SPATIAL TRANSCRIPTOMICS SUPER-RESOLUTION VIA DUAL-ADAPTIVE MULTIMODAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Conventional sequencing-based spatial transcriptomics offers broad gene coverage but aggregates signals from multiple cells, motivating histology-guided super-resolution reconstruction of finer spatial expression. Existing methods often use fixed context combinations or uniform-capacity prediction branches, limiting explicit adaptation of scale contributions and mapping capacity to local tissue and molecular information. To address this limitation, we introduce Dual-Adaptive Spatial Transcriptomics (DAST), which jointly adapts multiscale context integration and expression mapping to predict superpixel-level gene expression under spot-aggregated supervision. Specifically, DAST constructs multiscale contexts after pathology feature encoding, interacts each scale with local expression features, and dynamically learns scale contributions from the resulting multimodal representations. The resulting fused content and scale weights then jointly guide selection among experts with different hidden widths, adapting mapping capacity to the context composition. Across ten Visium HD and Xenium datasets, DAST achieves the lowest RMSE and highest SSIM compared with six other methods. Factorial comparisons and ablations provide evidence of synergistic gains from dynamic scale fusion and heterogeneous expert mapping. Downstream analyses further explore spatial domain identification and candidate tertiary lymphoid structure localization.

1 INTRODUCTION

Spatial transcriptomics measures gene expression while preserving spatial information, enabling the study of tissue heterogeneity, cellular microenvironments, and disease-associated spatial structures (Moses & Pachter, 2022). However, conventional sequencing-based platforms offer broad transcriptomic coverage but often aggregate signals from multiple cells (Ståhl et al., 2016), whereas imaging-based platforms provide finer localization but are constrained by predefined gene panels and the experimental cost of repeated labeling and imaging (Chen et al., 2015; Vu et al., 2022). Reconstructing finer-grained gene-expression distributions from existing low-resolution spatial expression is therefore valuable for localizing molecular signals more precisely and extending analyses of tissue structure and microenvironments.

We study super-resolution reconstruction of superpixel-level gene expression by combining low-resolution spatial expression with high-resolution H&E images. Cellular morphology, tissue boundaries, and spatial arrangements visible in H&E provide cues for distinguishing tissue components mixed within spots. Previous studies have used these cues to infer fine-grained expression without direct high-resolution reference-expression supervision (Bergenstråhle et al., 2022; Hu et al., 2023; Zhang et al., 2024). Yet different within-spot expression distributions can correspond to the same observed spot expression, so reconstruction quality still depends on how tissue morphology and local molecular signals are combined to constrain expression within spots.

Using this information effectively requires tissue context at spatial scales appropriate to the local environment. Local views capture cellular morphology and fine boundaries, whereas broader views reveal cellular arrangements, tissue compartments, and the surrounding environment. A field of view that is too small may omit useful structures, while indiscriminately expanding it may introduce signals from other tissue regions, motivating scale contributions that adapt to local tissue and molecular information. Prior work has expanded context through hierarchical feature extraction and neighbor-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

hood modeling, but obtaining multiscale features does not establish their relative value for gene-expression prediction. Representations formed during image-only encoding do not yet incorporate local expression information, while predetermined scale combinations lack explicit adaptation to local context. Incorporating molecular observations and dynamically learning scale contributions after encoding can therefore produce context combinations better suited to super-resolution reconstruction.

Changes in context composition also affect the feature relationships needed for prediction, making the benefit of multiscale integration dependent on how the resulting representations are mapped to expression. Existing approaches commonly use a shared nonlinear predictor, applying the same mapping capacity across feature combinations without an explicit mechanism for adapting that capacity to the fused content. Experts with different hidden widths offer alternative mappings: narrower experts organize inputs into fewer intermediate features, concentrating prediction on a limited set of feature combinations, whereas wider experts provide more nonlinear units to accommodate richer contextual relationships. To select a mapping suited to the current context, the fused content provides multimodal feature information, while the scale weights indicate the relative contributions of different tissue ranges. Using both allows mapping capacity to adapt to context composition, linking how multiscale information is integrated with how it is used for prediction.

Building on this idea, we propose DAST (Dual-Adaptive Spatial Transcriptomics), a multimodal framework for spatial transcriptomics super-resolution. DAST constructs multiscale contexts in the feature space of a pretrained pathology encoder and dynamically learns scale contributions from their interactions with local expression features. The resulting scale weights organize the fused context and, together with its content, guide the selection of experts with different hidden widths, adapting expression mapping to the context composition. This connection enables joint learning of multiscale integration and expression mapping under spot-aggregated supervision.

Across ten Visium HD and Xenium datasets, DAST achieves the lowest RMSE and highest SSIM compared with six baselines. Factorial comparisons, ablations, and routing interventions examine the coupled design, while expression visualization, spatial-domain identification, and tertiary lymphoid structure (TLS) analysis assess its downstream utility.

Our main contributions are:

1. **Post-encoder dynamic multiscale multimodal fusion.**
   We learn scale contributions from histology–expression interactions at each scale, adapting context composition to local morphological and molecular information.
2. **Fusion-conditioned heterogeneous expert mapping.**
   We use fused content and scale composition to select candidate mappings of different capacities, enabling expression mapping to adapt jointly with dynamic context integration.
3. **Cross-platform evaluation and downstream analysis.**
   We evaluate super-resolution reconstruction on ten datasets, examine the synergistic design through factorial comparisons, ablations, and interventions, and explore applications in spatial domain identification and TLS analysis.

2 Related Work

Spatial Transcriptomics Super-Resolution. Spatial transcriptomics super-resolution uses spatial priors or histological information to constrain expression detail not directly resolved by low-resolution observations. Spatial priors supplement these observations through neighborhood modeling in BayesSpace and expression–coordinate relationships in STAGE (Zhao et al., 2021; Li et al., 2024); high-resolution histology additionally provides morphological evidence for distinguishing tissue components within spots. This complementarity underlies joint image–expression modeling in XFuse and histology-guided imputation in TESLA (Bergenstråhle et al., 2022; Hu et al., 2023), as well as the combination of histological representations and low-resolution expression for spot-supervised prediction in iStar, scstGCN, and HISTEX (Zhang et al., 2024; Xue et al., 2025b;a). Unlike conditional diffusion in Diff-ST, which uses high-resolution reference-expression targets during training (Wang et al., 2024), we focus on reconstruction supervised only by spot expression, where

reconstruction quality particularly depends on how complementary morphological and molecular evidence is used.

Multiscale Histology and Multimodal Integration. Using this complementarity requires both tissue context at suitable spatial scales and effective interaction with molecular observations. Vision Transformers encode images as patch-token sequences (Dosovitskiy et al., 2021); HIPT and UNI provide hierarchical and large-scale pretrained histological representations, respectively (Chen et al., 2022b; 2024). CrossFormer models cross-scale visual interactions, while RegionViT exchanges information between regional and local tokens (Wang et al., 2022; Chen et al., 2022a). For spatial transcriptomics super-resolution, iStar combines local and global features, scstGCN extends neighborhood context through spatial graphs, and TransformerST combines adaptive graph and cross-scale modeling (Zhang et al., 2024; Xue et al., 2025b; Zhao et al., 2024). HISTEX uses bidirectional cross-attention (BCA) between histological features and local gene features derived from low-resolution expression (Xue et al., 2025a). Multiscale feature extraction alone does not determine scale contributions to expression prediction. DAST learns these contributions after each encoded context interacts with local expression.

Adaptive Expression Mapping. Differences in context composition also affect downstream expression mapping. iStar and HISTEX use nonlinear predictors shared across locations to process different feature combinations (Zhang et al., 2024; Xue et al., 2025a). Sparse gating enables input-dependent expert selection (Shazeer et al., 2017); Switch Transformer uses Top-1 routing (Fedus et al., 2022), and DeepSeekMoE combines routed and shared experts (Dai et al., 2024). Soft MoE uses differentiable soft assignments to pass weighted combinations of input tokens to experts (Puigcerver et al., 2024). DAST uses fused content and scale weights to select experts with different hidden widths, coupling mapping-capacity selection with multimodal context integration.

3 METHOD

3.1 FRAMEWORK OVERVIEW

DAST combines H&E images, spatial coordinates, and observed spot expression to predict superpixel-level expression on a regular grid (Figure 1). A base feature map $\mathbf{F}$, comprising UNI features and aligned RGB and coordinate channels, supports multiscale context construction, followed by dynamic fusion with local gene context $\mathbf{G}_b$ and expert mapping conditioned on fused content and scale composition. For a local prediction block b, the overall process is

$$\mathbf{F} \xrightarrow{\text{Pool}\to\text{crop}} \{\mathbf{H}_b^{(r)}\}_{r\in\mathcal{R}} \xrightarrow[\mathbf{G}_b]{\text{Dynamic fusion}} (\mathbf{Z}_b, \mathbf{A}_b) \xrightarrow{\text{Expert mapping}} \widehat{\mathbf{Y}}_b. \tag{1}$$

Here, $\mathcal{R} = \{1, 2, 4\}$, $\mathbf{Z}_b$ is the fused representation, $\mathbf{A}_b$ collects scale weights, and $\widehat{\mathbf{Y}}_b \in \mathbb{R}^{T\times G}$ contains predictions for T positions and G genes. An image patch is a histological input to UNI, whereas prediction block b is a local window in the encoded feature space. Training aggregates predictions to the spot scale; Figure 1 omits b from position-level labels for clarity.

3.2 POST-ENCODER DYNAMIC MULTISCALE MULTIMODAL FUSION

The relative predictive value of local morphology and broader tissue structure varies with the tissue environment. We therefore learn fusion weights after each scale interacts with local expression, adapting context composition to both morphological and molecular information (Figure 1b).

We arrange spatial tokens extracted by UNI (Chen et al., 2024) at their original image locations and concatenate aligned RGB and normalized xy coordinate channels to obtain $\mathbf{F}$. Using spatial pooling factors $r \in \{1, 2, 4\}$, with $r = 1$ retaining the base map, we crop 7×7 windows around corresponding block centers to obtain $\mathbf{H}_b^{(r)} \in \mathbb{R}^{T\times C}$, where $T = 49$ and C is the feature dimension. These doubling scales extend spatial coverage while keeping token counts fixed, spanning approximately 56, 112, and $224\,\mu\text{m}$ per side. This construction enables expression prediction to draw on both local morphological detail at the base scale and surrounding tissue organization at broader

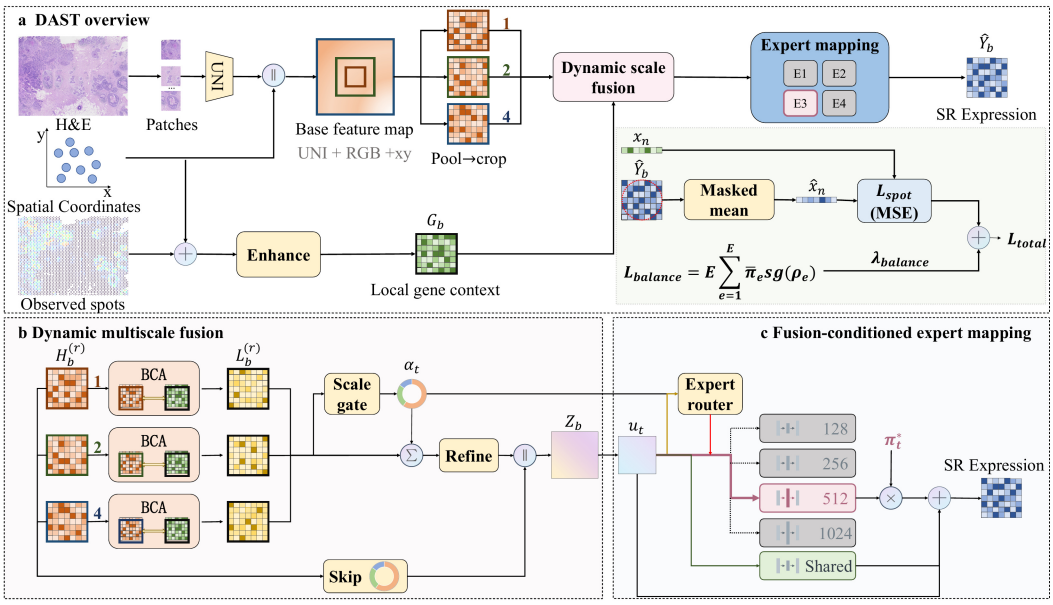

Figure 1: Overview of DAST. (a) Super-resolution reconstruction with spot-level supervision. (b) Dynamic multiscale fusion. (c) Expert mapping conditioned on content features and scale weights.

scales. Observed spot expression is spatially interpolated, and the K gene-context locations nearest the prediction-block center form $\mathbf{G}_b \in \mathbb{R}^{K\times G}$.

Each scale interacts with the same gene features through an independently parameterized bidirectional cross-attention (BCA) module (Xue et al., 2025a), yielding $\mathbf{L}_b^{(r)} = \mathrm{BCA}_r(\mathbf{H}_b^{(r)}, \mathbf{G}_b)$. At prediction-position index t, the gate computes scale weights from these multimodal representations and performs fusion:

$$\boldsymbol{\alpha}_{bt} = \mathrm{softmax}\left[g_{\mathrm{scale}}\left(\boldsymbol{\ell}_{bt}^{(1)} \| \boldsymbol{\ell}_{bt}^{(2)} \| \boldsymbol{\ell}_{bt}^{(4)}\right)\right], \tag{2a}$$

$$\bar{\boldsymbol{\ell}}_{bt} = \sum_{r\in\mathcal{R}} \alpha_{btr}\boldsymbol{\ell}_{bt}^{(r)}. \tag{2b}$$

Here, $\boldsymbol{\ell}_{bt}^{(r)}$ is the scale-specific representation at index t, and $\|$ denotes concatenation. Scale weights are nonnegative and sum to one; stacking $\boldsymbol{\alpha}_{bt}$ as rows gives $\mathbf{A}_b \in \mathbb{R}^{T\times|\mathcal{R}|}$, and stacking the fused vectors gives $\bar{\mathbf{L}}_b$.

The fused representation further interacts with gene context to refine the resulting multimodal combination, while a histological skip branch retains direct morphological information outside this interaction path:

$$\mathbf{V}_b = \mathrm{BCA}_{\mathrm{ref}}\left(\bar{\mathbf{L}}_b + R_1(\bar{\mathbf{L}}_b), \mathbf{G}_b\right), \tag{3a}$$

$$\widetilde{\mathbf{L}}_b = \mathbf{V}_b + R_2(\mathbf{V}_b), \tag{3b}$$

$$\mathbf{z}_{bt} = \widetilde{\boldsymbol{\ell}}_{bt} \| \left[\sum_{r\in\mathcal{R}} \alpha_{btr}\psi_r(\mathbf{h}_{bt}^{(r)})\right]. \tag{3c}$$

R_1 and R_2 are dimension-preserving linear maps, ψ_r combines an identity path with a low-dimensional nonlinear transform, and $\mathbf{h}_{bt}^{(r)}$ is the pre-interaction histological feature. The skip branch reuses the same scale weights to maintain context composition consistent with the multimodal branch. Stacking $\mathbf{z}_{bt}$ by index gives $\mathbf{Z}_b$.

3.3 FUSION-CONDITIONED HETEROGENEOUS EXPERT MAPPING

Dynamic scale fusion makes inputs depend on different context combinations, whose conversion to expression calls for different intermediate feature organization: some representations benefit from concentration on fewer predictive combinations, whereas others require more relationships to be retained simultaneously. A fixed-width prediction branch cannot select between these forms of organization according to the current input. To adapt expression mapping to the differences created by fusion, we provide candidate experts with different hidden widths and select among them using fused content and scale composition (Figure 1c).

The fused vector is first projected to $\mathbf{u}_{bt} = P(\mathbf{z}_{bt}) \in \mathbb{R}^d$ for an expert decoder with E candidate experts sharing input and output dimension d. Expert e transforms its input through an intermediate layer of width h_e:

$$f_e(\mathbf{u}) = \mathbf{W}_{2,e}\,\mathrm{Dropout}[\mathrm{GELU}(\mathbf{W}_{1,e}\mathbf{u} + \mathbf{b}_{1,e})] + \mathbf{b}_{2,e}. \tag{4}$$

Here, $\mathbf{W}_{1,e} \in \mathbb{R}^{h_e \times d}$ and $\mathbf{W}_{2,e} \in \mathbb{R}^{d \times h_e}$. Narrower experts transform inputs through fewer learned feature combinations, concentrating prediction on a limited set of combinations. Wider experts provide more nonlinear units, allowing multiple contextual relationships to be represented separately and subsequently combined. Heterogeneous widths thus offer both concentrated feature organization and richer mapping capacity for input-dependent selection. Uniformly increasing width expands capacity but does not require some inputs to be mapped through fewer intermediate features.

Under spot-level weak supervision, aggregated observations do not uniquely determine within-spot expression. This design provides an input-dependent structural constraint for fine-grained prediction, allowing concentrated feature combinations while retaining richer mapping capacity for representations requiring multiple relationships.

The router combines fused content and scale weights to compute expert probabilities and selects the most probable expert:

$$\boldsymbol{\pi}_{bt} = \mathrm{softmax}\left(\frac{\mathbf{W}_u\mathbf{u}_{bt} + \mathbf{W}_\alpha\boldsymbol{\alpha}_{bt} + \mathbf{b}_{\mathrm{gate}}}{\tau}\right), \tag{5a}$$

$$e^\star_{bt} = \arg\max_{1\le e\le E} \pi_{bte}, \qquad \pi^\star_{bt} = \pi_{bt,e^\star_{bt}}. \tag{5b}$$

Here, $\tau > 0$ is the temperature. Fused content describes multimodal information, while scale weights explicitly record contributions from different tissue ranges, distinguishing compositions that may produce similar fused representations. Expert width is not fixed to a particular scale: local information may contain complex relationships, whereas broader context may provide concentrated predictive evidence.

The selected expert retains its original routing-probability weight and is combined with the input residual and an always-active shared expert to predict gene expression:

$$\mathbf{m}_{bt} = \mathbf{u}_{bt} + \pi^\star_{bt} f_{e^\star_{bt}}(\mathbf{u}_{bt}) + f_{\mathrm{sh}}(\mathbf{u}_{bt}), \tag{6a}$$

$$\widehat{\mathbf{y}}_{bt} = \mathrm{Softplus}[\mathbf{W}_o\,\mathrm{LN}(\mathbf{m}_{bt}) + \mathbf{b}_o]\,. \tag{6b}$$

The final gene readout maps the hidden representation $\mathbf{m}_{bt}$ to gene dimension through layer normalization and a linear transform, with Softplus ensuring nonnegative predictions. The input residual preserves fused information, routed experts learn condition-dependent transformations, and the shared expert learns reusable transformations across locations.

3.4 WEAKLY SUPERVISED LEARNING

Training compares mean predictions within each circular spot footprint with normalized observed expression using MSE and adds an expert-balancing regularizer. For a training batch $\mathcal{B}$, the objective is

$$\mathcal{L} = \frac{1}{|\mathcal{B}|G} \sum_{n \in \mathcal{B}} \left\| \frac{1}{|\Omega_n|} \sum_{i \in \Omega_n} \widehat{\mathbf{y}}_{ni} - \mathbf{x}_n \right\|_2^2 + \lambda_{\text{balance}} E \sum_{e=1}^{E} \bar{\pi}_e \ \text{sg}(\rho_e). \tag{7}$$

Here, Ω_n is the circular mask of spot n, $\mathbf{x}_n$ is its gene-wise min–max-normalized expression, and $\widehat{\mathbf{y}}_{ni}$ is the prediction at position i in the block centered on that spot. $\bar{\pi}_e$ and ρ_e are the mean expert probability and Top-1 assignment fraction over all routed positions in the batch, including those outside the masks; the shared expert is excluded. sg stops gradients, and λ_{balance} weights the balance term.

The reconstruction term constrains aggregate expression without assigning identical labels to individual superpixels, whose variation is modeled through histology, local expression context, and adaptive mapping. The balance term discourages overly concentrated expert assignment.

4 Experiments

4.1 Experimental Setup

Data and evaluation setting. We evaluate DAST on five Visium HD and five Xenium datasets. Visium HD expression on $8 \times 8\,\mu\text{m}^2$ bins serves as the high-resolution reference; Xenium cell expression is assigned to equal-sized bins in proportion to cell–bin overlap area. Summing reference expression within spot footprints produces pseudo-Visium inputs. We use 1,000 highly variable genes selected from each Visium HD high-resolution reference matrix and the Xenium panel's Gene Expression genes. Models use spot expression and H&E images from the same section for training, with reconstruction evaluated against the high-resolution grid reference.

Baselines and metrics. We compare with STAGE, XFuse, TESLA, iStar, HISTEX, and scstGCN (Li et al., 2024; Bergenstråhle et al., 2022; Hu et al., 2023; Zhang et al., 2024; Xue et al., 2025a;b). Predicted and reference maps are independently min–max normalized for each gene before computing RMSE and SSIM and averaging over valid evaluation genes. These metrics therefore assess normalized spatial expression patterns.

DAST uses frozen UNI features (Chen et al., 2024), contexts at scales 1, 2, and 4, and routed experts with hidden widths of 128, 256, 512, and 1,024. For DAST and its trained controls, we take the locationwise, gene-wise median of five independently trained models' final-checkpoint predictions, restore the output scale, and assemble the super-resolved map by spatial coordinate.

Implementation and evaluation details, supplementary experiments, and additional visualizations are provided in the appendix.

4.2 Gene Expression Reconstruction

Quantitative results. DAST achieves the lowest RMSE and highest SSIM on all ten datasets (Table 1), improving both reconstruction error and spatial structural similarity of normalized expression patterns across the two platforms.

Spatial-expression patterns. Across the three genes in Figure 2, DAST shows the closest visual agreement with the high-resolution reference (GT), better preserving the ring-like boundaries of ACTG2, localized enrichment of CXCR4, and clustered high-expression structures of SCD. Enlarged views reveal more scattered noise in HISTEX, block-like artifacts in scstGCN, and broader, smoother patterns in iStar. XFuse exhibits larger spatial discrepancies, particularly failing to recover the prominent SCD structures in the selected region.

4.3 Ablation Studies and Routing Analysis

Scale fusion and expert mapping. To examine how the value of multiscale information depends on downstream expression mapping, we combine single-scale (S; scale 1) or dynamic multiscale (M; scales 1, 2, and 4) representations with dense (D) or heterogeneous-expert (M) mappings. The four

Table 1: Reconstruction on ten datasets. Lower RMSE and higher SSIM are better. Best and second-best results are bold and underlined, respectively.

Visium HD

Method	HBCHD		HOHD		MBHD		MEHD		MKHD	
	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM
STAGE	0.1950	0.3762	0.1892	0.4233	0.1638	0.3561	0.1582	0.4123	0.1947	0.4789
XFuse	0.1439	0.4268	0.1678	0.4565	0.1236	0.3894	0.1927	0.5163	0.1422	0.5662
TESLA	0.1264	0.6237	0.1064	0.7157	0.1004	0.6136	0.1736	0.6588	0.1039	0.6245
iStar	0.1183	0.6976	0.1186	0.5774	0.0797	0.6934	0.0693	0.7234	0.0854	0.6618
scstGCN	0.0428	0.8395	0.0417	0.8722	0.0412	0.8528	0.0381	0.9083	0.0347	0.8482
HISTEX	0.0492	0.8154	0.0235	0.9065	0.0488	0.8109	0.0434	0.8899	0.0265	0.8753
DAST	**0.0346**	**0.8526**	**0.0169**	**0.9273**	**0.0327**	**0.8690**	**0.0324**	**0.9171**	**0.0218**	**0.8990**

Xenium

Method	HBC_S1R1		HBC_S1R2		HBC_S2		HMTC		HPC	
	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM
STAGE	0.1483	0.4565	0.1648	0.3965	0.2130	0.4428	0.1457	0.3624	0.2145	0.3781
XFuse	0.1836	0.4114	0.1536	0.3452	0.1495	0.3834	0.1772	0.4565	0.1524	0.4551
TESLA	0.0984	0.5138	0.0943	0.6476	0.0862	0.5493	0.1043	0.5043	0.1384	0.4892
iStar	0.0817	0.5851	0.0746	0.6298	0.0672	0.6839	0.0611	0.7053	0.0854	0.5394
scstGCN	0.0452	0.6659	0.0449	0.7760	0.0473	0.8266	0.0304	0.8533	0.0362	0.7534
HISTEX	0.0498	0.6401	0.0489	0.7591	0.0452	0.8166	0.0282	0.8585	0.0365	0.7461
DAST	**0.0355**	**0.7215**	**0.0377**	**0.8028**	**0.0396**	**0.8525**	**0.0234**	**0.8683**	**0.0300**	**0.7908**

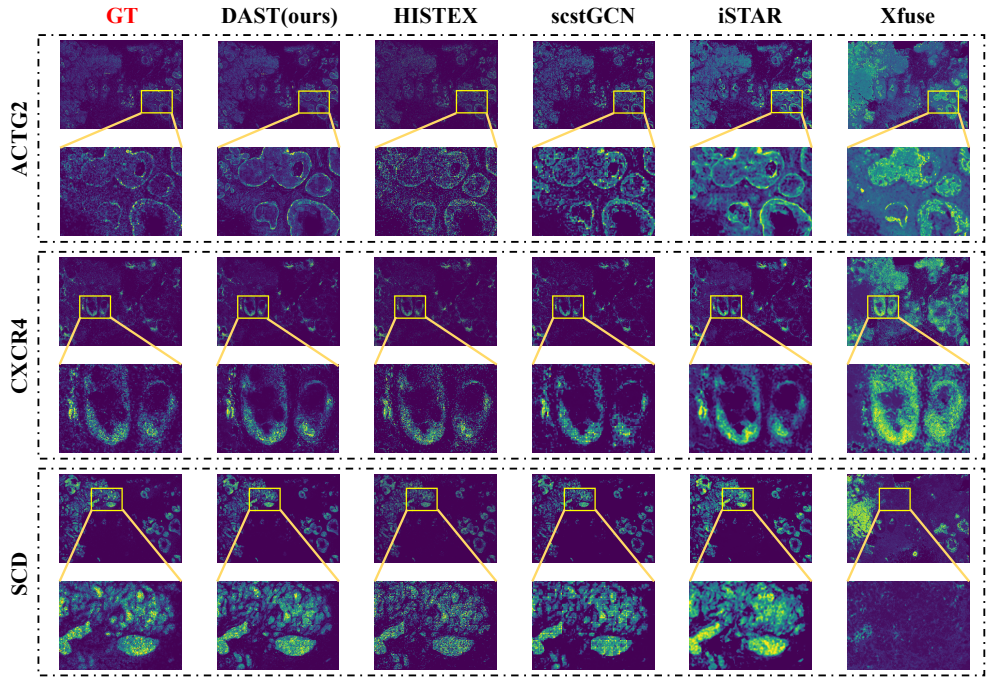

Figure 2: Expression reconstruction for ACTG2, CXCR4, and SCD against the high-resolution reference (GT). Yellow boxes indicate enlarged regions.

configurations share the training and evaluation protocol; dense mappings approximately match the total parameter count of the expert decoder, and both expert configurations retain the shared branch.

Table 2: Scale–mapping comparison. S–D: single-scale dense; S–M: single-scale experts; M–D: multiscale dense; M–M: full model. Best values at the displayed precision are bold.

Dataset	S–D		S–M		M–D		M–M (Full)	
	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM
HBCHD	0.0362	0.8472	0.0348	0.8505	0.0364	0.8455	**0.0346**	**0.8526**
HOHD	0.0174	0.9261	0.0174	0.9270	0.0181	0.9253	**0.0169**	**0.9273**
MBHD	0.0339	0.8674	0.0332	0.8683	0.0336	0.8676	**0.0327**	**0.8690**
MEHD	0.0326	0.9144	**0.0324**	0.9163	0.0344	0.9115	**0.0324**	**0.9171**
MKHD	**0.0213**	0.8972	0.0222	0.8974	0.0228	0.8956	0.0218	**0.8990**
HBC_S1R1	0.0369	0.7183	0.0367	0.7196	0.0378	0.7176	**0.0355**	**0.7215**
HBC_S1R2	0.0391	0.8005	0.0380	0.8022	0.0396	0.8004	**0.0377**	**0.8028**
HBC_S2	0.0414	0.8424	0.0402	0.8477	0.0413	0.8502	**0.0396**	**0.8525**
HMTC	0.0238	0.8647	**0.0233**	0.8670	0.0234	0.8663	0.0234	**0.8683**
HPC	0.0304	0.7881	**0.0300**	0.7886	0.0311	0.7796	**0.0300**	**0.7908**

The full model achieves the highest SSIM on all ten datasets and the lowest RMSE on most datasets (Table 2). On some datasets, multiscale dense mapping does not outperform single-scale dense mapping, but improves when expert mapping is introduced. Comparing the gains from experts under the two representation conditions, the gain difference is positive on 9/10 datasets for RMSE and 8/10 for SSIM. This trend supports adapting the downstream mapping to the context formed by dynamic fusion.

Fusion and mapping components. Component ablations on HBC_S1R1 examine the specific fusion and mapping choices (Table 3a). The full model outperforms all three single-scale variants and mean fusion, whereas mean fusion does not exceed scale 1, indicating that the benefit of additional contextual ranges depends on their integration.

Four homogeneous experts of width 512 underperform the heterogeneous experts despite having more routed-expert parameters. This supports offering mappings with different widths, allowing input-dependent selection between more concentrated feature combinations and richer combinations of relationships. The shared expert further improves both metrics.

Inference interventions. Permuting fusion weights or content-routing logits, or forcing any single expert in the trained model, reduces reconstruction quality (Table 3b). Even always using the widest expert does not match the original routing, indicating that input-dependent routing is more effective than a fixed choice within this trained expert set.

Table 3: Ablations and inference interventions on HBC_S1R1. Scale 1 is S–M in Table 2. Shuffle results average five permutations; forced experts use the trained model without retraining.

(a) Component ablations

Setting	RMSE	SSIM
Full model	**0.0355**	**0.7215**
Scale 1 only	0.0367	0.7196
Scale 2 only	0.0394	0.6924
Scale 4 only	0.0380	0.6850
Mean fusion	0.0380	0.7115
Homogeneous experts	0.0368	0.7191
No shared expert	0.0367	0.7177

(b) Inference interventions

Setting	RMSE	SSIM
Full model	**0.0355**	**0.7215**
Fusion-weight shuffle	0.0368	0.7153
Content-logit shuffle	0.0388	0.6995
Force E_1 (128)	0.0393	0.7057
Force E_2 (256)	0.0385	0.7092
Force E_3 (512)	0.0388	0.7086
Force E_4 (1024)	0.0367	0.7173

4.4 Spatial Organization and Downstream Analysis

Spatial-domain identification. To assess downstream spatial resolution, we applied k-means to DAST's pre-decoder superpixel-level multimodal representations on MBHD and compared the domains with STAGATE (Dong & Zhang, 2022) and GraphST (Long et al., 2023) on the original

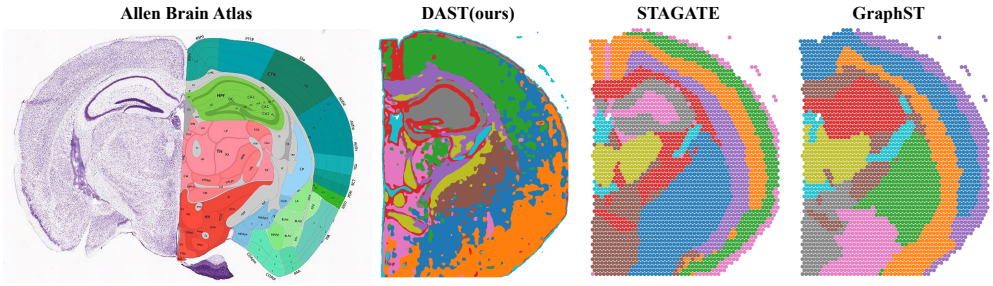

Figure 3: Spatial domain comparison: anatomical atlas reference, DAST with superpixel-level k-means, and spot-level STAGATE and GraphST. Cluster colors are independently assigned.

spots, using the Allen Brain Atlas as an anatomical reference (Allen Institute for Brain Science) (Figure 3). Despite using a simple clustering method, DAST resolves curved boundaries and narrow regional subdivisions more clearly, demonstrating the value of finer representations for delineating tissue domains.

TLS-associated spatial patterns. To examine TLS-associated immune organization (Fridman et al., 2022), we analyzed the HER2ST (Andersson et al., 2021) H_1 (H1_re) whole section using a composite score of CD74, CD79A, CD52, PTGDS, and TRBC2 (Figure 4). The original spot-level expression and superpixel-level predictions use the same five-gene set. The map computed from DAST-predicted expression shows elevated signals in the two manually annotated immune-infiltrated regions and reveals the locations, regional shapes, and local spatial detail of candidate TLS-associated enrichment.

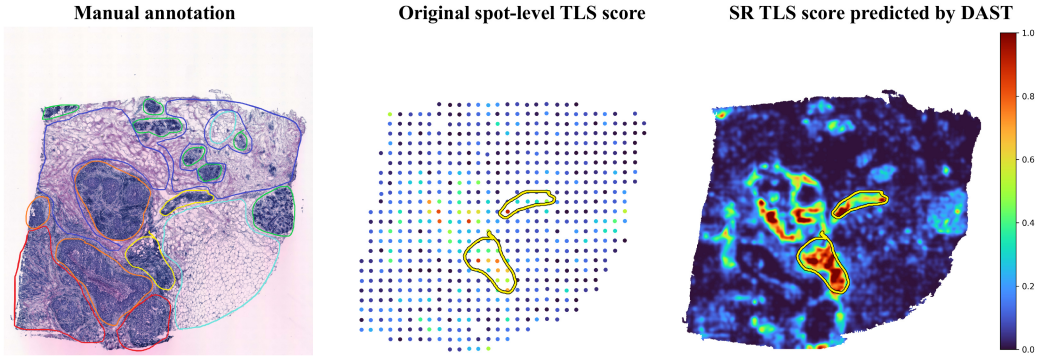

Figure 4: TLS-associated signals in H1_re. Panels show the H&E image, the five-gene combined signal at measured spot locations, and the signal derived from superpixel-level predictions. Yellow outlines mark immune-infiltrated regions.

5 Conclusion

We presented DAST for jointly adapting multiscale context composition and its mapping to superpixel-level gene expression under spot-aggregated supervision. Evaluation across ten datasets shows consistent reconstruction improvements, while factorial comparisons and ablations support selecting mappings of different capacities using fused content and scale composition. Spatial-domain identification and candidate TLS analysis further demonstrate the utility of finer representations for studying tissue organization and immune-related spatial patterns. Future work will investigate transfer across tissues and platforms and quantify reconstruction uncertainty to support the interpretation of predicted spatial details.

AI USE STATEMENT

AI tools were used only to assist with English-language polishing and manuscript-format checking.

REFERENCES

Allen Institute for Brain Science. Allen reference atlases: Atlas viewer. URL https://atlas.brain-map.org/. Accessed 2026-09-23.

Alma Andersson, Ludvig Larsson, Linnea Stenbeck, et al. Spatial deconvolution of HER2-positive breast cancer delineates tumor-associated cell type interactions. *Nature Communications*, 12: 6012, 2021. doi: 10.1038/s41467-021-26271-2.

Ludvig Bergenstråhle, Bryan He, Joseph Bergenstråhle, et al. Super-resolved spatial transcriptomics by deep data fusion. *Nature Biotechnology*, 40(4):476–479, 2022. doi: 10.1038/s41587-021-01075-3.

Chun-Fu Chen, Rameswar Panda, and Quanfu Fan. RegionViT: Regional-to-local attention for vision transformers. In *International Conference on Learning Representations*, 2022a. URL https://openreview.net/forum?id=T__V3uLix7V.

Kok Hao Chen, Alistair N. Boettiger, Jeffrey R. Moffitt, Siyuan Wang, and Xiaowei Zhuang. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, 348(6233):aaa6090, 2015. doi: 10.1126/science.aaa6090.

Richard J. Chen, Chengkuan Chen, Yicong Li, et al. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16144–16155, 2022b.

Richard J. Chen, Tong Ding, Ming Y. Lu, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024. doi: 10.1038/s41591-024-02857-3.

Damai Dai, Chengqi Deng, Chenggang Zhao, R.X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y.K. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1280–1297. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.70. URL https://aclanthology.org/2024.acl-long.70/.

Kangning Dong and Shihua Zhang. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nature Communications*, 13:1739, 2022. doi: 10.1038/s41467-022-29439-6.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Wolf H. Fridman, Maxime Meylan, Florent Petitprez, Cheng-Ming Sun, Antoine Italiano, and Catherine Sautès-Fridman. B cells and tertiary lymphoid structures as determinants of tumour immune contexture and clinical outcome. *Nature Reviews Clinical Oncology*, 19(7):441–457, 2022. doi: 10.1038/s41571-022-00619-z.

Jian Hu, Kyle Coleman, Daiwei Zhang, et al. Deciphering tumor ecosystems at super resolution from spatial transcriptomics with TESLA. *Cell Systems*, 14(5):404–417.e4, 2023. doi: 10.1016/j.cels.2023.03.008.

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. doi: 10.1093/nar/gkae294.

Yahui Long, Kok Siong Ang, Mengwei Li, Kian Long Kelvin Chong, Raman Sethi, Chengwei Zhong, Hang Xu, Zhiwei Ong, Karishma Sachaphibulkij, Ao Chen, et al. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST. *Nature Communications*, 14(1):1155, 2023. doi: 10.1038/s41467-023-36796-3.

Lambda Moses and Lior Pachter. Museum of spatial transcriptomics. *Nature Methods*, 19(5):534–546, 2022. doi: 10.1038/s41592-022-01409-2.

Joan Puigcerver, Carlos Riquelme Ruiz, Basil Mustafa, and Neil Houlsby. From sparse to soft mixtures of experts. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/79fea214543ba263952ac3f4e5452b14-Abstract-Conference.html.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=B1ckMDqlg.

Patrik L. Ståhl, Fredrik Salmén, Sanja Vickovic, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, 2016. doi: 10.1126/science.aaf2403.

Tam Vu, Alexander Vallmitjana, Joshua Gu, et al. Spatial transcriptomics using combinatorial fluorescence spectral and lifetime encoding, imaging and analysis. *Nature Communications*, 13:169, 2022. doi: 10.1038/s41467-021-27798-0.

Wenxiao Wang, Lu Yao, Long Chen, Binbin Lin, Deng Cai, Xiaofei He, and Wei Liu. CrossFormer: A versatile vision transformer hinging on cross-scale attention. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=_PHymLIxuI.

Xiaofei Wang, Xingxu Huang, Stephen Price, and Chao Li. Cross-modal diffusion modelling for super-resolved spatial transcriptomics. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15003 of *Lecture Notes in Computer Science*, pp. 98–108, 2024. doi: 10.1007/978-3-031-72384-1_10.

Shuailin Xue, Changmiao Wang, Xiaomao Fan, and Wenwen Min. Inferring super-resolved gene expression by integrating histology images and spatial transcriptomics with HISTEX. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15972 of *Lecture Notes in Computer Science*, pp. 296–306, 2025a. doi: 10.1007/978-3-032-05169-1_29. URL https://papers.miccai.org/miccai-2025/0449-Paper1633.html.

Shuailin Xue, Fangfang Zhu, Jinyu Chen, and Wenwen Min. Inferring single-cell resolution spatial gene expression via fusing spot-based spatial transcriptomics, location, and histology using GCN. *Briefings in Bioinformatics*, 26(1):bbae630, 2025b. doi: 10.1093/bib/bbae630.

Daiwei Zhang, Amelia Schroeder, Hanying Yan, et al. Inferring super-resolution tissue architecture by integrating spatial transcriptomics with histology. *Nature Biotechnology*, 42(9):1372–1377, 2024. doi: 10.1038/s41587-023-02019-9.

Chongyue Zhao, Zhongli Xu, Xinjun Wang, et al. Innovative super-resolution in spatial transcriptomics: a transformer model exploiting histology images and spatial gene expression. *Briefings in Bioinformatics*, 25(2):bbae052, 2024. doi: 10.1093/bib/bbae052.

Edward Zhao, Matthew R. Stone, Xing Ren, et al. Spatial transcriptomics at subspot resolution with BayesSpace. *Nature Biotechnology*, 39(11):1375–1384, 2021. doi: 10.1038/s41587-021-00935-2.

A Code Availability

An anonymized implementation of DAST will be provided in the supplementary materials.

B Data Preparation and Evaluation

B.1 Data Sources and Gene Sets

The benchmark comprises HBCHD, HOHD, MBHD, MEHD, and MKHD from Visium HD, and HBC_S1R1, HBC_S1R2, HBC_S2, HMTC, and HPC from Xenium. Each section provides a high-resolution grid reference and pseudo-Visium input for within-section expression reconstruction.

Table B.1: Datasets used for reconstruction and downstream analysis. HER2ST is used only for the TLS analysis and is not part of the ten-dataset reconstruction benchmark.

Dataset	Platform	Tissue/sample	Role
HBCHD	Visium HD	Human breast cancer, fresh frozen	Main benchmark
HOHD	Visium HD	Human ovarian cancer, fresh frozen	Main benchmark
MBHD	Visium HD	Mouse brain, fresh frozen	Main benchmark
MEHD	Visium HD	Mouse embryo, FFPE	Main benchmark
MKHD	Visium HD	Mouse kidney	Main benchmark
HBC_S1R1	Xenium	Human breast cancer	Main benchmark
HBC_S1R2	Xenium	Human breast cancer	Main benchmark
HBC_S2	Xenium	Human breast cancer	Main benchmark
HMTC	Xenium	Human heart	Main benchmark
HPC	Xenium	FFPE human pancreas	Main benchmark
HER2ST H_1 / H1_re	Spatial Transcriptomics	HER2-positive human breast cancer	TLS analysis

For Visium HD, we select 1,000 highly variable genes from the high-resolution reference matrix before constructing pseudo-Visium inputs. Selection uses Scanpy's Seurat v3 method. For Xenium, we retain panel features annotated as Gene Expression, without further highly variable gene selection. High-resolution expression thus contributes to benchmark construction and Visium HD gene-set determination, while network optimization uses spot-aggregated expression as its supervision target.

B.2 Grid References and Pseudo-Visium Inputs

At an image scale of $0.5\,\mu\mathrm{m}/\mathrm{pixel}$, a 16×16-pixel region corresponds to an $8 \times 8\,\mu\mathrm{m}^2$ superpixel. Visium HD references are constructed from binned expression and spatial coordinates. Xenium expression is distributed from each cell in proportion to its overlap area with each bin, and contributions from different cells are summed within each bin.

Pseudo-Visium spot centers follow a staggered layout with a $100\,\mu\mathrm{m}$ coordinate step and a $50\,\mu\mathrm{m}$ vertical offset between adjacent columns. The default spot radius is 55 image pixels ($27.5\,\mu\mathrm{m}$), or $55/16$ on the superpixel grid. Its discrete circular mask contains 37 locations within a 7×7 window. Spot expression is the sum of reference expression within this mask, with negative values at invalid locations set to zero before aggregation.

Candidate spots are filtered by the preprocessing tissue mask; the current mask query integer-divides image coordinates by 16 and subtracts three grid units. Expression evaluation uses the saved mask.png, while masks for histological-feature analysis are stored separately.

B.3 Training Normalization

Before training, raw spot expression and interpolation-enhanced expression are independently min–max normalized per gene using their respective minima and maxima, with 10^{-12} added to the denominator.

No additional library-size normalization or log transformation precedes this step. Normalized spot expression provides the training target.

B.4 Evaluation Metrics

We evaluate super-resolution reconstruction on the Visium HD and Xenium benchmarks using root mean square error (RMSE) and structural similarity index measure (SSIM), which assess pointwise expression error and spatial structural similarity, respectively.

For each gene, the predicted expression and high-resolution reference are represented as two-dimensional spatial maps, with each location's intensity corresponding to its expression value.

Predictions and references are paired using the saved gene order. Bottom and right padding is removed from predictions using feature-map metadata. Remaining size mismatches are resolved by bilinear interpolation for predictions and nearest-neighbor interpolation for tissue masks.

Negative reference values are set to zero. Each predicted and reference map is then independently min–max normalized over its complete two-dimensional array to the range $[0, 1]$, with constant maps mapped to zero. Tissue masking is applied after normalization.

Root mean square error. Let P_g and T_g denote the normalized prediction and reference maps for gene g, and let M be the binary tissue mask. The tissue locations are $\mathcal{T} = \{i : M_i = 1\}$.

RMSE is computed over corresponding tissue locations:

$$\mathrm{RMSE}_g = \sqrt{\frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} (P_{g,i} - T_{g,i})^2}. \tag{8}$$

Here, $|\mathcal{T}|$ is the number of tissue locations. A lower RMSE indicates smaller pointwise differences between normalized predicted and reference expression.

Structural similarity. SSIM complements pointwise error by comparing spatial expression patterns through local means, variances, and covariance.

For corresponding local windows x and y, the local SSIM is

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \tag{9}$$

Here, μ_x and μ_y are local means, σ_x^2 and σ_y^2 are local variances, and σ_{xy} is the local covariance. The stabilizing constants are $C_1 = (K_1 L)^2$ and $C_2 = (K_2 L)^2$, where L is the input data range.

We compute SSIM on the complete rectangular maps $M \odot P_g$ and $M \odot T_g$, with non-tissue locations set to zero. We use skimage.metrics.structural_similarity with data_range=1 and the remaining parameters at their implementation defaults.

The returned image-level score averages local-window similarities. Higher SSIM indicates closer agreement with the reference spatial expression structure.

Dataset-level aggregation. Both metrics are computed per gene and then averaged separately over valid gene-level values, excluding NaNs. Genes without prediction files are omitted, and the evaluated gene set is recorded.

Component ablations and factorial comparisons use the same valid-gene set for paired comparisons.

Expression visualizations illustrate spatial patterns, while quantitative results follow the numerical evaluation procedure above.

B.5 Comparison Methods

The reconstruction benchmark compares DAST with STAGE, XFuse, TESLA, iStar, HISTEX, and scstGCN, with quantitative results reported in Table 1.

C Construction of Spatial Context

C.1 Multiscale Histological Features

Frozen UNI extracts spatial tokens from 224×224-pixel image patches, with each patch token corresponding to a 16×16-pixel grid cell; tokens are arranged by their original-image positions. The 1,024-dimensional UNI feature is concatenated with aligned mean RGB values and two normalized spatial-coordinate channels to form a 1,029-dimensional base feature. RGB values are divided by 255, and coordinates specify the grid center's relative position in the full image. Each prediction block has $T = 49$ output positions, and its local gene context has $K = 5$ locations.

We retain the base scale for local morphological detail and progressively expand coverage to incorporate nearby and broader tissue context.

The doubling configuration covers these complementary ranges with three branches, while fixed-size windows control the token count within each branch.

Scale 1 retains the base map; scales 2 and 4 apply nonoverlapping average pooling with kernel and stride equal to the scale factor, padding the bottom and right with zeros when needed. Image coordinates are converted to base-grid coordinates using metadata, divided by the scale factor, and rounded with NumPy's round-to-nearest operation to obtain each scale's center. A 7×7 window is extracted at each scale with zero-padding outside the map, covering approximately 56×56, 112×112, and $224 \times 224\,\mu\mathrm{m}^2$.

Each scale represents its tissue window with a 7×7 token layout. Fusion combines scale-specific features at corresponding relative indices within these windows, followed by expression decoding on the base-grid prediction block.

C.2 Local Gene Context

Expression enhancement constructs a staggered grid from spot coordinates, fills observed expression by nearest-neighbor interpolation, zeros corresponding non-tissue positions, and increases grid density by bilinear interpolation without aligning corner-pixel centers. Zero-padding and adjacency-support filtering generate the enhanced expression matrix and coordinates. For each training spot or inference-block center, a k-d tree search with Euclidean distance selects the five nearest enhanced-expression locations to form a $5 \times G$ matrix shared by all prediction locations and scale-specific modules within the block.

D Model Configuration, Training, and Inference

D.1 Network Configuration

Table D.1: Network configuration.

Component	Configuration
Scale / gene inputs	49 tokens of 1,029 dimensions per scale; 5 gene-context locations
Per-scale BCA	Independent parameters; dimension 128; 8 attention heads
BCA computation	H&E queries genes, genes query H&E, then latent representations interact
Scale gate	$384 \to 3$, softmax; weights initialized to zero, bias $(2, 0, 0)$
Fusion refinement	128-dimensional linear residual, BCA, then a second linear residual
H&E skip adapters	Independent $1{,}029 \to 32 \to 1{,}029$ maps; GELU and input residual
Decoder projection	$1{,}157 \to 512$, LayerNorm, GELU, Dropout
Routed experts	$512 \to h_e \to 512$; $h_e \in \{128, 256, 512, 1{,}024\}$; GELU, Dropout
Shared expert	$512 \to 128 \to 512$; GELU, Dropout
Router	512-dimensional content and 3-dimensional scale branches; 4 logits; Top-1
Output head	LayerNorm; $512 \to G$; Softplus

The expert input and output dimension is $d = 512$; the always-active shared expert has parameters independent of the width-128 routed expert. Both linear layers in each skip adapter are bias-free, with the final layer initialized to zero so the adapter starts as an identity transform. The two refinement residual maps are bias-free 128-dimensional linear layers. The router's content branch includes a bias and its scale branch does not. The selected expert retains its original softmax probability weight without renormalization to one.

D.2 Training and Ensemble Prediction

Table D.2: Training configuration for DAST and its trained controls.

Setting	Value or implementation
Optimizer / learning rate	Adam / 10^{-4}; no scheduler
Epochs / runs	500 epochs per run; 5 runs with seeds 0–4
Batch size	$\max(1, \min(32, \lfloor N/16 \rfloor))$, with N training spots
Data loading	Fixed sample order (no shuffling)
Spot target	MSE between mean prediction in the circular mask and normalized spot expression
Balance weight	0.01
Temperature / dropout	1.0 / 0.1
CUDA setup	Single device; mixed 16-bit precision; deterministic mode
Checkpoint	Final checkpoint of each run
Framework	PyTorch, PyTorch Lightning, Python 3.11

The five models are independently initialized and trained on the same section. Each final checkpoint is retained without selection by validation metrics against high-resolution reference expression. Their median ensemble produces one final prediction per configuration; table entries are therefore metrics of the ensemble prediction.

At inference, the base grid is padded at the bottom and right to multiples of seven and processed in nonoverlapping 7×7 blocks. Each block extracts multiscale histological and local gene contexts around its center. Predictions are combined by a locationwise, gene-wise median, transformed to the export scale, assembled by grid position, and cropped to remove padding.

For normalized prediction $v_{ig}^{(s)}$ from run s, gene-wise minima and maxima a_g, b_g of raw spot expression, and M locations in the spot mask, the exported expression is

$$\widehat{y}_{ig}^{\text{export}} = \frac{a_g + (b_g - a_g)\,\text{median}_{s=1,\ldots,5}(v_{ig}^{(s)})}{M}, \qquad M = 37. \tag{10}$$

Training averages within-spot predictions on the normalized scale; division by M at export converts the restored spot-expression scale to the grid scale. Whole-tissue predictions receive no additional post hoc spot-consistency calibration.

E Controlled Comparisons and Inference Interventions

E.1 Factorial Controls and Gain Differences

Table 2 pairs single-scale or dynamic multiscale representations with either dense mapping or heterogeneous experts. Both S–M and M–M retain the shared expert. Data processing, spot-aggregated targets, training duration, five-model ensembling, and valid evaluation-gene sets are shared across configurations.

The dense mapper retains the same form of input projection, output normalization, and expression head, replaces expert mapping with a residual feed-forward network, and adjusts hidden width to approximately match total decoder parameters. This control matches total parameter capacity; per-location active parameters and compute are not matched.

Let R and S denote RMSE and SSIM for each configuration. We compare the improvement from replacing dense mapping with experts under multiscale versus single-scale representations:

$$\begin{aligned}\Delta_{\mathrm{RMSE}} &= (R_{\mathrm{M\text{–}D}} - R_{\mathrm{M\text{–}M}}) - (R_{\mathrm{S\text{–}D}} - R_{\mathrm{S\text{–}M}}),\\ \Delta_{\mathrm{SSIM}} &= (S_{\mathrm{M\text{–}M}} - S_{\mathrm{M\text{–}D}}) - (S_{\mathrm{S\text{–}M}} - S_{\mathrm{S\text{–}D}}).\end{aligned} \tag{11}$$

Positive values indicate larger gains from experts under multiscale representations. Their counts summarize an interaction trend across datasets and do not constitute a statistical significance test.

E.2 Component Ablations

The single-scale variants in Table 3a use scale 1, 2, or 4 while retaining expert mapping; the scale-1 result is the S–M entry in Table 2. Mean fusion replaces learned scale fusion with equal contributions from all scales. Homogeneous experts set all four routed hidden widths to 512. The no-shared-expert variant removes the always-active shared branch.

E.3 Additional Ablations on Visium HD

Table E.1: Additional ablations on HOHD. Lower RMSE and higher SSIM are better.

Setting	Scales	RMSE	SSIM
Scale 1	1	0.0174	0.9270
Scale 2	2	0.0197	0.9247
Scale 4	4	0.0179	0.9215
Mean fusion	1, 2, 4	0.0182	0.9244
Full model	**1, 2, 4**	**0.0169**	**0.9273**
HISTEX-style dense decoder	1, 2, 4	0.0224	0.9118
Parameter-matched dense decoder	1, 2, 4	0.0181	0.9253
Homogeneous MoE	1, 2, 4	0.0177	0.9262
Content-only router	1, 2, 4	0.0177	0.9261
Without shared expert	1, 2, 4	0.0178	0.9257

Among the HOHD configurations in Table E.1, the full model achieves the lowest RMSE and highest SSIM. It also outperforms the single-scale and mean-fusion configurations on both metrics.

The table further reports the dense-decoder and alternative expert-routing results.

E.4 Supplementary Router Ablation

Content-only routing retains dynamic fusion and removes the router's scale branch. On HBC_S1R1, RMSE changes from 0.0355 to 0.0364 and SSIM from 0.7215 to 0.7201, providing supplementary evidence for including scale composition alongside fused content.

E.5 Fusion-Weight and Content-Logit Permutations

Purpose and setup. These inference-time interventions examine whether the correspondence of learned scale weights and content-routing signals with local inputs contributes to reconstruction. All trained parameters remain fixed, and no retraining is performed.

Let i index tissue positions participating in the intervention and let σ denote a random permutation of these positions. Each permutation moves complete vectors between positions without changing the ordering of scale or expert components within a vector.

Fusion-weight permutation. At each position, we retain the local features at all scales and replace the complete scale-weight vector with that from another tissue position:

$$\bar{\boldsymbol{\ell}}_i^{\mathrm{perm}} = \sum_{r\in\mathcal{R}} \alpha_{\sigma(i),r}\boldsymbol{\ell}_i^{(r)}. \tag{12}$$

Table E.2: Supplementary router ablation on HBC_S1R1.

Setting	RMSE	SSIM
Full model	**0.0355**	**0.7215**
Content-only router	0.0364	0.7201

This preserves the collection of learned weight vectors and their normalization while changing their correspondence with local features. Scale features, scale identities, and tissue coordinates remain unchanged.

The permuted weights are used in both multimodal fusion and the histological skip branch, after which downstream representations and predictions are recomputed.

The router's explicit scale input retains the original local scale weights, but the altered fused content can change the content-routing term and expert selection. This intervention therefore includes downstream routing responses to the altered fusion.

Content-logit permutation. This intervention changes the correspondence between content-routing signals and local positions while retaining the local fused representation and expert input.

Following the routing formulation in the main text, let $\mathbf{c}_i = \mathbf{W}_u \mathbf{u}_i$ and $\mathbf{s}_i = \mathbf{W}_\alpha \boldsymbol{\alpha}_i$ denote the content and scale terms. Only the content term is permuted:

$$\boldsymbol{\pi}_i^{\text{perm}} = \text{softmax}\left(\frac{\mathbf{c}_{\sigma(i)} + \mathbf{s}_i + \mathbf{b}_{\text{gate}}}{\tau}\right). \tag{13}$$

The Top-1 expert and its probability weight are recomputed from the modified routing probabilities. The selected expert still processes the local input $\mathbf{u}_i$, rather than the input at the position supplying the permuted content logits.

The input residual and always-active shared expert are retained. This intervention does not directly shuffle expert indices or force all positions to use a single expert.

Repeats and aggregation. The two interventions are evaluated separately, each using five permutations with seeds 0–4. Within each repeat, the five trained models share the same spatial permutation, applied to each model's own weights or logits.

For each repeat, predictions are first median-ensembled across the five models. RMSE and SSIM are then computed and averaged over the same 313 genes. The reported results are the means of these metrics across the five permutations.

These repeats evaluate permutations with fixed trained models and are not independent training runs.

Results and interpretation. As reported in Table 3, the full model achieves an RMSE of 0.0355 and an SSIM of 0.7215. Fusion-weight permutation changes these values to 0.0368 and 0.7153, whereas content-logit permutation yields 0.0388 and 0.6995.

Both interventions reduce reconstruction quality, supporting the usefulness of matching learned scale contributions and content-routing signals to local inputs.

Content-logit permutation produces a larger degradation in this experiment. Because the interventions act on different computational paths, their magnitudes do not directly rank the independent importance of the two modules.

E.6 Forced-Expert Selection

For each trained model, all valid locations are assigned in turn to E_1, E_2, E_3, or E_4, with hidden widths 128, 256, 512, and 1,024, respectively. Each location retains the original maximum routing probability as the expert-output weight, together with the shared branch and input residual. No retraining is performed. This intervention compares fixed choices with location-dependent selection within the existing trained expert set.

F Spatial Clustering and TLS Analysis

F.1 Multimodal-Representation Clustering

This experiment evaluates whether the super-resolution pipeline supports finer spatial-domain analysis. DAST uses simple K-means on superpixel-level representations, while STAGATE (Dong & Zhang, 2022) and GraphST (Long et al., 2023) identify domains from the original spot data. The two resolutions represent the analysis settings before and after super-resolution processing.

The MBHD clustering input combines the 128-dimensional cross-modal refinement representation, 1,024-dimensional histological features, RGB, and two spatial coordinates, giving 1,157 dimensions. The cross-modal representation precedes the expert decoder and is combined by an elementwise median across the five models. Clustering therefore uses the multimodal representations learned by the super-resolution model rather than the final predicted expression matrix.

The three method maps each display ten clusters. DAST applies PCA, Gaussian spatial smoothing, and k-means to the multimodal representations, using a separate tissue mask.

Colors identify clusters within each method. The Allen Brain Atlas provides a qualitative anatomical reference (Allen Institute for Brain Science) for examining spatial organization and regional boundaries in the spot-level and superpixel-level analyses.

F.2 TLS-Associated Five-Gene Analysis

For the H1_re whole-section data from HER2ST H_1 (Andersson et al., 2021), we use five TLS-associated genes available in both the original expression data and the super-resolved results: CD74, CD79A, CD52, PTGDS, and TRBC2.

The same gene set defines the combined signal at both resolutions for comparing immune-related spatial patterns. Two manually annotated immune-infiltrated regions in H&E provide the spatial reference.

The middle panel of Figure 4 displays the combined signal at measured spot locations. The right panel displays the signal derived from super-resolved expression, with smoothing and background suppression for visualization.

The two panels are normalized separately for display to show the spatial distribution at each resolution.

The comparison examines the locations and shapes of candidate TLS-associated enrichment and their spatial correspondence with the annotated regions.

G Additional Gene-Expression Visualizations

G.1 Additional Comparisons on MBHD

Figure G.1 presents additional comparisons for five genes. DAST preserves the curved expression bands of Atp2b4 and Rimbp2, the narrow band of Ccn2, and the regional expression patterns of Cnr1 and Rgs4.

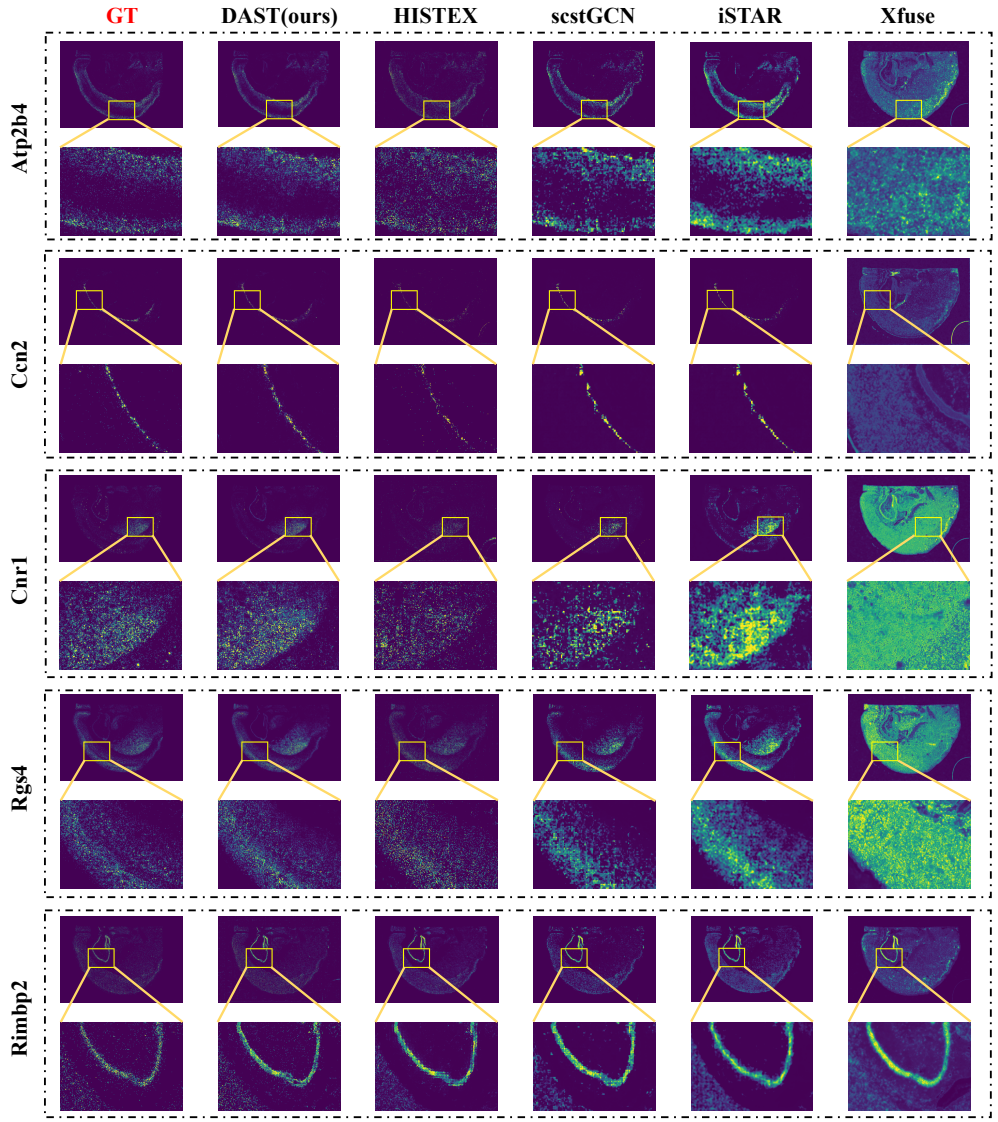

Figure G.1: Additional gene-expression comparisons on MBHD. Gene groups, from top to bottom, are Atp2b4, Ccn2, Cnr1, Rgs4, and Rimbp2. Columns show GT, DAST, HISTEX, scstGCN, iSTAR, and Xfuse, from left to right. Yellow boxes indicate the regions shown in the enlarged views.

G.2 Additional Comparisons on HBC_1

Figure G.2 presents additional reconstructions of LUM and KRT5. DAST closely follows the spatial distribution of LUM and the localized enrichment patterns of KRT5, with the enlarged regions showing local structures consistent with the reference maps.

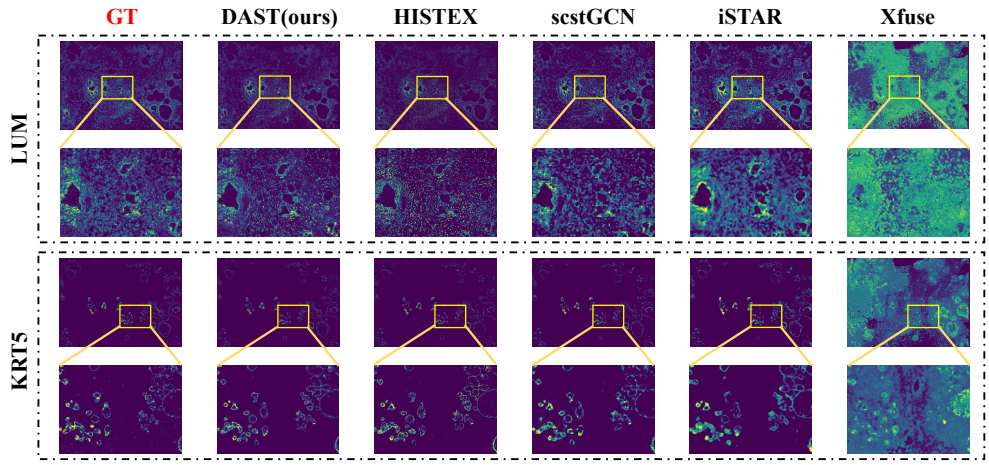

Figure G.2: Additional gene-expression comparisons on HBC_1 for LUM and KRT5. Columns show the high-resolution reference (GT), DAST, HISTEX, scstGCN, iSTAR, and Xfuse, from left to right. For each gene, overview maps are paired with enlarged views of the regions indicated by yellow boxes.