we propose HiFusion, a novel deep learning framework that integrates two complementary components. First, we introduce the Hierarchical Intra-Spot Modeling module that extracts fine-grained morphological representations through multi-resolution sub-patch decomposition, guided by a feature alignment loss to ensure semantic consistency across scales. Concurrently, we present the Context-aware Cross-scale Fusion module, which employs cross-attention to selectively incorporate biologically relevant regional context, thereby enhancing representational capacity. This architecture enables comprehensive modeling of both cellular-level features and tissue microenvironmental cues, which are essential for accurate gene expression prediction. Extensive experiments on two benchmark ST datasets demonstrate that HiFusion achieves state-of-the-art performance across both 2D slide-wise cross-validation and more challenging 3D sample-specific scenarios. These results underscore HiFusion's potential as a robust, accurate, and scalable solution for ST inference from routine histopathology.

Code — https://github.com/Advanced-AI-in-Medicine-and-Physics-Lab/HiFusion

Extended version — http://arxiv.org/abs/2511.12969

Introduction

Spatial transcriptomics (ST) has emerged as a transformative technology that enables genome-wide gene expression profiling while preserving spatial localization within tissue sections, offering near-cellular resolution of molecular activity (Zhu et al. 2025; He et al. 2020; Pang, Su, and Li 2021). By integrating high-throughput RNA sequencing with spatial barcoding, ST maps transcriptomes to precise histological coordinates, thereby revealing spatial heterogeneity, tissue architecture, and cell–cell interactions across diverse biological systems (He et al. 2020; Pang, Su, and Li 2021). Most ST platforms divide tissue sections into discrete spots, typically 55–100 μm in diameter (Lin et al. 2024; Niu et al. 2025). These spatially barcoded spots collectively generate expression matrices that bridge molecular phenotypes with tissue morphology (Ståhl et al. 2016).

Despite its high-resolution potential, widespread adoption of ST remains hindered by practical constraints, including high experimental costs, specialized instrumentation, and limited scalability in clinical workflows (Zhu et al. 2025; Ruiz et al. 2025; Yang et al. 2023). Consequently, large-scale diagnostic or population-level applications remain rare.

In contrast, hematoxylin and eosin (H&E)-stained whole-slide images (WSIs) are routinely acquired in clinical pathology, are cost-effective, and encapsulate rich morphological features closely associated with gene expression patterns (Zhu et al. 2025; Niu et al. 2025). For example, overexpression of tumor markers such as *ERBB2* in HER2-positive breast cancer has been linked to distinct morphological phenotypes in H&E images (Chen et al. 2024). This has motivated a growing body of research on computational models, particularly deep learning-based approaches, to infer transcriptomic profiles directly from WSIs (He et al. 2020; Pang, Su, and Li 2021).

Recent advances in deep neural networks, including convolutional neural networks (CNNs), graph neural networks (GNNs), and transformer-based architectures, have enabled the prediction of spatially resolved gene expression at the spot level directly from WSI-derived image patches (Zhu et al. 2025; Xie et al. 2023). These models typically take spot-aligned image patches as input and aim to predict the expression levels of hundreds to thousands of genes by learning complex associations between tissue morphology and molecular profiles (Chen et al. 2024; Pang, Su, and Li 2021). For instance, ST-Net (He et al. 2020) employs a DenseNet backbone to generate spot-level predictions, while more recent approaches, such as HisToGene, Hist2ST, EGN, TRIPLEX, and ASIGN, enhance performance by incorporating spatial dependencies via long-range modeling, multi-resolution inputs, and inter-spot context integration (Pang, Su, and Li 2021; Zeng et al. 2022; Chung et al. 2024; Zhu et al. 2025).

Despite these advances, several key limitations remain.

*Corresponding author.

First, most existing methods struggle to capture both fine-grained morphological details and global tissue context simultaneously (Chen and Huang 2025; Chung et al. 2024). They typically treat each spot as homogeneous, overlooking the hierarchical structure within the spot. In reality, a single spot often contains diverse microstructures, such as distinct cell types, nuclear textures, and subcellular patterns, that are directly associated with gene expression (Pang, Su, and Li 2021; He et al. 2020). However, current architectures often fail to exploit these multi-scale intra-spot cues.

Moreover, broader contextual information is often used merely as auxiliary input, without explicitly modeling the semantic correlations between a spot and its surrounding tissue. This decoupled design limits the effective integration of region-aware signals, potentially leading to suboptimal representations for spatial gene prediction. Although recent models like TRIPLEX and ASIGN have adopted large regional patches (e.g., exceeding 1000×1000 pixels) to incorporate spatial context, the utility of such high-resolution inputs remains empirically underexplored. In fact, expanding the receptive field may introduce morphological noise or irrelevant signals, especially when adjacent regions lack biological relevance to the target spot.

To address these challenges, we propose ***HiFusion*** (Hierarchical Intra-Spot Alignment and Context-aware Fusion Network), a dual-branch framework for robust spatial gene expression prediction from histopathology images. HiFusion comprises two key components: Hierarchical Intra-Spot Modeling (HISM) and Context-Aware Cross-Scale Fusion (CCF).

The HISM module explicitly captures intra-spot morphological heterogeneity by decomposing each spot image into a hierarchy of non-overlapping sub-patches at multiple spatial resolutions, down to the cellular level. These multi-scale patches, along with the full-spot image, are processed through a shared encoder to extract features reflecting tissue-, cellular-, and subcellular-level structures. A feature alignment loss ensures semantic consistency across scales, encouraging coherent multi-scale representations.

The CCF module incorporates broader tissue context by encoding neighboring regions with a lightweight encoder. Contextual features act as queries in a cross-attention module, while the adaptively fused multi-scale spot representations from HISM serve as keys and values. This design allows the model to selectively attend to biologically relevant contextual information while suppressing spatial noise, thereby enhancing the robustness and expressiveness of the learned features.

Together, these two components enable HiFusion to jointly model fine-grained intra-spot morphology and spatial context, overcoming limitations of coarse spot-level representations and simplistic context integration. Our key contributions can be summarized as follows:

- We propose HiFusion, a novel framework for spatial gene expression prediction from whole-slide images. HiFusion explicitly integrates multi-scale intra-spot representations with regional tissue context, effectively capturing spatial and biological heterogeneity across scales.
- Our approach introduces a hierarchical intra-spot modeling module that extracts rich, fine-grained features from multiple spatial resolutions, coupled with a feature alignment loss to ensure semantic consistency across scales. A context-aware cross-scale fusion module further integrates these intra-spot features with neighboring regional context via a residual cross-attention mechanism, enhancing representational expressiveness and robustness.
- Extensive evaluation on two public ST datasets demonstrates that HiFusion consistently outperforms state-of-the-art methods under both conventional 2D slide-wise cross-validation and a recent 3D sample-specific evaluation protocol, establishing new benchmarks for spatial gene expression inference. Comprehensive ablation studies further validate the effectiveness of each module and analyze the impact of spatial context size on prediction performance.

Related Works

Spatial transcriptomics captures spatially resolved mRNA using microarray chips, followed by next-generation sequencing and spatial mapping onto histological images to generate high-resolution gene expression landscapes (Zhang et al. 2022; Niu et al. 2025). Recent efforts have shifted toward learning-based approaches that infer spatial gene expression directly from H&E-stained WSIs, formulating it as a multi-output regression task over spot-aligned image patches.

ST-Net (He et al. 2020) initiated this direction by mapping spot-level patches to gene expression using a DenseNet-121 backbone, treating each spot independently and neglecting contextual cues. HisToGene (Pang, Su, and Li 2021) incorporated long-range dependencies via Vision Transformers, while Hist2ST (Zeng et al. 2022) added local feature extraction (ConvMixer) and neighborhood modeling with Graph Neural Networks. Image similarity–based models like EGN (Yang et al. 2023) and BLEEP (Xie et al. 2023) retrieve exemplar patches or learn contrastive embeddings, but are sensitive to staining variations and generalize poorly across samples. To integrate broader context, TRIPLEX (Chung et al. 2024) extracts features from the spot, its neighborhood, and the full slide using a three-branch architecture, while ASIGN (Zhu et al. 2025) aligns adjacent tissue sections in 3D with a graph-based model. However, both methods rely on multi-resolution inputs and complex alignment pipelines. In contrast, our proposed HiFusion explicitly captures intra-spot hierarchical structure via multiscale patch decomposition and enforces cross-scale consistency through feature alignment. It further enhances context integration by fusing region-level features with fine-grained spot representations via cross-attention, offering a more efficient and generalizable framework for spatial gene expression prediction.

Method

Problem Formulation

We formulate spatial gene expression prediction as a multi-output regression task over a set of spatially arranged spots on a whole-slide image (WSI). Let $X^S \in \mathbb{R}^{n \times H_S \times W_S \times 3}$

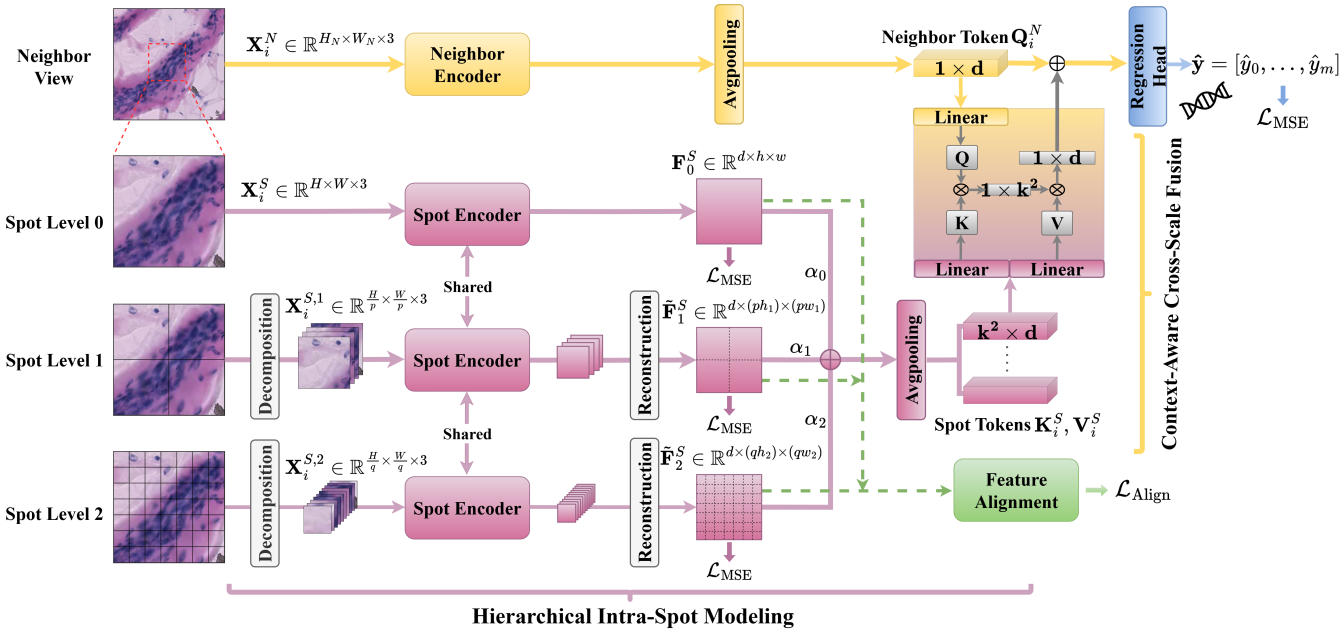


Figure 1: Schematic of the proposed **HiFusion** framework, which integrates *Hierarchical Intra-Spot Modeling (HISM)* and *Context-Aware Cross-Scale Fusion (CCF)*. HISM hierarchically decomposes each spot into multi-scale patches to extract fine-grained features with semantic alignment. CCF fuses contextual region features with multi-scale spot representations via residual cross-attention for gene expression prediction.

denote the collection of cropped spot-level image patches, where n is the number of spots, and (H_S, W_S) denotes the height and width of each spot image. The corresponding normalized gene expression profiles are represented as $Y \in \mathbb{R}^{n\times m}$, where m is the number of genes to be predicted. To incorporate local spatial context, we additionally extract regional neighbor patches for each spot, denoted as $X^N \in \mathbb{R}^{n\times H_N\times W_N\times 3}$, where (H_N, W_N) denotes the size of each neighbor patch. Our objective is to learn a predictive function $\phi : \{X^S, X^N\} \to Y$ that maps each spot and its surrounding context to its corresponding gene expression vector.

Overview of HiFusion

The overall workflow of HiFusion is shown in Figure 1, comprising two components: Hierarchical Intra-Spot Modeling (HISM) and Context-Aware Cross-Scale Fusion (CCF). In HISM, each spot image is decomposed into non-overlapping sub-patches at multiple resolutions, including the cellular scale, to capture fine-grained morphological cues such as nuclear structure and cell-type variation. The full spot and its sub-patches are processed by a shared encoder to extract multi-scale features. A feature alignment loss encourages semantic consistency across scales, leveraging the translation invariance of CNNs. In CCF, multi-scale spot features are adaptively fused to form the keys and values, while a region-level feature extracted by a lightweight encoder serves as the query. These are integrated via residual multi-head cross-attention to produce the final representation for gene expression prediction.

Hierarchical Intra-Spot Modeling and Alignment

To capture the rich intra-spot morphological heterogeneity, we propose a hierarchical modeling strategy that encodes visual patterns from tissue- to subcellular-level resolution. Given a spot image $\mathbf{X}_i^S \in \mathbb{R}^{H\times W\times 3}$, we define this as the Level-0 input. A shared encoder $f_\theta(\cdot)$ extracts a global feature map $\mathbf{F}_0^S = f_\theta(\mathbf{X}_i^S) \in \mathbb{R}^{d\times h\times w}$, which captures coarse tissue-level context.

To obtain finer-scale features, we decompose the input into $p \times p$ and $q \times q$ non-overlapping patches ($q > p$), forming Level-1 and Level-2 inputs $\{\mathbf{X}_{i,j}^{S,1}\}_{j=1}^{p^2}$ and $\{\mathbf{X}_{i,k}^{S,2}\}_{k=1}^{q^2}$, where $\mathbf{X}_{i,j}^{S,1} \in \mathbb{R}^{\frac{H}{p}\times\frac{W}{p}\times 3}$ and $\mathbf{X}_{i,k}^{S,2} \in \mathbb{R}^{\frac{H}{q}\times\frac{W}{q}\times 3}$, respectively. These patches are passed through the same encoder to yield multi-scale representations: $\mathbf{F}_1^S \in \mathbb{R}^{p^2\times d\times h_1\times w_1}$ and $\mathbf{F}_2^S \in \mathbb{R}^{q^2\times d\times h_2\times w_2}$.

We then reconstruct the spatial layout of patch features based on their original positions, resulting in $\tilde{\mathbf{F}}_1^S \in \mathbb{R}^{d\times(ph_1)\times(pw_1)}$ and $\tilde{\mathbf{F}}_2^S \in \mathbb{R}^{d\times(qh_2)\times(qw_2)}$. If the reconstructed resolutions do not match that of $\mathbf{F}_0^S$, bilinear interpolation is applied to align them accordingly.

To enforce cross-scale semantic consistency, we define a feature alignment loss that encourages the fine-scale features to preserve the global semantics of the full spot representation:

$$\mathcal{L}_{\text{align}} = \sum_{s=1}^{2} \left\| \tilde{\mathbf{F}}_s^S - \mathbf{F}_0^S \right\|_1 \tag{1}$$

This hierarchical design enables the model to learn

both coarse and fine-grained morphological representations within each spot. By explicitly aligning multi-scale features at the pixel level, the network is encouraged to maintain semantic consistency across spatial resolutions, thereby enhancing the robustness of the learned representations.

Context-Aware Cross-Scale Fusion

To incorporate broader tissue context while preserving intra-spot structural fidelity, we introduce a cross-scale fusion module guided by region-level information. For each spot $\mathbf{X}_i^S$, we extract a surrounding tissue region $\mathbf{X}_i^N \in \mathbb{R}^{H_N \times W_N \times 3}$, which is encoded by a lightweight encoder $f_\psi(\cdot)$ followed by global average pooling, yielding a condensed regional representation $\mathbf{Q}_i^N \in \mathbb{R}^{1\times d}$:

$$\mathbf{Q}_i^N = \text{AvgPool}(f_\psi(\mathbf{X}_i^N)) \tag{2}$$

To integrate multi-scale intra-spot features $\{\mathbf{F}_s^S\}_{s=0}^2$ derived from spot image decomposition at three scales ($s = 0, 1, 2$), we adopt a learnable weighted fusion strategy. The fused spot representation is computed as:

$$\mathbf{F}_{\text{fused}}^S = \sum_{s=0}^{2} \omega_s \cdot \mathbf{F}_s^S, \tag{3}$$

where the weight ω_s for each scale is generated via a softmax over learnable parameters $\{\alpha_s\}_{s=0}^2$:

$$\omega_s = \frac{\exp(\alpha_s)}{\sum_{j=0}^{2}\exp(\alpha_j)}, \quad s \in \{0,1,2\} \tag{4}$$

To prepare the fused intra-spot features for cross-attention, we apply adaptive average pooling to obtain $\bar{\mathbf{F}}^S = \text{AvgPool}(\mathbf{F}_{\text{fused}}^S, (k,k)) \in \mathbb{R}^{d\times k\times k}$, which is then reshaped into key and value matrices:

$$\mathbf{K}_i^S = \mathbf{V}_i^S = \text{reshape}(\bar{\mathbf{F}}^S) \in \mathbb{R}^{k^2\times d} \tag{5}$$

We then integrate the region-level and multi-scale intra-spot features through a residual cross-attention mechanism. Specifically, the region token $\mathbf{Q}^N$ is used as the query, while the fused intra-spot representation serves as the keys and values. The cross-attention output is computed as:

$$\phi_{\text{ca}}(\cdot) = \text{softmax}\left(\frac{(\mathbf{Q}_i^N\mathbf{W}_Q)(\mathbf{K}_i^S\mathbf{W}_K)^\top}{\sqrt{d_K}}\right)(\mathbf{V}_i^S\mathbf{W}_V), \tag{6}$$

where $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ are learnable linear projections, and d_K is the dimensionality of the key vectors used for scaling. Here we present the single-head formulation; our implementation uses the standard multi-head attention.

Finally, we predict the gene expression vector $\hat{\mathbf{y}}_i \in \mathbb{R}^m$ using the fused output and a residual connection, followed by a prediction head composed of LayerNorm and a fully connected layer:

$$\hat{\mathbf{y}}_i = \text{FC}(\text{LayerNorm}(\mathbf{Q}_i^N + \phi_{\text{ca}}(\mathbf{Q}_i^N, \mathbf{K}_i^S, \mathbf{V}_i^S))) \tag{7}$$

This context-aware fusion strategy allows the model to selectively attend to semantically relevant intra-spot features while mitigating high-frequency noise introduced by overly fine-grained morphological details.

Loss Function

To enhance context-aware gene expression prediction, we employ a composite loss combining regression objectives with feature alignment.

Let $\hat{\mathbf{y}}_i, \mathbf{y}_i \in \mathbb{R}^m$ denote the predicted and ground-truth expression for the i-th spot. The primary prediction loss is:

$$\mathcal{L}_{\text{main}} = \frac{1}{n}\sum_{i=1}^{n}\|\hat{\mathbf{y}}_i - \mathbf{y}_i\|_2^2 \tag{8}$$

For multi-scale supervision, each level's features ($\mathbf{F}_0^S$, $\tilde{\mathbf{F}}_1^S$, $\tilde{\mathbf{F}}_2^S$) generate auxiliary predictions $\hat{\mathbf{y}}_i^{(s)}$ via shared FC layers:

$$\mathcal{L}_{\text{aux}} = \frac{1}{3n}\sum_{i=1}^{n}\sum_{s=0}^{2}\|\hat{\mathbf{y}}_i^{(s)} - \mathbf{y}_i\|_2^2 \tag{9}$$

The total training objective combines regression and alignment losses:

$$\mathcal{L}_{\text{total}} = \underbrace{\mathcal{L}_{\text{main}} + \mathcal{L}_{\text{aux}}}_{\mathcal{L}_{\text{reg}}} + \lambda\mathcal{L}_{\text{align}}, \tag{10}$$

where λ balances the alignment regularization. This multi-level supervision strategy ensures that the network captures both global and fine-grained morphological patterns relevant to gene expression, while the alignment term enforces semantic consistency across scales to improve training stability and generalization.

Experiments and Results

Dataset and Pre-processing

In our experiments, we utilized two publicly available datasets used in (Zhu et al. 2025) to evaluate the performance of the proposed *HiFusion* model. The first is the HER2-positive breast tumor dataset (Andersson et al. 2021), denoted as HER2, which includes 36 whole-slide images comprising 4 samples with 6-layer tissue sections and 4 samples with 3-layer sections, totaling 13,620 ST spots. The second dataset, referred to as ST-Data, is a breast cancer dataset introduced in ST-Net (He et al. 2020). Following the protocol in (Zhu et al. 2025), we selected 16 samples with three-layer tissue sections, yielding 41,544 spots in total. In both datasets, the spots have a diameter of 100 µm and are arranged on a grid with a center-to-center distance of 200 µm.

Given the high dimensionality of gene expression data (exceeding 15,000 genes), it is impractical to predict the expression of all genes from histological patches. Hence, following (He et al. 2020), we select the top 250 genes with the highest average expression levels for prediction. The selected genes are listed in the supplementary material. To normalize gene expression values, we first perform a spot-wise normalization by dividing each gene count x_i by the total expression count across all genes within the same spot (after adding 1 to avoid division by zero), and then apply a logarithmic transformation $x^{\text{norm}} = \log\left((x+1)/\sum_{i=1}^{m}(x_i+1)\right)$, where m denotes the number of genes and all operations are applied element-wise.

Method	HER2						ST-Data					
	2D slide-wise			3D sample-specific			2D slide-wise			3D sample-specific		
	MSE	MAE	PCC	MSE	MAE	PCC	MSE	MAE	PCC	MSE	MAE	PCC
ST-Net	0.6523	0.6255	0.4621	0.5323	0.5747	0.7042	0.5798	0.5943	0.5304	0.4939	0.5514	0.7443
HisToGene	0.6105	0.6063	0.4294	0.4851	0.4899	0.7028	0.5310	0.5694	0.5427	0.4841	0.5121	0.7280
His2ST	0.5843	0.5885	0.4478	0.3174	0.4438	0.7276	0.5230	0.5636	0.5442	0.2877	0.4240	0.7725
EGN	0.5845	0.5940	0.4723	0.2917	0.4258	0.7441	0.5568	0.5800	0.5103	0.2755	0.4136	0.7800
TRIPLEX	0.5715	0.5918	0.4750	0.2899	0.4268	0.7471	0.5389	0.5769	0.5387	0.2857	0.4255	0.7780
ASIGN-2D	0.5830	0.5901	0.4601	0.3116	0.4415	0.7316	0.5449	0.5764	0.5373	0.2822	0.4204	0.7741
ASIGN-3D	\	\	\	0.4163	0.4987	0.7019	\	\	\	0.3141	0.4445	0.7524
HiFusion* (Ours)	**0.5459**	**0.5699**	**0.4961**	**0.2846**	**0.4205**	**0.7492**	**0.5095**	**0.5557**	**0.5613**	**0.2711**	**0.4102**	**0.7838**

Table 1: Comparison of MSE, MAE, and PCC on HER2 and ST-Data under two evaluation settings. Best and second-best results are shown in bold and underlined, respectively. * denotes significant improvement over the second-best baseline ($p < 0.05$).

Experiment Setup and Evaluation Metrics

We evaluate our method under two distinct testing scenarios to comprehensively assess its performance.

The first, 2D slide-wise cross-validation, follows the protocol commonly adopted in prior studies. We conduct 4-fold cross-validation on both datasets, ensuring that samples from the same patient are assigned exclusively to either the training or test set to avoid data leakage.

The second, 3D sample-specific validation, represents a more challenging and recently proposed paradigm (Qian et al. 2025; Fang et al. 2025), where training and testing are performed within individual patients. Specifically, the first histological layer from each patient is used for training, while the remaining sections are reserved for testing. This setup emphasizes within-sample generalization and minimizes inter-patient domain shifts.

Model performance is evaluated using three metrics: Mean Squared Error (MSE), which reflects the average squared deviation between predictions and ground truth; Mean Absolute Error (MAE), capturing the average magnitude of errors; and the Pearson Correlation Coefficient (PCC), which quantifies the linear correlation between predicted and actual gene expression values. Lower MSE and MAE indicate higher predictive accuracy, while higher PCC suggests stronger consistency with ground-truth profiles.

Implementation Details

For all datasets, each spot image (Level-0 input) is cropped to 224×224 pixels, corresponding to approximately 150 μm $\times$ 150 μm in the original pathology image, using the center coordinates of each spot. To model intra-spot spatial hierarchy, each spot image is decomposed into 2×2 and 7×7 non-overlapping patches (Level-1 and Level-2 inputs, respectively), where $p = 2$ and $q = 7$. To incorporate regional context, neighboring patches are extracted by cropping a 448×448 image centered at each spot. We adopt ResNet-18 as the backbone encoder for spot-level feature extraction and ResNet-10 as a lightweight encoder for neighboring regions. The number of key and value tokens after adaptive average pooling is set to $k \times k = 2 \times 2$. The loss weight λ is empirically set to 1. The model is optimized using Adam with a momentum of 0.9 and a weight decay of 10^{-5}. The initial learning rate is set to 3×10^{-4} and adjusted dynamically using a cosine annealing scheduler with a minimum learning rate (η_{min}) of 1×10^{-6}. Training is conducted for 50 epochs with a batch size of 32. The reported results are the average performance across all patient samples. All experiments are performed on a single NVIDIA RTX 4090 GPU.

Baselines

We benchmarked the performance of our model against several representative baseline methods, including: 1) local-based models (ST-Net (He et al. 2020), EGN (Yang et al. 2023)) and 2) global-based models (TRIPLEX (Chung et al. 2024), ASIGN (Zhu et al. 2025)). Notably, ASIGN represents the current state-of-the-art (SOTA) approach. We reproduced ST-Net and EGN following the implementations described in the TRIPLEX and ASIGN papers. Specifically, for EGN, we employed ResNet-18 as the feature extractor, while ST-Net utilized a pretrained DenseNet-121. For TRIPLEX and ASIGN, we retained the original network architectures and hyperparameter settings as reported in their respective papers. All baseline models were trained and evaluated under identical conditions for fair comparison.

Comparison between HiFusion and Baselines

Table 1 demonstrates the superior performance of HiFusion across both HER2 and ST-Data datasets under 2D slide-wise and 3D sample-specific evaluations. In slide-wise testing, HiFusion achieved MSE/MAE/PCC scores of 0.5459/0.5699/0.4961 on HER2, outperforming TRIPLEX (second-best) by 2.1–2.6% and ASIGN (SOTA) by 2.0–3.7%, with over 10% MSE improvement versus ST-Net. Similar superiority was observed on ST-Data, confirming HiFusion's robust cross-patient generalization capability for heterogeneous spatial transcriptomics. The 3D sample-specific evaluation revealed consistent advantages, particularly showing 22–25% improvement over ST-Net. While less pronounced than 2D slide-wise gains, these results validate HiFusion's effectiveness in modeling intra-patient spatial structures. The consistent superiority across both evaluation paradigms highlights HiFusion's versatility in addressing distinct challenges in ST analysis.

Level Combination	MSE ↓	MAE ↓	PCC ↑
1×1	0.5718	0.5840	0.4763
1×1 + 2×2	0.5561	0.5774	0.4902
1×1 + 4×4	0.5538	0.5769	0.4854
1×1 + 7×7	0.5541	0.5725	0.4777
1×1 + 2×2 + 4×4	0.5478	0.5732	0.4935
1×1 + 2×2 + 7×7	**0.5459**	**0.5699**	**0.4961**
1×1 + 4×4 + 7×7	0.5566	0.5765	0.4756
1×1 + 2×2 + 4×4 + 7×7	0.5477	0.5719	0.4871

Table 2: Ablation study for image decomposition levels

Method	HER2			ST		
	MSE	MAE	PCC	MSE	MAE	PCC
Our	**0.5459**	**0.5699**	**0.4961**	**0.5095**	**0.5557**	**0.5613**
w/o FA	0.5642	0.5798	0.4730	0.5150	0.5594	0.5559

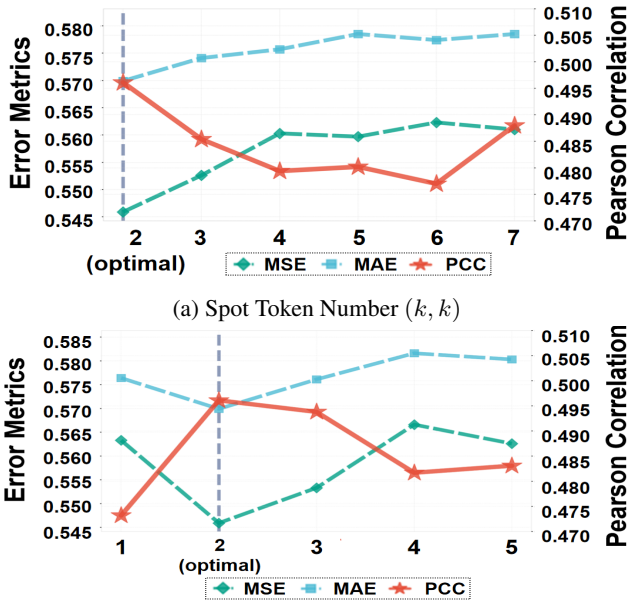
Table 3: Ablation study for feature alignment

Comparative Analysis of 3D Prediction Strategies

ASIGN-3D was designed for 3D sample-wise prediction, using known-label layers to improve accuracy. However, this sophisticated strategy underperformed compared to the simpler intra-sample learning paradigm, which trains solely on a single labeled section and directly predicts gene expression in adjacent layers. We attribute this unexpected outcome to two key factors: (1) substantial inter-patient variability in histopathology and gene expression, which introduces noise during multi-sample training; and (2) potential feature degradation caused by global 3D registeration during preprocessing, which may distort biologically meaningful tissue structures. These results demonstrate the advantages of 3D intra-sample learning, offering improved accuracy through patient-specific patterns while reducing computational costs. This performance-efficiency balance is promising for clinical spatial transcriptomics applications. Additional 3D-2D comparisons are available in the supplementary material.

Ablation Study on HISM

We evaluate the impact of different image decomposition level combinations within our HISM framework on the HER2 dataset. Given the original spot image size of 224×224, it can be decomposed into patch grids of 2×2, 4×4, and 7×7. Here, 1×1 refers to using the original image without decomposition. As shown in Table 2, incorporating any level of decomposition consistently improves performance compared to using only the original image. Even without decomposition, the result is already comparable to the second-best baseline (e.g., TRIPLEX in Table 1), highlighting the strength of our base framework. Among all configurations, the combination of $1 \times 1 + 2 \times 2 + 7 \times 7$ achieves the best overall performance. We attribute this to the complementary spatial granularity captured at different scales: the 1×1 input preserves global tissue-level context, the 2×2 patches capture sub-tissue or regional structures, and the fine-grained 7×7 decomposition provides detailed cellular or subcellular information. This multi-scale representation supports more accurate and biologically meaningful gene expression prediction.



(a) Spot Token Number (k, k)

(b) Neighbor Patch Size $(224 \times N, 224 \times N)$

Figure 2: Ablation study for (a) spot token number and (b) neighbor patch size.

We further investigate the role of feature alignment in HISM on both the HER2 and ST-Data datasets. As shown in Table 3, incorporating the feature alignment loss consistently improves performance across both datasets. On HER2, it reduces MSE by nearly 2% and increases PCC by over 2%. These results demonstrate that the feature alignment loss effectively enforces cross-scale semantic and predictive consistency, synergizing with hierarchical image decomposition to enhance model performance.

Ablation Study on CCF

We conduct two experiments on the HER2 dataset to evaluate the effect of **(a)** spot token number and **(b)** neighbor image size in the CCF module, as shown in Figure 2.

In Figure 2(a), we vary the number of tokens by applying adaptive average pooling to the fused multi-scale spot features, using grid sizes from 2×2 to 7×7 (7×7 corresponds to no pooling). Overall, MSE and MAE tend to increase and PCC slightly declines as token count grows. The best performance is observed with the 2×2 configuration. We hypothesize that fewer tokens help suppress spatial noise while preserving key contextual signals, and are better suited to the capacity of the cross-attention module. In contrast, longer sequences may fragment attention, introduce redundancy, or exceed the module's effective modeling range, leading to performance degradation.

As discussed in the Introduction, prior methods such as TRIPLEX and ASIGN use large patches (e.g., 1120×1120) to incorporate global context, though their utility remains empirically unclear. To investigate this, Figure 2(b) shows

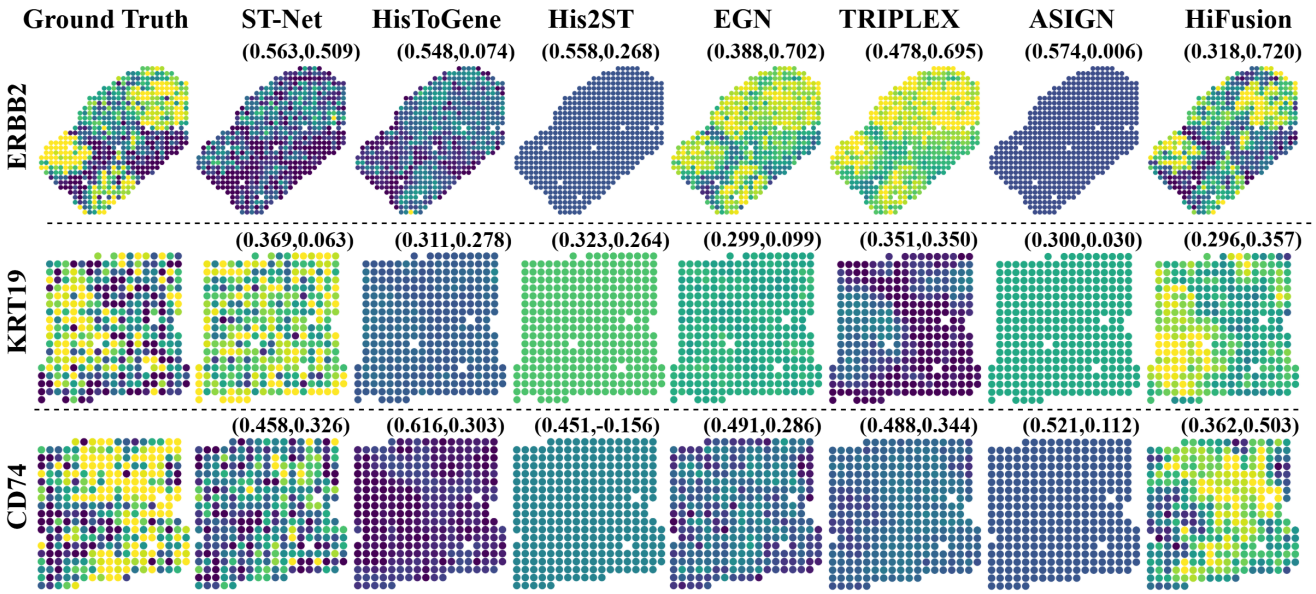


Figure 3: Predicted spatial expression of ERBB2, KRT19 and CD74 by different models. (MAE, PCC) values with the ground truth are shown. Brighter regions indicate higher gene expression levels, while darker regions represent lower expression. HiFusion achieves the best visual and quantitative alignment.

results using different neighbor image sizes. The optimal performance occurs when the neighborhood is twice the spot size ($N = 2$), while larger contexts lead to degradation. This suggests that moderate expansion captures informative spatial cues, whereas excessive enlargement introduces irrelevant tissue regions, increasing noise and diminishing neighbor query quality. Further ablation results on ST-Data are in the supplementary material.

Visualization of Cancer Marker Genes

To evaluate the clinical applicability of our model, we examined three well-established cancer marker genes with significant relevance to HER2-positive breast cancer: *ERBB2* (HER2) (Mehta and Tripathy 2014; Dent et al. 2013; Conley et al. 2016), *KRT19* (Saha et al. 2017; Tang et al. 2014), and *CD74* (Su et al. 2017; Borghese and Clanchy 2011). Cross-validation on the HER2 dataset shows that HiFusion consistently outperforms existing methods in predicting spatial gene expression. For *ERBB2*, HiFusion achieved an MAE of 0.711 (PCC = 0.518), substantially outperforming ASIGN (1.074, PCC = -0.035), TRIPLEX (0.900, 0.486), EGN (0.778, 0.401), His2ST (1.265, -0.029), HisToGene (1.072, 0.267), and ST-Net (1.127, 0.378). Similar performance gains were observed for KRT19, where HiFusion obtained an MAE of 0.446 and PCC of 0.230, surpassing the best competing MAE of 0.450 and PCC of 0.201. For CD74, HiFusion again led with an MAE of 0.584 and PCC of 0.357, outperforming the best alternatives (MAE = 0.594, PCC = 0.253). Detailed results for these marker genes are provided in the supplementary material. These results collectively highlight HiFusion's robust and consistent superiority across diverse gene targets.

Figure 3 illustrates the predicted spatial distributions of ERBB2, KRT19, and CD74 in three WSI samples, along with corresponding MAE and PCC scores for each model. Notably, HiFusion not only achieves the highest quantitative agreement with the ground truth, but also provides visually accurate localization of high-expression regions (highlighted by brighter colors), demonstrating its robustness in capturing complex spatial gene expression patterns.

Discussion and Conclusion

We propose HiFusion, a novel framework for spatial gene expression prediction from whole-slide images. To overcome the limitations of coarse spot-level modeling and insufficient contextual integration, HiFusion combines hierarchical intra-spot modeling with context-aware cross-scale fusion, effectively capturing multiscale spatial and biological heterogeneity. The HISM module extracts fine-grained features across multiple resolutions, reinforced by a feature alignment loss to ensure semantic consistency. The CCF module further enhances representation through residual cross-attention, dynamically integrating intra-spot features with neighboring regional context. Extensive experiments demonstrate that HiFusion consistently outperforms competing methods and generalizes robustly across both 2D slide-wise cross-validation and 3D sample-specific evaluation. Importantly, we found that the 3D learning paradigm achieves strong intra-patient generalization at minimal computational and labeling cost, underscoring its practical potential for clinical spatial transcriptomics. While HiFusion effectively incorporates regional context via a single-branch design, future work may explore more expressive and efficient mechanisms to extract biologically-relevant fine-grained features from tissue regions and integrate them more effectively with intra-spot representations.

Acknowledgments

Ziqiao Weng was supported by Australian Government Research Training Program (RTP) scholarship.

References

Andersson, A.; Larsson, L.; Stenbeck, L.; Salmén, F.; Ehinger, A.; Wu, S. Z.; Al-Eryani, G.; Roden, D.; Swarbrick, A.; Borg, Å.; et al. 2021. Spatial deconvolution of HER2-positive breast cancer delineates tumor-associated cell type interactions. *Nature Communications*, 12(1): 6012.

Borghese, F.; and Clanchy, F. I. 2011. CD74: an emerging opportunity as a therapeutic target in cancer and autoimmune disease. *Expert Opinion on Therapeutic Targets*, 15(3): 237–251.

Chen, J.; Zhou, M.; Wu, W.; Zhang, J.; Li, Y.; and Li, D. 2024. STimage-1K4M: A histopathology image-gene expression dataset for spatial transcriptomics. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 35796–35823. Curran Associates, Inc.

Chen, X.; and Huang, J. 2025. DELST: Dual Entailment Learning for Hyperbolic Image-Gene Pretraining in Spatial Transcriptomics. arXiv:2503.00804.

Chung, Y.; Ha, J. H.; Im, K. C.; and Lee, J. S. 2024. Accurate spatial gene expression prediction by integrating multi-resolution features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11591–11600.

Conley, S.; Bosco, E.; Tice, D.; Hollingsworth, R.; Herbst, R.; and Xiao, Z. 2016. HER2 drives Mucin-like 1 to control proliferation in breast cancer cells. *Oncogene*, 35(32): 4225–4234.

Dent, S.; Oyan, B.; Honig, A.; Mano, M.; and Howell, S. 2013. HER2-targeted therapy in breast cancer: a systematic review of neoadjuvant trials. *Cancer Treatment Reviews*, 39(6): 622–631.

Fang, Y.; Qian, J.; Wang, X.; Cooper, L. A.; and Zhou, B. 2025. Sparser2Sparse: Single-shot Sparser-to-Sparse Learning for Spatial Transcriptomics Imputation with Natural Image Co-learning. arXiv:2507.16886.

He, B.; Bergenstråhle, L.; Stenbeck, L.; Abid, A.; Andersson, A.; Borg, Å.; Maaskola, J.; Lundeberg, J.; and Zou, J. 2020. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature Biomedical Engineering*, 4(8): 827–834.

Lin, Y.; Luo, L.; Chen, Y.; et al. 2024. ST-Align: A Multimodal Foundation Model for Image-Gene Alignment in Spatial Transcriptomics. arXiv:2411.16793.

Mehta, A.; and Tripathy, D. 2014. Co-targeting estrogen receptor and HER2 pathways in breast cancer. *The Breast*, 23(1): 2–9.

Niu, Y.; Liu, J.; Zhan, Y.; Shi, J.; Zhang, D.; Reinius, M.; Machado, I.; Crispin-Ortuzar, M.; Wu, J.; Li, C.; et al. 2025. PH2ST: ST-Prompt Guided Histological Hypergraph Learning for Spatial Gene Expression Prediction. arXiv:2503.16816.

Pang, M.; Su, K.; and Li, M. 2021. Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. *BioRxiv*, 2021–11.

Qian, J.; Fang, Y.; Wang, X.; Cooper, L. A.; and Zhou, B. 2025. ST-DAI: Single-shot 2.5D Spatial Transcriptomics with Intra-Sample Domain Adaptive Imputation for Cost-efficient 3D Reconstruction. arXiv:2507.21516.

Ruiz, D.; Cardenas, P.; Manrique, L.; Vega, D.; Mejia, G.; and Arbelaez, P. 2025. Completing Spatial Transcriptomics Data for Gene Expression Prediction Benchmarking. arXiv:2505.02980.

Saha, S.; Choi, H.; Kim, B.; Dayem, A.; Yang, G.; Kim, K.; Yin, Y.; and Cho, S. 2017. KRT19 directly interacts with β-catenin/RAC1 complex to regulate NUMB-dependent NOTCH signaling pathway and breast cancer properties. *Oncogene*, 36(3): 332–349.

Ståhl, P. L.; Salmén, F.; Vickovic, S.; Lundmark, A.; Navarro, J. F.; Magnusson, J.; Giacomello, S.; Asp, M.; Westholm, J. O.; Huss, M.; et al. 2016. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294): 78–82.

Su, H.; Na, N.; Zhang, X.; and Zhao, Y. 2017. The biological function and significance of CD74 in immune diseases. *Inflammation Research*, 66(3): 209–216.

Tang, J.; Zhuo, H.; Zhang, X.; Jiang, R.; Ji, J.; Deng, L.; Qian, X.; Zhang, F.; and Sun, B. 2014. A novel biomarker Linc00974 interacting with KRT19 promotes proliferation and metastasis in hepatocellular carcinoma. *Cell Death & Disease*, 5(12): e1549–e1549.

Xie, R.; Pang, K.; Chung, S.; Perciani, C.; MacParland, S.; Wang, B.; and Bader, G. 2023. Spatially resolved gene expression prediction from histology images via bi-modal contrastive learning. *Advances in Neural Information Processing Systems*, 36: 70626–70637.

Yang, Y.; Hossain, M. Z.; Stone, E. A.; and Rahman, S. 2023. Exemplar guided deep neural network for spatial transcriptomics analysis of gene expression prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5039–5048.

Zeng, Y.; Wei, Z.; Yu, W.; Yin, R.; Yuan, Y.; Li, B.; Tang, Z.; Lu, Y.; and Yang, Y. 2022. Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. *Briefings in Bioinformatics*, 23(5): bbac297.

Zhang, L.; Chen, D.; Song, D.; Liu, X.; Zhang, Y.; Xu, X.; and Wang, X. 2022. Clinical and translational values of spatial transcriptomics. *Signal Transduction and Targeted Therapy*, 7(1): 111.

Zhu, J.; Deng, R.; Yao, T.; Xiong, J.; Qu, C.; Guo, J.; Lu, S.; Yin, M.; Wang, Y.; Zhao, S.; et al. 2025. ASIGN: an anatomy-aware spatial imputation graphic network for 3D spatial transcriptomics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 30829–30838.