SPATIAL GENE EXPRESSION PREDICTION AS FIELD RECONSTRUCTION FROM HISTOLOGY

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial transcriptomics links gene expression to tissue morphology, but sequencing-based assays remain too costly and labor-intensive for routine profiling at scale. This has motivated a growing line of work on inferring spatial expression from standard hematoxylin-and-eosin (H&E) histology. Current methods, however, remain largely tied to assay-indexed prediction: they estimate expression at measured locations, with spatial context primarily used to inform predictions defined at those locations. This view is misaligned with how molecular signal is organized in tissue. Gene expression is not a set of isolated measurements, but a morphology-conditioned expression field shaped by regional tissue structure and resolved through local transcriptional variation. We formulate histology-based spatial gene expression prediction as *field reconstruction*. **Carta** reconstructs a morphology-conditioned expression field over the measured spatial coordinates of a spatial transcriptomics assay. It first estimates regional expression states from histology and then refines local expression by propagating information across nearby measured locations. To move beyond losses evaluated independently at sampled coordinates, Carta uses a bidirectional conditional transport objective that shapes the distributional structure of the reconstructed expression field across the spatial hierarchy. Across 16 organ-specific prediction tasks — 10 based on HEST-1k and 6 constructed from STImage-1K4M, Carta improves average PCC by 12.9 and 2.3 absolute PCC points over the UNI pathology foundation model baseline and the strongest baseline. It is particularly effective for sparse genes and spatially localized expression patterns, whose weak but structured variations can be under-represented by coordinate-level regression. These results show that morphology-conditioned field reconstruction provides a principled formulation for spatial expression prediction when direct molecular profiling is impractical.

1 INTRODUCTION

Spatial transcriptomics (ST) measures gene expression while preserving the spatial organization of tissue (Ståhl et al., 2016; Xiao & Yu, 2021; Zeng et al., 2022; Jia et al., 2024). By linking molecular activity to histological structure, ST has become a powerful tool for studying developmental programs, tumor microenvironments, and disease progression (Van Dam et al., 2018; Wu et al., 2024; Li et al., 2024b). Yet sequencing-based ST assays remain expensive, labor-intensive, and difficult to deploy routinely at scale (Shi et al., 2024a; Williams et al., 2022; Ståhl et al., 2016). Hematoxylin-and-eosin (H&E) stained histology, in contrast, is inexpensive, standardized, and widely available in both research and clinical workflows. This gap has motivated a growing line of work on inferring spatial gene expression directly from histology images.

Existing methods approach this problem from several directions. Early approaches learn supervised mappings from local image patches to expression values at measured locations (He et al., 2020; Monjo et al., 2022). Later methods incorporate neighborhood graphs (Zeng et al., 2022; Mejia et al., 2023; Yang et al., 2024; Wang et al., 2025; Niu et al., 2026; Sun et al., 2025; Ganguly et al., 2025) or multi-resolution whole-slide features (Chung et al., 2024) to capture broader morphology. Retrieval-based methods estimate expression by aggregating profiles from visually similar reference regions (Xie et al., 2023; Shi et al., 2024b; Min et al., 2024; Chen et al., 2025; Wu et al., 2025), while generative approaches based on diffusion (Li et al., 2024a; Zhu et al., 2025; Liu et al., 2025)

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.

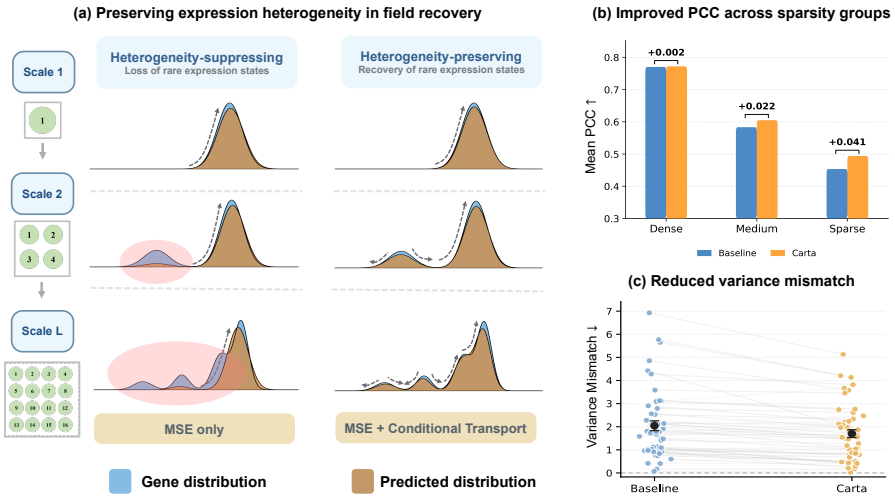

Figure 1: **Carta improves the recovery of sparse expression patterns beyond coordinate-level supervision.** (a) For sparse expression profiles, weak or localized variations can contribute little to MSE relative to larger residuals elsewhere. Conditional transport complements point-wise regression by shaping the distributional structure of the predicted expression field, improving the recovery of sparse expression states and local heterogeneity. (b) The benefit is reflected in real data: compared with STFlow, Carta improves mean PCC across all sparsity groups, with larger gains for medium and sparse genes. (c) Carta also reduces variance mismatch between predicted and measured expression, suggesting improved recovery of gene-wise expression variability.

or flow matching (Huang et al., 2025) model conditional expression distributions given histology. Together, these studies show that tissue morphology carries recoverable molecular signal, and that spatial information is essential for extracting it.

Despite these architectural differences, the supervised target often remains expression at locations measured by the spatial assay (He et al., 2020; Mejia et al., 2023; Xie et al., 2023; Pang et al., 2021; Monjo et al., 2022). Even when multiple locations are processed jointly (Zhu et al., 2025; Huang et al., 2025; Chung et al., 2024), the loss is usually evaluated at the measured coordinates. The task is therefore still anchored to the assay's sampled locations. Gene expression in tissue is organized differently: it varies as a morphology-conditioned field, shaped by regional tissue structure and resolved through local cellular neighborhoods and transcriptional variation.

Coordinate-level losses supervise the field only through values at sampled locations. They therefore treat two different errors in the same way: errors that change individual expression vectors and errors that distort the spatial organization of the field (Zhu et al., 2025; Huang et al., 2025; He et al., 2020). A model can reduce average error at the measured locations while attenuating sparse genes, blurring boundaries between tissue regions, and washing out molecular niches spread across nearby locations (Zhu et al., 2025; Xie et al., 2023; Mathieu et al., 2015). For sparse expression profiles, weak or localized deviations may contribute only marginally to the aggregate coordinate-level objective relative to larger errors elsewhere in the field. Consequently, fitting sampled values alone does not ensure that the spatial organization of the reconstructed field is faithfully preserved (Ledig et al., 2017; Zhao et al., 2016). As illustrated in Figure 1, this limitation becomes particularly relevant at finer spatial scales, where increasingly localized expression variation must be resolved.

We therefore formulate spatial gene expression prediction from histology as *field reconstruction*. The measured ST locations are treated as irregular samples from an underlying morphology-conditioned expression field, and the goal is to reconstruct that field over the measured spatial layout of the assay. In this formulation, spatial organization is not only context for local prediction; it is part of the reconstruction target.

In this paper, we introduce **Carta**, which instantiates this field-reconstruction view through a regional-to-local prediction factorization. Carta first infers regional expression states from tissue morphology and uses them to condition progressively finer expression predictions. It then refines local expression

by propagating information among nearby locations, allowing broad tissue context to inform fine-grained variation. To reduce field-level distortions left by coordinate-level losses, Carta uses a bidirectional conditional transport objective (Tanwisuth et al., 2021) that shapes the distributional structure of the reconstructed field across the spatial hierarchy. Across 16 organ-specific prediction tasks, *i.e.*, 10 based on HEST (Jaume et al., 2024) and 6 constructed from STImage-1K4M (Chen et al., 2024a), Carta consistently improves histology-based spatial gene expression prediction over strong baselines. It is particularly effective for sparse genes and spatially localized expression patterns, for which weak but structured variations can be under-represented by coordinate-level regression. Together, these results show that morphology-conditioned field reconstruction provides a principled formulation for spatial expression prediction when direct molecular profiling is impractical. Our contributions are as follows:

- We formulate histology-based spatial gene expression prediction as morphology-conditioned field reconstruction, shifting the modeling perspective and prediction factorization from assay-indexed measurements to an underlying molecular field sampled on irregular spatial layouts.
- We introduce Carta, which predicts regional expression states and propagates them to condition progressively finer expression estimates over the observed sampling geometry.
- We develop a bidirectional conditional transport objective for shaping the distributional structure of the reconstructed field, encouraging the reconstructed landscape to preserve both regional organization and local heterogeneity beyond coordinate-level prediction error.
- We demonstrate consistent improvements over strong baselines across 16 organ-specific prediction tasks, tissues, and sequencing platforms, with especially large gains on sparse genes and spatially localized expression patterns.

2 Related Work

Spatial gene expression prediction from histology images. Histology-based spatial gene expression prediction has been studied along several directions (Huang et al., 2025). Early methods predict expression from local tissue morphology, typically by extracting image patches around measured locations and training supervised predictors on the corresponding expression profiles (Xie et al., 2023; He et al., 2020). Other methods expand the visual field of view with neighboring locations, graph structure, slide-level representations, or multi-resolution whole-slide features, allowing predictions to use broader tissue context (Pang et al., 2021; Chung et al., 2024). More recent work has introduced retrieval-based estimation from visually similar reference regions and generative modeling through diffusion or flow-based conditional distributions (Huang et al., 2025; Pang et al., 2021; Zhu et al., 2025; He et al., 2020; Xie et al., 2023). These methods differ in architecture and learning objective, but they are usually trained around the same observed quantity: expression values at locations measured by the assay. Spatial information is incorporated through patches, graphs, slide context, retrieval sets, or generative priors, yet the target remains a set of sampled measurements (Zhu et al., 2025; Xie et al., 2023; He et al., 2020). In contrast, Carta changes how predictions are factorized over the sampled geometry. Rather than using spatial structure primarily to contextualize predictions on a fixed support, Carta constructs a sequence of coordinate-derived supports, predicts regional expression states, and propagates these predicted states to condition progressively finer layouts.

Hierarchical autoregressive generation in vision. Autoregressive visual models factorize image generation into a sequence of conditional predictions (Tian et al., 2024; Sun et al., 2024; Esser et al., 2021). Early methods often operate on discrete visual tokens and generate them under a fixed order (Esser et al., 2021; Yu et al., 2022). Recent work moves beyond a single flat sequence by introducing multi-scale tokenizations or scale-wise prediction orders, so that global layout is modeled before local texture and detail (Tian et al., 2024; Ren et al., 2024). In this line of work, the ordering is part of the model: it determines which information is available when later predictions are made. In Carta, the ordering is defined by a coordinate-derived spatial hierarchy rather than image resolution. Spatial gene expression is organized across tissue regions and local cellular neighborhoods (Zhu et al., 2025; Williams et al., 2022). Carta first predicts coarse regional gene-expression states from histology and propagates these states across the spatial hierarchy to condition progressively finer gene-expression estimates at measured locations. This regional-to-local factorization enables local predictions to incorporate regional molecular context while preserving fine-scale spatial variation.

Conditional transport and distributional structure. Most histology-based expression prediction methods are trained with paired regression losses, since each measured location has a corresponding expression vector. This supervision is natural, but it only compares predictions and observations at matched samples. At the level of a reconstructed field, what remains unconstrained is how predicted profiles are distributed within and across tissue regions. A model can fit many paired profiles while under-representing the tails of the expression distribution, especially for sparse genes, low-expression programs, or region-specific heterogeneity (Bishop, 1994; Guilhoto et al., 2026; Zhu et al., 2025). Conditional transport models directional conditional assignments between two sets, providing an efficient alternative to a balanced transport coupling with prescribed shared marginals (Li et al., 2023a; Tanwisuth et al., 2021; Zheng & Zhou, 2021). In histology-based spatial expression prediction, however, transport objectives have rarely been used to model the structure of the reconstructed expression field (Ren et al., 2025; Feng et al., 2024; Halmos et al., 2025; Liu et al., 2019). Motivated by this, we adapt the bidirectional conditional transport construction to relate node prediction features to gene-specific anchors defined by the shared expression decoder. BCT is applied across the prediction hierarchy, acting on regional-node features at coarse scales and spot-level features at the finest scale. This gene-anchored regularization complements coordinate-level regression and provides larger gains for sparse genes while improving field-level reconstruction quality.

3 METHOD

A spatial transcriptomics assay measures gene expression at a finite, irregular set of spatial locations. For histology-based prediction, each measured coordinate is paired with an H&E patch centered at the same location. We model these paired observations as samples from a morphology-conditioned expression field, rather than as independent labels attached to spatial coordinates.

Carta estimates this field by separating regional state estimation from local prediction. At coarse scales, regional nodes aggregate morphology over nearby measured locations and produce expression states that capture broad tissue context. At finer scales, predictions at measured coordinates are conditioned on these regional states and updated using neighboring locations. Training uses coordinate-level regression for the measured expression values, together with a bidirectional conditional transport objective that shapes the distributional structure of the reconstructed field. The overall framework is illustrated in Figure 2. Additional method details are provided in Appendix B. We next define the field reconstruction problem, describe the regional-to-local predictor, and introduce the bidirectional conditional transport objective.

3.1 FIELD RECONSTRUCTION FORMULATION

Consider a whole-slide histology image paired with a spatial transcriptomics sample. Let $\mathbf{C} = \{c_i\}_{i=1}^N$ denote the measured spatial coordinates, with $c_i \in \mathbb{R}^2$, and let $\mathbf{Y} = \{y_i\}_{i=1}^N$ denote the corresponding expression profiles, with $y_i \in \mathbb{R}^G$ for G target genes. For each coordinate c_i, we extract a local H&E patch x_i and write $\mathbf{X} = \{x_i\}_{i=1}^N$.

We model these observations as samples from a morphology-conditioned expression field,

$$\mathcal{F}_{\mathbf{X}} : \Omega \to \mathbb{R}^G, \tag{1}$$

where $\Omega \subset \mathbb{R}^2$ is the tissue domain and $y_i = \mathcal{F}_{\mathbf{X}}(c_i)$ is the measured field value. Given the patch set and measured coordinates, Carta estimates field values at the assayed locations:

$$\hat{\mathbf{Y}} = \hat{\mathcal{F}}_\theta(\mathbf{X}, \mathbf{C}), \qquad \hat{y}_i = \hat{\mathcal{F}}_\theta(\mathbf{X}, \mathbf{C})_i. \tag{2}$$

This highlights that prediction at one location may depend on the full histology-coordinate set, not only on the local patch. Carta represents the field on a sequence of coarse-to-fine spatial layouts $\{\mathcal{V}^{(1)}, \ldots, \mathcal{V}^{(L)}\}$, where $\mathcal{V}^{(1)}$ contains regional nodes and $\mathcal{V}^{(L)}$ corresponds to the original measured coordinates. The model produces estimates: $\{\hat{\mathbf{Y}}^{(1)}, \ldots, \hat{\mathbf{Y}}^{(L)}\}$, with $\hat{\mathbf{Y}}^{(L)} = \hat{\mathbf{Y}}$.

3.2 REGIONAL-TO-LOCAL RECONSTRUCTION

Carta builds the spatial hierarchy from the measured coordinates rather than from an image grid. We encode each H&E patch with a pathology foundation model (PFM) (Ciga et al., 2022; Xu et al., 2024;

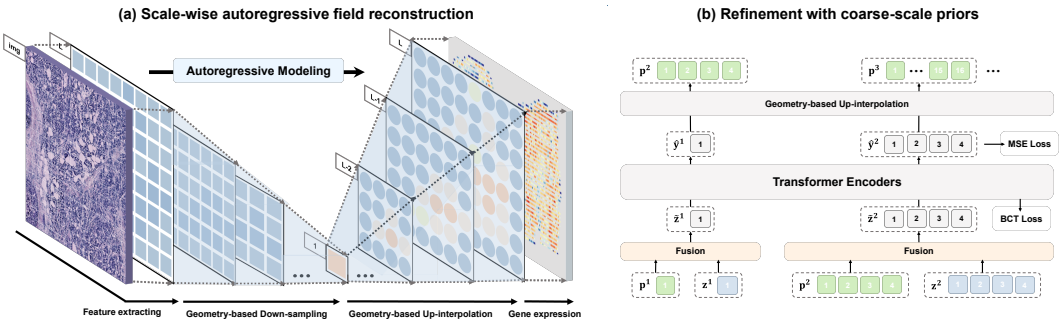

Figure 2: **Overview of Carta.** Carta builds a spatial hierarchy from measured ST coordinates and reconstructs expression from regional states to local predictions. At each scale, propagated expression states condition local histology features; the final prediction is supervised at the measured coordinates, while BCT shapes the distributional structure of the reconstructed field.

Chen et al., 2024b):

$$h_i = f_{\mathrm{PFM}}(x_i), \qquad h_i \in \mathbb{R}^{d_h}. \tag{3}$$

For the finest scale, each node corresponds to one measured ST location, so $\mathcal{V}^{(L)} = \{(h_i, c_i)\}_{i=1}^{N}$. For each coarser scale $l < L$, we partition the normalized coordinate frame into bins and keep only bins containing measured locations. The j-th retained bin defines a regional node $R_j^{(l)}$, with feature and coordinate:

$$h_j^{(l)} = \frac{1}{|R_j^{(l)}|} \sum_{i \in R_j^{(l)}} h_i, \qquad c_j^{(l)} = \frac{1}{|R_j^{(l)}|} \sum_{i \in R_j^{(l)}} c_i. \tag{4}$$

All distances are computed in the normalized coordinate frame; we keep the notation c_i for simplicity. Additional binning details are given in Appendix B.1. For $l \geq 2$, expression states are transferred from scale $l-1$ to scale l by distance-weighted interpolation. For node i at scale l, let $\mathcal{N}_i^{(l)}$ denote its nearest nodes from scale $l-1$, with $|\mathcal{N}_i^{(l)}| = \min(K, N_{l-1})$; if fewer than K valid nodes exist, all of them are used. The interpolated expression state is:

$$p_i^{(l)} = \sum_{j \in \mathcal{N}_i^{(l)}} w_{ij}^{(l)} \hat{y}_j^{(l-1)}, \tag{5}$$

where,

$$w_{ij}^{(l)} = \frac{\exp\left(-\|c_i^{(l)} - c_j^{(l-1)}\|_2^2 / \lambda_u\right)}{\sum_{j' \in \mathcal{N}_i^{(l)}} \exp\left(-\|c_i^{(l)} - c_{j'}^{(l-1)}\|_2^2 / \lambda_u\right)}. \tag{6}$$

Here $\lambda_u > 0$ is the interpolation bandwidth. At the coarsest scale, Carta initializes the expression condition with a learnable gene-wise Gaussian prior:

$$p^{(1)} = \mu + \exp(\log \sigma) \odot \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I), \tag{7}$$

where $\mu, \log \sigma \in \mathbb{R}^G$. For every coarsest-scale node i, we set $p_i^{(1)} = p^{(1)}$ during training and use $p_i^{(1)} = \mu$ at evaluation unless otherwise specified. Further details of cross-scale expression-state propagation and coarsest-scale initialization are provided in Appendix B.2.

Conditional prediction at each scale. At scale l, Carta predicts expression from the histology representation at the current node and the expression condition $p_i^{(l)}$. We project both into the model latent space:

$$z_i^{(l)} = W_h h_i^{(l)} \in \mathbb{R}^d, \qquad e_i^{(l)} = W_p p_i^{(l)} \in \mathbb{R}^d.$$

Let $\mathcal{M}_i^{(l)}$ denote the within-scale neighborhood of node i. Carta conditions the histology state on nearby expression conditions by local cross-attention:

$$\tilde{z}_i^{(l)} = z_i^{(l)} + \mathrm{CrossAttn}\left(q = z_i^{(l)}, K = \{e_j^{(l)} : j \in \mathcal{M}_i^{(l)}\}, V = \{e_j^{(l)} : j \in \mathcal{M}_i^{(l)}\}\right). \tag{8}$$

A local Transformer encoder then produces the contextualized node representation

$$u_i^{(l)} = \mathrm{LocalTransformer}\big(\{\tilde{z}_j^{(l)} : j \in \mathcal{M}_i^{(l)}\}\big)_i. \tag{9}$$

The local attention operator follows STFlow (Huang et al., 2025); its geometry-aware implementation is described in Appendix B.3. A two-layer prediction head maps $u_i^{(l)}$ to expression values. We obtain an intermediate prediction feature $f_i^{(l)} = \phi_1(u_i^{(l)}) \in \mathbb{R}^{d_f}$ and predict expression by:

$$\hat{y}_i^{(l)} = \psi\Big(W_o f_i^{(l)} + b_o\Big), \tag{10}$$

where $\phi_1(\cdot)$ is the intermediate projection, $W_o \in \mathbb{R}^{G\times d_f}$, and $b_o \in \mathbb{R}^G$. The activation ψ is $\mathrm{softplus}$ for non-negative expression targets and the identity map for standardized targets.

Predictions at intermediate scales are passed to the next scale by the interpolation operator above. At the finest scale, the nodes are the original measured coordinates, and $\hat{\mathbf{Y}}^{(L)}$ is the final reconstructed field.

3.3 Bidirectional Conditional Transport Objective

Coordinate-level regression matches each predicted expression vector to the measured vector at the same location, but it does not constrain how gene-specific responses are distributed across the reconstructed field. Carta therefore adds a Bidirectional Conditional Transport (BCT) objective (Tanwisuth et al., 2021), which relates node prediction features to gene anchors defined by the output layer.

Let $\mathbf{F}^{(l)} = \{f_i^{(l)}\}_{i=1}^{N^{(l)}}$ denote the node prediction features at scale l, where $N^{(l)} = |\mathcal{V}^{(l)}|$. We use the rows of the shared output projection W_o as gene anchors:

$$\boldsymbol{g} = \{g_k\}_{k=1}^G, \qquad g_k = \mathrm{sg}(W_o[k,:])\,.$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient within BCT; W_o remains trainable through the expression regression objective. The similarity between gene anchor k and node feature i is: $S_{ki}^{(l)} = {g_k}^\top f_i^{(l)}$, and the transport cost is the cosine distance:

$$D_{ki}^{(l)} = 1 - \frac{{g_k}^\top f_i^{(l)}}{\|g_k\|_2 \|f_i^{(l)}\|_2}. \tag{11}$$

To account for uneven gene activity in a minibatch, BCT maintains a dynamic prior $\mathbf{q} \in \Delta^{G-1}$ over gene anchors, where $\Delta^{G-1} = \{\mathbf{q} \in \mathbb{R}_+^G : \sum_{k=1}^G q_k = 1\}$. The prior is initialized uniformly and updated online from batch matching statistics; details are provided in Appendix B.4. Given $\mathbf{q}$, the node-to-anchor transport weights are:

$$\pi_{ki}^{s,(l)} = \mathrm{softmax}_k\left(\frac{S_{ki}^{(l)}}{\tau} + \log(q_k + \delta)\right), \tag{12}$$

where τ is the temperature and $\delta > 0$ is used for numerical stability. The node-to-anchor loss is:

$$\mathcal{L}_s^{(l)} = \frac{\alpha}{N^{(l)}} \sum_{i=1}^{N^{(l)}} \sum_{k=1}^{G} D_{ki}^{(l)} \pi_{ki}^{s,(l)}. \tag{13}$$

The reverse transport weights assign each gene anchor to spatial nodes:

$$\pi_{ki}^{t,(l)} = \mathrm{softmax}_i\left(\frac{S_{ki}^{(l)}}{\tau}\right), \tag{14}$$

with loss,

$$\mathcal{L}_t^{(l)} = (1-\alpha) \sum_{k=1}^{G} q_k \sum_{i=1}^{N^{(l)}} D_{ki}^{(l)} \pi_{ki}^{t,(l)}. \tag{15}$$

Here $\alpha \in [0,1]$ balances the two transport directions. The BCT loss at scale l is $\mathcal{L}_{\mathrm{BCT}}^{(l)} = \mathcal{L}_s^{(l)} + \mathcal{L}_t^{(l)}$, and the multi-scale regularizer is: $\mathcal{L}_{\mathrm{BCT}} = \frac{1}{L}\sum_{l=1}^{L} \mathcal{L}_{\mathrm{BCT}}^{(l)}$. Regression supervision is applied at the finest scale: $\mathcal{L}_{\mathrm{reg}} = \mathrm{MSE}(\hat{\mathbf{Y}}^{(L)}, \mathbf{Y})$. The final training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{reg}} + \lambda_{\mathrm{BCT}} \mathcal{L}_{\mathrm{BCT}}, \tag{16}$$

where λ_{BCT} controls the strength of transport regularization.

4 EXPERIMENTS

In this section, we evaluate Carta on spatial gene expression prediction from H&E histology. Our experiments cover two paired histology–spatial-transcriptomics resources and compare Carta against methods that differ in visual representation, use of reference data, spatial context modeling, and generative modeling. Additional analyses in Appendix C compare Carta with recent context-aware histology-to-ST methods (C.1) and examine the predictive capacity of frozen pathology features (C.2). We then evaluate prediction at held-out spatial coordinates (C.3), spatial field diagnostics (C.4), and sparsity-wise performance (C.5). Further analyses examine scale-schedule selection (C.6), cross-scale state propagation (C.7), cross-scale fusion (C.8), coarsest-scale initialization (C.9), BCT hyperparameter sensitivity (C.10), gene-anchor design (C.11), and the relationship between gene-anchor geometry and gene co-expression (C.12). Finally, we report computational efficiency (C.13), with qualitative analyses provided in Appendix D.

4.1 SETUP

Datasets. We evaluate on two H&E-paired spatial transcriptomics resources: `HEST-1k` (Jaume et al., 2024) and `STImage-1K4M` (Chen et al., 2024a). From these resources, we define 16 prediction tasks: 10 cancer-specific tasks from `HEST-1k` and 6 organ-level tasks constructed from `STImage-1K4M`. Together, these tasks span diverse tissue morphologies, disease contexts, sequencing platforms, and expression patterns.

The `HEST-1k` tasks include IDC, PRAD, PAAD, SKCM, COAD, READ, CCRCC, HCC, LUNG, and LYMPH. For `HEST-1k`, the performance is evaluated using the patient-disjoint cross-validation setup of prior work (Huang et al., 2025; Jaume et al., 2024). The `STImage-1K4M` tasks include Brain, Breast, Heart, Liver, Pancreas, and Skin. For each `STImage-1K4M` task, we construct a patient-disjoint 80/20 train/test split at the slide level, ensuring that slides from the same patient do not appear in both partitions. For all methods, architecture and hyperparameters are selected on `HEST-1k` and then fixed for all `STImage-1K4M` tasks, without further hyperparameter tuning. For each evaluation split, we apply a per-spot log1p transformation and select the top 50 highly variable genes using only the training partition. The resulting gene set is then fixed for all compared methods on that split. All methods are evaluated using identical HEST patient-disjoint folds or STImage patient-disjoint partitions. The final evaluation contains 800 task-specific gene targets across 16 prediction tasks. Further dataset statistics, platform compositions, and evaluation protocol details are provided in Appendix A.2.

Baselines. We compare Carta with nine baselines: Ciga (Ciga et al., 2022), Gigapath (Xu et al., 2024), UNI (Chen et al., 2024b), BLEEP (Xie et al., 2023), DANet (Wu et al., 2025), HisToGene (Pang et al., 2021), Stem (Zhu et al., 2025), STFlow (Huang et al., 2025), and STNet (He et al., 2020). These baselines cover four modeling families: pathology-foundation-model feature regression (Ciga et al., 2022; Xu et al., 2024; Chen et al., 2024b), image–expression alignment (Xie et al., 2023; Wu et al., 2025), task-specific neural regression models (Pang et al., 2021; He et al., 2020), and generative expression modeling (Zhu et al., 2025; Huang et al., 2025). Further comparisons with recent methods that incorporate broader tissue context are provided in Appendix C.1. For Ciga, Gigapath, and UNI, we extract histology features at measured locations using the corresponding pathology foundation model and train a Random Forest regressor for gene expression prediction. Further analyses with additional prediction heads are provided in Appendix C.2 to assess the predictive capacity of the frozen pathology features. For methods that require a pathology image encoder, we use UNI as the default encoder unless otherwise specified. In particular, Stem, STFlow, and Carta all use UNI features at measured locations. This controls for the visual backbone and makes the comparison focus on how each method models spatial expression from the same histology representation. Additional preprocessing details, baseline settings, and training configurations are provided in Appendix A.1.

4.2 GENE EXPRESSION PREDICTION

Overall prediction performance. Table 1 reports task-level PCC on HEST and STImage-1K4M. Carta attains the highest average PCC on both resources, reaching 0.425 on HEST and 0.653 on STImage-1K4M. Compared with the strongest baseline, STFlow, Carta improves average PCC by 0.024 and 0.021, respectively. The gains are broad across tasks: Carta ranks first on all ten HEST

Table 1: **Organ-wise comparison across two datasets.** Methods are arranged in columns, and PCC is reported for each organ. Subscripts indicate the standard deviation across multiple random seeds. The best results are highlighted in bold, and the second-best results are underlined.

Dataset	Organ	Ciga	Gigapath	UNI	BLEEP	DANet	HisToGene	Stem	STFlow	STNet	Carta
HEST	CCRCC	$0.126_{.001}$	$0.182_{.001}$	$0.186_{.000}$	$0.286_{.001}$	$0.251_{.002}$	$0.155_{.011}$	$0.242_{.019}$	$\underline{0.329}_{.022}$	$0.166_{.006}$	$\mathbf{0.335}_{.007}$
	COAD	$0.250_{.000}$	$0.261_{.003}$	$0.295_{.001}$	$0.289_{.004}$	$0.225_{.003}$	$0.215_{.019}$	$0.309_{.006}$	$\underline{0.327}_{.017}$	$0.276_{.003}$	$\mathbf{0.363}_{.008}$
	HCC	$0.043_{.001}$	$0.054_{.000}$	$0.052_{.001}$	$0.057_{.006}$	$0.061_{.003}$	$0.038_{.007}$	$0.064_{.007}$	$\underline{0.121}_{.010}$	$0.097_{.006}$	$\mathbf{0.132}_{.006}$
	IDC	$0.405_{.000}$	$0.491_{.001}$	$0.487_{.002}$	$0.521_{.005}$	$0.538_{.001}$	$0.311_{.012}$	$0.517_{.022}$	$\underline{0.547}_{.020}$	$0.407_{.006}$	$\mathbf{0.595}_{.009}$
	LUNG	$0.514_{.000}$	$0.546_{.002}$	$0.512_{.001}$	$0.597_{.002}$	$0.545_{.003}$	$0.412_{.005}$	$0.572_{.012}$	$\underline{0.606}_{.006}$	$0.566_{.003}$	$\mathbf{0.621}_{.004}$
	LYMPH	$0.217_{.001}$	$0.248_{.001}$	$0.233_{.000}$	$0.221_{.002}$	$0.255_{.002}$	$0.242_{.013}$	$0.238_{.018}$	$0.260_{.007}$	$\underline{0.264}_{.002}$	$\mathbf{0.315}_{.005}$
	PAAD	$0.396_{.000}$	$0.426_{.000}$	$0.423_{.000}$	$0.407_{.008}$	$0.387_{.002}$	$0.234_{.011}$	$0.420_{.012}$	$\underline{0.486}_{.009}$	$0.427_{.009}$	$\mathbf{0.516}_{.010}$
	PRAD	$0.331_{.001}$	$0.370_{.000}$	$0.357_{.002}$	$0.369_{.005}$	$0.384_{.004}$	$0.258_{.009}$	$0.350_{.010}$	$\underline{0.400}_{.008}$	$0.372_{.005}$	$\mathbf{0.422}_{.004}$
	READ	$0.046_{.000}$	$0.155_{.001}$	$0.163_{.000}$	$0.236_{.024}$	$0.205_{.022}$	$0.031_{.018}$	$0.227_{.017}$	$\underline{0.240}_{.008}$	$0.158_{.029}$	$\mathbf{0.245}_{.033}$
	SKCM	$0.484_{.000}$	$0.540_{.002}$	$0.613_{.001}$	$0.518_{.008}$	$0.611_{.001}$	$0.233_{.008}$	$0.652_{.005}$	$\underline{0.695}_{.004}$	$0.439_{.010}$	$\mathbf{0.703}_{.006}$
	Avg	0.281	0.327	0.332	0.350	0.346	0.213	0.359	$\underline{0.401}$	0.317	**0.425**
STImage	Brain	$0.410_{.001}$	$0.463_{.001}$	$0.483_{.002}$	$0.497_{.006}$	$0.535_{.009}$	$0.232_{.004}$	$0.515_{.016}$	$\underline{0.568}_{.002}$	$0.474_{.009}$	$\mathbf{0.586}_{.016}$
	Breast	$0.489_{.000}$	$0.523_{.002}$	$0.486_{.001}$	$0.442_{.007}$	$0.433_{.012}$	$0.391_{.013}$	$0.513_{.018}$	$\underline{0.535}_{.023}$	$0.454_{.012}$	$\mathbf{0.595}_{.013}$
	Heart	$0.581_{.000}$	$0.643_{.001}$	$0.638_{.002}$	$0.632_{.005}$	$0.629_{.004}$	$0.562_{.006}$	$0.689_{.007}$	$\mathbf{0.768}_{.004}$	$0.693_{.011}$	$\underline{0.763}_{.006}$
	Liver	$0.407_{.001}$	$0.424_{.001}$	$0.434_{.000}$	$0.566_{.006}$	$0.572_{.005}$	$0.543_{.013}$	$0.586_{.009}$	$0.615_{.012}$	$\underline{0.618}_{.008}$	$\mathbf{0.659}_{.022}$
	Pancreas	$0.543_{.002}$	$0.608_{.001}$	$0.585_{.001}$	$0.713_{.011}$	$0.754_{.013}$	$0.511_{.018}$	$0.739_{.030}$	$\underline{0.798}_{.005}$	$0.623_{.014}$	$\mathbf{0.806}_{.017}$
	Skin	$0.331_{.001}$	$0.343_{.000}$	$0.302_{.002}$	$0.487_{.012}$	$0.450_{.008}$	$0.351_{.007}$	$0.444_{.013}$	$\underline{0.508}_{.004}$	$0.424_{.010}$	$\mathbf{0.509}_{.005}$
	Avg	0.460	0.501	0.488	0.556	0.562	0.432	0.581	$\underline{0.632}$	0.548	**0.653**

cancer-specific tasks and on five of the six STImage-1K4M organ-level tasks, with Heart as the only task where it ranks second.

These results demonstrate the efficacy of field reconstruction. Because Carta and STFlow use the same UNI histology features (Chen et al., 2024b), their comparison isolates the expression model rather than the visual backbone. Stem predicts expression from local histology patches, while STFlow adds slide-level modeling through local spatial attention and flow matching. Carta changes how expression is predicted: regional expression states are estimated first and then used to condition local predictions at the measured coordinates. The improvement over STFlow therefore reflects the value of regional-to-local field reconstruction, rather than a stronger image encoder. This interpretation is further supported by the sparse-gene analysis in Appendix C.5, where Carta shows clearer gains on genes that are difficult to recover with coordinate-level supervision alone.

Biomarker prediction. We further evaluate Carta on six marker genes: UBE2C (Ma et al., 2023), MET (Gibney et al., 2013), KRTDAP (Tsuchida et al., 2004; Su et al., 2021), PTPRC (Al Barashdi et al., 2021; Li et al., 2023b), GATA3 (Mehra et al., 2005), and ERBB2 (Revillion et al., 1998). Table 2 reports PCC for each gene. Carta obtains the highest PCC on all six markers and reaches an average PCC of 0.842, compared with 0.825 for STFlow, the strongest baseline. Marker genes provide a useful stress test because their expression is often localized to specific tissue regions or cell populations; errors that are diluted in average gene-level metrics can become visible at this scale.

Table 2: **Biomarker prediction.** PCC is reported for six representative biomarkers; subscripts denote standard deviation.

Method	UBE2C	MET	KRTDAP	PTPRC	GATA3	ERBB2
BLEEP	$0.618_{.005}$	$0.769_{.006}$	$0.365_{.001}$	$0.634_{.007}$	$0.789_{.012}$	$0.724_{.009}$
HisToGene	$0.667_{.002}$	$0.561_{.004}$	$0.343_{.001}$	$0.597_{.005}$	$0.721_{.008}$	$0.664_{.003}$
Stem	$0.453_{.014}$	$0.765_{.006}$	$0.820_{.008}$	$0.812_{.005}$	$0.786_{.007}$	$0.754_{.005}$
STFlow	$\underline{0.749}_{.006}$	$\underline{0.816}_{.007}$	$\underline{0.872}_{.016}$	$\underline{0.851}_{.009}$	$\underline{0.837}_{.004}$	$\underline{0.826}_{.007}$
STNet	$0.674_{.004}$	$0.764_{.001}$	$0.433_{.002}$	$0.824_{.003}$	$0.743_{.002}$	$0.707_{.004}$
UNI	$0.701_{.001}$	$0.745_{.002}$	$0.810_{.005}$	$0.746_{.001}$	$0.822_{.003}$	$0.793_{.006}$
Carta	$\mathbf{0.760}_{.004}$	$\mathbf{0.839}_{.002}$	$\mathbf{0.897}_{.011}$	$\mathbf{0.866}_{.005}$	$\mathbf{0.853}_{.007}$	$\mathbf{0.835}_{.006}$

Figure 3 shows representative spatial predictions. For KRTDAP in TENX115, the ground truth contains a localized high-expression region near the lower-right tissue area. Carta recovers this activation while avoiding spurious signal in the low-expression background. For PTPRC in TENX141, Carta preserves the broad high-expression pattern and maintains sharper low-expression boundaries than competing methods.

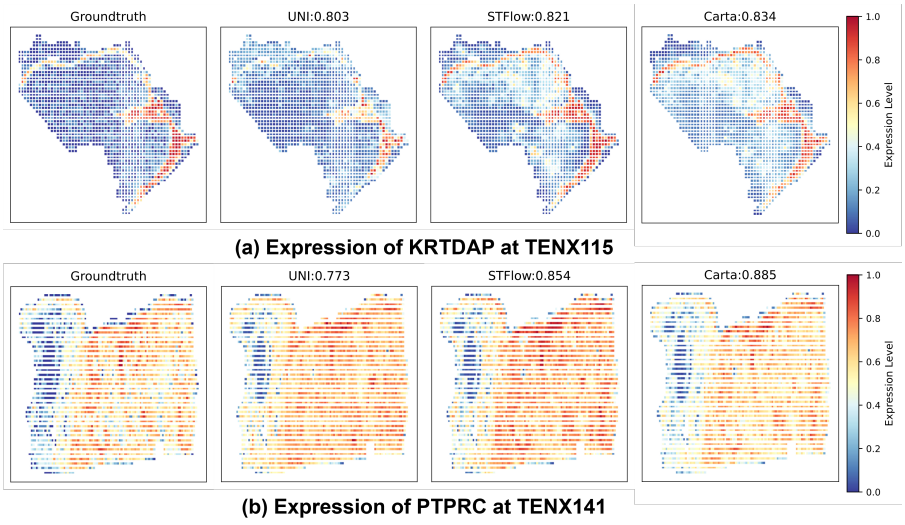

Figure 3: **Biomarker expression maps.** We visualize KRTDAP at TENX115 and PTPRC at TENX141, with PCC reported above each prediction. Carta better preserves localized activation and structured spatial expression patterns.

These visualizations show that the biomarker gains are not only due to higher global correlation. Carta better preserves where marker expression appears, and where it should remain low, which is consistent with reconstructing expression as a spatial field rather than fitting each measured location independently. Additional biomarker expression maps and coarse-to-fine reconstruction visualizations are provided in Appendix D.1 and Appendix D.2.

4.3 Ablation Studies

Table 3 isolates the regional-to-local hierarchy and BCT regularization under the same training protocol. With the UNI encoder fixed, adding BCT to the variant without either component improves the average PCC from 0.398 to 0.406, while adding the multi-scale hierarchy alone yields a larger improvement to 0.413. Combining both components further increases the PCC to 0.425, demonstrating that both components contribute and have complementary effects. The same ordering holds with Ciga and Gigapath features, suggesting that Carta improves expression modeling rather than relying on a specific pathology foundation model.

Table 3: **Component ablation of Carta.** We report the average PCC on HEST.

Variant	Ciga	UNI	Gigapath
w/o MS & BCT	0.340	0.398	0.386
w/o MS	0.346	0.406	0.395
w/o BCT	0.351	0.413	0.401
Carta	**0.363**	**0.425**	**0.412**

5 Conclusion

We formulated spatial gene expression prediction from histology as field reconstruction, where assayed ST locations are treated as irregular observations of a morphology-conditioned expression field. The goal is not only to match expression values at measured coordinates, but to preserve the spatial organization of the field they sample. We presented Carta, which estimates regional expression states from H&E morphology, refines local expression over measured coordinates, and uses bidirectional conditional transport to shape the distributional structure of the reconstructed field. Across 16 organ-specific prediction tasks based on HEST and STImage-1K4M, Carta improves over strong baselines, with gains concentrated on sparse genes and spatially localized expression patterns. These results suggest that a key challenge is not only missing spatial context, but also supervision limited to sampled values. By treating spatial expression as a field, Carta incorporates regional organization and local heterogeneity into the reconstruction process, leading to more faithful spatial expression reconstruction. Limitations and broader impact are discussed in Appendix E.

Reproducibility Statement

We provide comprehensive details of the datasets, preprocessing, model architecture, training procedure, and evaluation protocols in the main text and Appendix. Mathematical formulations, implementation details, hyperparameter settings, and additional experimental results are included in the Appendix. The source code will be released upon acceptance.

AI Use Statement

In accordance with the ICLR policy, we disclose the utilization of Large Language Models (LLMs) in the preparation of this manuscript. The application of these tools was strictly confined to linguistic and formatting support. Specifically, an LLM was employed to proofread the text, correct grammatical errors, and enhance the clarity and readability of the prose. The LLM played no role in any substantive scientific components of this work, including the conception of research ideas, the design of methodologies, the execution or analysis of experiments, or the generation of results and conclusions. All intellectual contributions and the essential content of this paper are exclusively attributable to the authors.

Ethics Statement

This work adheres to the ICLR Code of Ethics. This work is intended to support exploratory spatial profiling and hypothesis generation from histology. Its predictions are computational estimates and should not be treated as measured molecular profiles or used for clinical decision-making without biological validation. Incorrect predictions may mislead downstream interpretation, particularly in clinically sensitive settings. We therefore advocate its responsible use as a research tool and disclose that we have no conflicts of interest.

References

Maryam Ahmed Al Barashdi, Ahlam Ali, Mary Frances McMullin, and Ken Mills. Protein tyrosine phosphatase receptor type c (ptprc or cd45). *Journal of clinical pathology*, 74(9):548–552, 2021.

Christopher M. Bishop. Mixture density networks. Technical Report NCRG/94/004, Aston University, Birmingham, UK, 1994.

Jiawen Chen, Muqing Zhou, Wenrong Wu, Jinwei Zhang, Yun Li, and Didong Li. Stimage-1k4m: A histopathology image-gene expression dataset for spatial transcriptomics. *Advances in neural information processing systems*, 37:35796–35823, 2024a.

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

Weiqing Chen, Pengzhi Zhang, Tu N Tran, Yiwei Xiao, Shengyu Li, Vrutant V Shah, Hao Cheng, Kristopher W Brannan, Keith Youker, Li Lai, et al. A visual–omics foundation model to bridge histopathology with spatial transcriptomics. *Nature Methods*, 22(7):1568–1582, 2025.

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

Ozan Ciga, Tony Xu, and Anne Louise Martel. Self supervised contrastive learning for digital histopathology. *Machine learning with applications*, 7:100198, 2022.

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

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Lan Feng, Fan Nie, Yuejiang Liu, and Alexandre Alahi. Tarot: Targeted data selection via optimal transport. *arXiv preprint arXiv:2412.00420*, 2024.

Aniruddha Ganguly, Debolina Chatterjee, Wentao Huang, Jie Zhang, Alisa Yurovsky, Travis Steele Johnson, and Chao Chen. Merge: Multi-faceted hierarchical graph-based gnn for gene expression prediction from whole slide histopathology images. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15611–15620. IEEE, 2025.

GT Gibney, SA Aziz, RL Camp, P Conrad, BE Schwartz, CR Chen, WK Kelly, and HM Kluger. c-met is a prognostic marker and potential therapeutic target in clear cell renal cell carcinoma. *Annals of Oncology*, 24(2):343–349, 2013.

Leonardo Ferreira Guilhoto, Akshat Kaushal, and Paris Perdikaris. Multimodal scientific learning beyond diffusions and flows. *arXiv preprint arXiv:2602.00960*, 2026.

Peter Halmos, Julian Gold, Xinhao Liu, and Benjamin J Raphael. Hierarchical refinement: Optimal transport to infinity and beyond. *arXiv preprint arXiv:2503.03025*, 2025.

Bryan He, Ludvig Bergenstråhle, Linnea Stenbeck, Abubakar Abid, Alma Andersson, Åke Borg, Jonas Maaskola, Joakim Lundeberg, and James Zou. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature biomedical engineering*, 4(8):827–834, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Tinglin Huang, Tianyu Liu, Mehrtash Babadi, Wengong Jin, and Rex Ying. Scalable generation of spatial transcriptomics from histology images via whole-slide flow matching. *arXiv preprint arXiv:2506.05361*, 2025.

Guillaume Jaume, Paul Doucet, Andrew H Song, Ming Y Lu, Cristina Almagro-Pérez, Sophia J Wagner, Anurag J Vaidya, Richard J Chen, Drew F Williamson, Ahrong Kim, et al. Hest-1k: A dataset for spatial transcriptomics and histology image analysis. *Advances in Neural Information Processing Systems*, 37:53798–53833, 2024.

Yuran Jia, Junliang Liu, Li Chen, Tianyi Zhao, and Yadong Wang. Thitogene: a deep learning method for predicting spatial transcriptomics from histological images. *Briefings in Bioinformatics*, 25(1): bbad464, 2024.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681–4690, 2017.

Kongming Li, Jiahao Li, Yuhao Tao, and Fei Wang. stdiff: a diffusion model for imputing spatial transcriptomics through single-cell transcriptomics. *Briefings in Bioinformatics*, 25(3):bbae171, 2024a.

Miaoge Li, Dongsheng Wang, Xinyang Liu, Zequn Zeng, Ruiying Lu, Bo Chen, and Mingyuan Zhou. Patchct: Aligning patch set and label set with conditional transport for multi-label image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15348–15358, 2023a.

Xiaoyu Li, Fangfang Zhu, and Wenwen Min. Spadit: diffusion transformer for spatial gene expression prediction using scrna-seq. *Briefings in Bioinformatics*, 25(6):bbae571, 2024b.

Xuemei Li, Zhanghui Yue, Dan Wang, and Lu Zhou. Ptprc functions as a prognosis biomarker in the tumor microenvironment of cutaneous melanoma. *Scientific reports*, 13(1):20617, 2023b.

Huidong Liu, Xianfeng Gu, and Dimitris Samaras. Wasserstein gan with quadratic transport cost. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4832–4841, 2019.

Xuepeng Liu, Zheng Jiang, Pinan Zhu, Hanyu Liu, and Chao Li. Hadm-st: Histology-assisted differential modeling for spatial transcriptomics generation. In *International Workshop on Computational Mathematics Modeling in Cancer Analysis*, pp. 129–138. Springer, 2025.

Sijia Ma, Qian Chen, Xu Li, Jing Fu, and Le Zhao. Ube2c serves as a prognosis biomarker of uterine corpus endometrial carcinoma via promoting tumor migration and invasion. *Scientific Reports*, 13 (1):16899, 2023.

Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015.

Rohit Mehra, Sooryanarayana Varambally, Lei Ding, Ronglai Shen, Michael S Sabel, Debashis Ghosh, Arul M Chinnaiyan, and Celina G Kleer. Identification of gata3 as a breast cancer prognostic marker by global gene expression meta-analysis. *Cancer research*, 65(24):11259–11264, 2005.

Gabriel Mejia, Paula Cárdenas, Daniela Ruiz, Angela Castillo, and Pablo Arbeláez. Sepal: spatial gene expression prediction from local graphs. In *Proceedings of the IEEE/CVF International Conference on computer vision*, pp. 2294–2303, 2023.

Wenwen Min, Zhiceng Shi, Jun Zhang, Jun Wan, and Changmiao Wang. Multimodal contrastive learning for spatial gene expression prediction using histology images. *Briefings in Bioinformatics*, 25(6):bbae551, 2024.

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

Yi Niu, Jiashuai Liu, Yingkang Zhan, Jiangbo Shi, Di Zhang, Marika Reinius, Ines Machado, Mireia Crispin-Ortuzar, Jialun Wu, Chen Li, et al. Ph2st: Prompt-guided hypergraph learning for spatial transcriptomics prediction in whole slide images. *Medical Image Analysis*, pp. 104008, 2026.

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

Qin Ren, Yifan Wang, Ruogu Fang, Haibin Ling, and Chenyu You. Otsurv: A novel multiple instance learning framework for survival prediction with heterogeneity-aware optimal transport. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 439–449. Springer, 2025.

Sucheng Ren, Qihang Yu, Ju He, Xiaohui Shen, Alan Yuille, and Liang-Chieh Chen. Flowar: Scale-wise autoregressive image generation meets flow matching. *arXiv preprint arXiv:2412.15205*, 2024.

F Revillion, J Bonneterre, and JP Peyrat. Erbb2 oncogene in human breast cancer and its clinical significance. *European Journal of Cancer*, 34(6):791–808, 1998.

Weixiong Shi, Jing Zhang, Shanqing Huang, Qian Fan, Jiao Cao, Jun Zeng, Lingling Wu, and Chaoyong Yang. Next-generation sequencing-based spatial transcriptomics: A perspective from barcoding chemistry. *JACS Au*, 4(5):1723, 2024a.

Zhiceng Shi, Fangfang Zhu, Changmiao Wang, and Wenwen Min. Spatial gene expression prediction from histology images with stco. In *International Symposium on Bioinformatics Research and Applications*, pp. 89–100. Springer, 2024b.

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

Yushen Su, Zhirui Zeng, Dongyun Rong, Yushi Yang, Bei Wu, and Yu Cao. Psmc2, orc5 and krtdap are specific biomarkers for hpv-negative head and neck squamous cell carcinoma. *Oncology letters*, 21(4):289, 2021.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

Weicheng Sun, Ping Zhang, Jinsheng Xu, Weihan Zhang, Yongbin Zeng, and Li Li. Bridging attention fusion-based multi-view graph neural networks for spatial gene expression prediction. *Pattern Recognition*, pp. 112207, 2025.

Korawat Tanwisuth, Xinjie Fan, Huangjie Zheng, Shujian Zhang, Hao Zhang, Bo Chen, and Mingyuan Zhou. A prototype-oriented framework for unsupervised domain adaptation. *Advances in Neural Information Processing Systems*, 34:17194–17208, 2021.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

Shuichi Tsuchida, Makoto Bonkobara, James R McMillan, Masashi Akiyama, Tatsuo Yudate, Yoshinori Aragane, Tadashi Tezuka, Hiroshi Shimizu, Ponciano D Cruz Jr, and Kiyoshi Ariizumi. Characterization of kdap, a protein secreted by keratinocytes. *Journal of investigative dermatology*, 122(5):1225–1234, 2004.

Sipko Van Dam, Urmo Vosa, Adriaan Van der Graaf, Lude Franke, and Joao Pedro de Magalhaes. Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in bioinformatics*, 19(4):575–592, 2018.

Yuequn Wang, Jun Wang, Yanyu Xu, Ning Liu, Bin Liu, Yuliang Li, and Guoxian Yu. Fmh2st: foundation model-based spatial transcriptomics generation from histological images. *Nucleic Acids Research*, 53(17):gkaf865, 2025.

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

Jiashuo Wu, Xilong Zhao, Yalan He, Bingyue Pan, Jiyin Lai, Miao Ji, Siyuan Li, Junling Huang, and Junwei Han. Idmir: identification of dysregulated mirnas associated with disease based on a mirna–mirna interaction network constructed through gene expression data. *Briefings in Bioinformatics*, 25(4):bbae258, 2024.

Yulong Wu, Jin Xie, Jing Nie, Jiale Cao, Yuansong Zeng, and Zheng Wang. Danet: spatial gene expression prediction from h&e histology images through dynamic alignment. *Briefings in Bioinformatics*, 26(4):bbaf422, 2025.

Yi Xiao and Dihua Yu. Tumor microenvironment as a therapeutic target in cancer. *Pharmacology & therapeutics*, 221:107753, 2021.

Ronald Xie, Kuan Pang, Sai Chung, Catia Perciani, Sonya MacParland, Bo Wang, and Gary Bader. Spatially resolved gene expression prediction from histology images via bi-modal contrastive learning. *Advances in Neural Information Processing Systems*, 36:70626–70637, 2023.

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, 2024.

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

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

Yuansong Zeng, Zhuoyi Wei, Weijiang Yu, Rui Yin, Yuchen Yuan, Bingling Li, Zhonghui Tang, Yutong Lu, and Yuedong Yang. Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. *Briefings in Bioinformatics*, 23(5):bbac297, 2022.

Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on computational imaging*, 3(1):47–57, 2016.

Huangjie Zheng and Mingyuan Zhou. Exploiting chain rule and bayes' theorem to compare probability distributions. *Advances in Neural Information Processing Systems*, 34:14993–15006, 2021.

Sichen Zhu, Yuchen Zhu, Molei Tao, and Peng Qiu. Diffusion generative modeling for spatially resolved gene expression inference from histology images. *arXiv preprint arXiv:2501.15598*, 2025.

APPENDIX

A IMPLEMENTATION DETAILS

This section provides supplementary experimental details for Carta. We first describe the implementation and training configurations, then introduce the baseline methods used for comparison, and finally present the datasets and evaluation protocols.

A.1 TRAINING DETAILS AND BASELINE SETTINGS

All experiments follow the preprocessing pipeline described in the main text. For each measured spatial location, we crop a 224×224 histology patch centered at that location and extract its visual representation using a frozen pathology foundation model. The resulting patch-level features are used as the input to Carta. Unless otherwise specified, we use UNI as the default pathology image encoder.

For all trainable neural models, we search the learning rate over $\{10^{-3}, 5 \times 10^{-4}, 10^{-4}\}$ and the dropout rate over $\{0, 0.2, 0.5\}$ on `HEST-1k` under the same cross-validation protocol. All models are optimized using Adam (Kingma & Ba, 2014) for up to 100 epochs, with early stopping using a patience of 20 epochs. For `HEST-1k`, we follow the patient-disjoint cross-validation and checkpoint-selection protocol of STFlow (Huang et al., 2025) under the official HEST-Bench splits (Jaume et al., 2024). The same procedure is applied to Carta and all re-trained baselines. For `STImage-1K4M`, the architecture and hyperparameters selected on `HEST-1k` are kept fixed without further tuning. All trainable models are optimized using Adam for a fixed 100 epochs under the same training budget, and the final-epoch checkpoint is used for evaluation.

To stabilize training, gradient norms are clipped to 1.0 for each model. All pathology foundation model parameters are frozen during training. All experiments are conducted on a single NVIDIA RTX A6000 GPU. To reduce randomness, we repeat the main experiments with 3 random seeds and report the averaged results.

For Carta, we set the batch size to 2. Unless otherwise specified, we use the default multi-scale schedule $\mathcal{S}_{\text{def}} = \{1^2, 2^2, 4^2, 8^2, 12^2, 16^2, 20^2, N\}$, where N denotes the native spot-level scale of each tissue section. For a scale denoted by s^2, we use a square spatial partition with $S_x^{(l)} = S_y^{(l)} = s$, and retain only non-empty bins as regional nodes. The interpolation bandwidth in cross-scale expression-state propagation is set to $\lambda_u = 0.2$. The model hidden dimension is set to $d = 256$, and the prediction-head intermediate dimension is set to $d_f = 256$. The model uses one cross-attention layer for fusing histology features and propagated expression states, followed by 4 self-attention Transformer blocks. The number of attention heads is set to 4 for both cross-attention and self-attention layers. The number of nearest neighbors is set to 8 for both local attention and cross-scale KNN interpolation.

For the BCT objective, we set the loss weight to $\lambda_{\text{BCT}} = 0.001$, the transport temperature to $\tau = 1$, the numerical stability constant to $\delta = 10^{-6}$, the source-target balancing coefficient to $\alpha = 0.5$, and the momentum coefficient of the dynamic prior to $\beta = 0.05$. The dynamic prior is initialized as a uniform distribution over gene anchors and updated online during training. These values are used as the default BCT configuration unless otherwise specified.

For the Ciga, Gigapath, and UNI feature baselines, we use the official pretrained weights of the corresponding pathology foundation models and train a Random Forest regressor for gene expression prediction. To examine whether their performance is limited by the choice of prediction head, we additionally evaluate linear regression, ridge regression, and a two-layer MLP in Appendix C.2. The number of trees is set to 70. For the remaining baselines, we follow their official implementations whenever available. Stem and STFlow also use UNI features by default and take slide-level inputs with batch size 2. For DANet and BLEEP, we construct the expression library using all training-set gene expression profiles, and use the linear combination of the top-50 retrieved profiles as the prediction. For STNet, we use ResNet-50 (He et al., 2016) as the default image encoder and set the batch size to 128. For HisToGene, we follow the official design and use its ViT-based (Dosovitskiy et al., 2020) encoder with slide-level inputs and batch size 2.

We additionally evaluate TRIPLEX (Chung et al., 2024), FmH2ST (Wang et al., 2025), and MERGE (Ganguly et al., 2025), three recent histology-to-ST methods that incorporate broader

tissue context. For all three additional methods, we use the public implementations and preserve their core spatial or multi-resolution architectures. They are evaluated using the same samples, data partitions, target genes, expression preprocessing, and evaluation metrics as the other baselines. Their quantitative comparison with Carta is reported in Appendix C.1.

Table 4: **Dataset statistics of HEST-Bench.** For each cancer-specific subset, we report the organ type, ST technology, number of patients and samples, evaluation splits, and average number of measured spots.

	IDC	PRAD	PAAD	SKCM	COAD	READ	CCRCC	HCC	LUNG	LYMPH
Organ	Breast	Prostate	Pancreas	Skin	Colon	Rectum	Kidney	Liver	Lung	Axillary Lymph Nodes
Technology	Xenium	Visium	Xenium	Xenium	Visium	Visium	Visium	Visium	Xenium	Visium
#Patients	4	2	3	2	3	2	24	2	2	4
#Samples	4	23	3	2	6	4	24	2	2	4
#Splits	4	2	3	2	2	2	6	2	2	4
Avg. spots	4925	2454	2780	1741	5079	1909	2792	1941	1944	4990

Table 5: **Dataset statistics of our STImage-derived benchmark.** We summarize the ST technology, number of patients and tissue samples, and average number of measured spots for each organ-level subset.

	Brain	Breast	Heart	Liver	Pancreas	Skin
Technology	Visium	Visium / ST	Visium / ST	Visium	Visium	Visium
#Patients	42	84	29	36	31	26
#Samples	90	195	50	44	40	49
Avg. spots	3259	1073	2171	2319	1635	1235

A.2 Dataset

Table 4 presents the dataset statistics of HEST-Bench. We follow the HEST-Bench (Jaume et al., 2024) setting used in prior work, which is constructed from HEST-1k and contains ten organ-specific subsets. HEST-Bench covers diverse tissue types and spatial transcriptomics technologies, including Visium and Xenium. Following the standard evaluation protocol, we adopt patient-stratified cross-validation on HEST-Bench to avoid data leakage caused by samples from the same patient appearing in both training and test sets. For each dataset, the model predicts the top 50 highly variable genes, and Pearson correlation coefficient is used as the primary evaluation metric.

In addition to HEST-Bench, we construct an STImage-derived benchmark based on STImage-1K4M (Chen et al., 2024a). Specifically, we select samples for each organ and use a patient-disjoint 80/20 train/test split. Different from the strictly single-platform setting used in prior work, our constructed benchmark contains organ-specific subsets with both single-platform and partially mixed-platform compositions. Most organ subsets are collected from a single spatial transcriptomics technology, while some subsets contain samples from both ST and Visium platforms. This setting introduces moderate platform heterogeneity in part of the benchmark and provides a more challenging evaluation scenario for testing model robustness across tissue and measurement conditions.

To ensure fair comparison, all samples are processed with the same preprocessing pipeline, including image patch extraction, coordinate normalization, gene expression normalization, gene selection, and train/test splitting. The platform information is not used as an additional supervision signal, but is retained as part of the benchmark difficulty. For transparency, we report the technology composition of each organ subset in Table 5. As in HEST-Bench, the model predicts the top 50 highly variable genes, and Pearson correlation coefficient is used as the evaluation metric.

B Supplementary Derivations of the Method

This section gives implementation details for Carta that are omitted from the main text. We describe coordinate normalization and regional node construction, distance-weighted propagation across

scales, the local geometry-aware attention operator, and the gene-anchor prior used in the bidirectional conditional transport objective.

B.1 Coordinate Binning and Regional Node Aggregation

For each tissue section, Carta constructs regional nodes by grouping measured locations at different spatial scales. Let the raw coordinate of spot i be $\mathbf{c}_i^{\mathrm{raw}} = (a_i, b_i)$. We first normalize coordinates within each tissue section. Let

$$s = \max\left(a_{\max} - a_{\min},\, b_{\max} - b_{\min}\right). \tag{17}$$

Then,

$$\tilde{a}_i = \frac{a_i - a_{\min}}{s + \epsilon}, \qquad \tilde{b}_i = \frac{b_i - b_{\min}}{s + \epsilon}. \tag{18}$$

where $\epsilon = 10^{-6}$ is used for numerical stability. We write $\tilde{\mathbf{c}}_i = (\tilde{a}_i, \tilde{b}_i)$. All binning and distance computations below use the normalized coordinates. At scale l, the normalized coordinate frame is divided into $S_x^{(l)} \times S_y^{(l)}$ bins. The bin indices for spot i are:

$$b_{i,x}^{(l)} = \min\Big(\Big\lfloor S_x^{(l)} \tilde{a}_i \Big\rfloor, S_x^{(l)} - 1\Big), \qquad b_{i,y}^{(l)} = \min\Big(\Big\lfloor S_y^{(l)} \tilde{b}_i \Big\rfloor, S_y^{(l)} - 1\Big). \tag{19}$$

Using zero-based indexing, the cell index is:

$$\mathrm{cell}^{(l)}(i) = b_{i,y}^{(l)} S_x^{(l)} + b_{i,x}^{(l)}. \tag{20}$$

We keep cells that contain at least one measured location. The j-th retained cell at scale l defines a regional node: $R_j^{(l)} = \{i : \mathrm{cell}^{(l)}(i) = j\}$. For each regional node, we average the histology embeddings and normalized coordinates of the locations it contains:

$$h_j^{(l)} = \frac{1}{|R_j^{(l)}|} \sum_{i \in R_j^{(l)}} h_i, \qquad \tilde{c}_j^{(l)} = \frac{1}{|R_j^{(l)}|} \sum_{i \in R_j^{(l)}} \tilde{c}_i. \tag{21}$$

For the finest scale, each node corresponds to one measured ST location, so $h_i^{(L)} = h_i$, $\tilde{c}_i^{(L)} = \tilde{c}_i$. This construction keeps the hierarchy tied to measured assay coordinates rather than to a regular image grid.

B.2 Cross-Scale Expression-State Propagation

For $l \geq 2$, expression states are passed from scale $l-1$ to scale l by distance-weighted interpolation. For node i at scale l, let $\mathcal{N}_i^{(l)}$ denote its K nearest nodes from scale $l-1$, using normalized coordinates. The propagated expression state is:

$$p_i^{(l)} = \sum_{j \in \mathcal{N}_i^{(l)}} w_{ij}^{(l)} \hat{y}_j^{(l-1)}, \tag{22}$$

where:

$$w_{ij}^{(l)} = \frac{\exp\Big(-\|\tilde{c}_i^{(l)} - \tilde{c}_j^{(l-1)}\|_2^2 / \lambda_u\Big)}{\sum_{j' \in \mathcal{N}_i^{(l)}} \exp\Big(-\|\tilde{c}_i^{(l)} - \tilde{c}_{j'}^{(l-1)}\|_2^2 / \lambda_u\Big)}. \tag{23}$$

The bandwidth $\lambda_u > 0$ controls the locality of interpolation. Smaller values emphasize nearby nodes, while larger values average over a wider neighborhood.

At the coarsest scale, no previous scale is available. As described in the main text, Carta initializes the expression condition with a learnable gene-wise Gaussian prior:

$$p^{(1)} = \mu + \exp(\log \sigma) \odot \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I). \tag{24}$$

For every coarsest-scale node i, we set $p_i^{(1)} = p^{(1)}$ during training. At evaluation time, we use the mean state $p_i^{(1)} = \mu$ unless otherwise specified.

B.3 Local Geometry-Aware Attention Module

Carta uses local geometry-aware attention to exchange information among nearby nodes. The operator follows the locality-constrained attention design used in STFlow (Huang et al., 2025); in Carta it is used in two places: cross-scale conditional fusion and within-scale self-attention. The two modules share the same geometric attention form but differ in how queries, keys, and values are constructed.

Let $\mathcal{M}_i^{(l)}$ denote the within-scale spatial neighborhood of node i at scale l. For a neighbor $j \in \mathcal{M}_i^{(l)}$, we first compute the relative coordinate and distance:

$$r_{ij}^{(l)} = \tilde{c}_j^{(l)} - \tilde{c}_i^{(l)}, \qquad d_{ij}^{(l)} = \|r_{ij}^{(l)}\|_2. \tag{25}$$

We then transform $r_{ij}^{(l)}$ into a set of frame-aware coordinate features $\{\phi_m(r_{ij}^{(l)})\}_{m=1}^{M}$. The geometric edge feature is then computed as

$$\rho_{ij}^{(l)} = \frac{1}{M} \sum_{m=1}^{M} \psi_e \left(\left[\phi_m(r_{ij}^{(l)}), d_{ij}^{(l)} \right] \right), \tag{26}$$

where $\psi_e(\cdot)$ is an MLP edge encoder and M is the number of frames. We use $\rho_{ij}^{(l)}$ for geometric edge features to avoid confusion with the expression-condition embedding defined below.

For cross-scale conditional fusion, the histology representation provides the query, and the propagated expression state provides keys and values. Since $p_i^{(l)} \in \mathbb{R}^G$ and the attention module operates in the model dimension, we first project the expression condition:

$$e_i^{(l)} = W_p p_i^{(l)} \in \mathbb{R}^d. \tag{27}$$

The query, key, and value are then:

$$q_i^{\text{cross}} = W_q^{\text{cross}} \text{LN}(z_i^{(l)}), \quad [k_j^{\text{cross}}, v_j^{\text{cross}}] = W_{kv}^{\text{cross}} \text{LN}(e_j^{(l)}), \quad j \in \mathcal{M}_i^{(l)}. \tag{28}$$

For within-scale self-attention, query, key, and value are all generated from the current node features:

$$[q_i^{\text{self}}, k_i^{\text{self}}, v_i^{\text{self}}] = W_{qkv}^{\text{self}} \text{LN}(z_i^{(l)}). \tag{29}$$

Given a query q_i, neighbor keys $\{k_j\}$, values $\{v_j\}$, and edge features $\{\rho_{ij}^{(l)}\}$, the local geometry-aware attention logit is:

$$a_{ij} = \psi_a([q_i, k_j, \rho_{ij}^{(l)}]), \tag{30}$$

where $\psi_a(\cdot)$ is an MLP that outputs a scalar logit. The attention weights are normalized within the local neighborhood:

$$A_{ij} = \frac{\exp(a_{ij})}{\sum_{j' \in \mathcal{M}_i^{(l)}} \exp(a_{ij'})}. \tag{31}$$

The local message aggregates both value features and geometric edge features:

$$m_i = \psi_o \left(\left[\sum_{j \in \mathcal{M}_i^{(l)}} A_{ij} v_j, \sum_{j \in \mathcal{M}_i^{(l)}} A_{ij} \rho_{ij}^{(l)} \right] \right), \tag{32}$$

where $\psi_o(\cdot)$ maps the concatenated vector back to the model dimension.

In cross-scale conditional fusion, this message is denoted by m_i^{cross} and is used to inject the propagated expression condition into the histology representation:

$$\bar{z}_i^{(l)} = z_i^{(l)} + m_i^{\text{cross}}, \qquad \tilde{z}_i^{(l)} = \bar{z}_i^{(l)} + \text{FFN}_{\text{cross}} \left(\text{LN}(\bar{z}_i^{(l)}) \right). \tag{33}$$

Here, $\tilde{z}_i^{(l)}$ denotes the node feature after conditioning on the transferred expression state.

The conditioned features are then passed to the within-scale encoder. Here we set:

$$z_{i,0}^{(l)} = \tilde{z}_i^{(l)}. \tag{34}$$

For the r-th within-scale encoder layer, local geometry-aware self-attention produces a message $m_{i,r}^{\text{self}}$, followed by an FFN update:

$$z_{i,r+\frac{1}{2}}^{(l)} = z_{i,r}^{(l)} + m_{i,r}^{\text{self}}, \qquad z_{i,r+1}^{(l)} = z_{i,r+\frac{1}{2}}^{(l)} + \text{FFN}_{\text{self}}\Big(\text{LN}(z_{i,r+\frac{1}{2}}^{(l)})\Big). \tag{35}$$

After R within-scale encoder layers, the output feature is:

$$u_i^{(l)} = z_{i,R}^{(l)}. \tag{36}$$

This feature is passed to the prediction head at scale l.

B.4 Gene Anchors and Dynamic Prior Updates in BCT

BCT is applied between the intermediate features of the prediction head and gene anchors defined by the output layer. At scale l, the prediction head first produces an intermediate feature:

$$f_i^{(l)} = \phi_1(u_i^{(l)}) = \text{LN}\left(\text{GELU}\left(W_1 u_i^{(l)} + b_1\right)\right) \in \mathbb{R}^{d_f}. \tag{37}$$

The expression prediction is:

$$\hat{y}_i^{(l)} = \psi\left(W_o f_i^{(l)} + b_o\right), \tag{38}$$

where $W_1 \in \mathbb{R}^{d_f \times d}$, $b_1 \in \mathbb{R}^{d_f}$, $W_o \in \mathbb{R}^{G \times d_f}$, and $b_o \in \mathbb{R}^G$. The activation ψ is set to softplus for non-negative expression targets and to the identity map for standardized targets.

Each row of the shared output projection is treated as a gene anchor:

$$g_k = \text{sg}(W_o[k,:]), \quad k = 1, \ldots, G. \tag{39}$$

Thus, the same latent directions used to predict gene expression also define the gene-anchor space in BCT. Given node feature $f_i^{(l)}$ and gene anchor g_k, the similarity matrix is:

$$S_{ki}^{(l)} = g_k^\top f_i^{(l)}. \tag{40}$$

The corresponding cosine transport cost is:

$$D_{ki}^{(l)} = 1 - \frac{g_k^\top f_i^{(l)}}{\|g_k\|_2 \|f_i^{(l)}\|_2}. \tag{41}$$

Because gene activity can be uneven within a mini-batch, BCT maintains a dynamic prior $\mathbf{q} \in \Delta^{G-1}$ over gene anchors, where:

$$\Delta^{G-1} = \left\{\mathbf{q} \in \mathbb{R}_+^G : \sum_{k=1}^G q_k = 1\right\}.$$

The prior is initialized uniformly:

$$q_k^{(0)} = \frac{1}{G}, \qquad k = 1, \ldots, G. \tag{42}$$

During each forward pass, we compute detached node-to-anchor assignments using the current prior:

$$\pi_{ki}^{\text{old},(l)} = \text{softmax}_k\left(\frac{\text{sg}(S_{ki}^{(l)})}{\tau} + \log(q_k + \delta)\right), \tag{43}$$

where τ is the temperature, and $\delta = 10^{-6}$ is used for numerical stability.

We estimate the batch-level gene-anchor usage by averaging these detached assignments over the nodes used in the current BCT computation at scale l:

$$\bar{q}_k^{(l)} = \frac{1}{N^{(l)}} \sum_{i=1}^{N^{(l)}} \pi_{ki}^{\text{old},(l)}. \tag{44}$$

The prior is updated by momentum outside backpropagation:

$$q_k \leftarrow (1-\beta)q_k + \beta \bar{q}_k^{(l)}. \tag{45}$$

Since both $\mathbf{q}$ and $\bar{\mathbf{q}}^{(l)}$ lie in the probability simplex, this update preserves normalization up to numerical precision. In implementation, we renormalize $\mathbf{q}$ after the update. The updated prior is used in the node-to-anchor transport weights:

$$\pi_{ki}^{s,(l)} = \mathrm{softmax}_k\left(\frac{S_{ki}^{(l)}}{\tau} + \log(q_k + \delta)\right). \tag{46}$$

The corresponding loss is:

$$\mathcal{L}_s^{(l)} = \frac{\alpha}{N^{(l)}} \sum_{i=1}^{N^{(l)}} \sum_{k=1}^{G} D_{ki}^{(l)} \pi_{ki}^{s,(l)}. \tag{47}$$

This term assigns each spatial node to gene anchors while accounting for gene-level imbalance through the dynamic prior. The reverse transport weights assign each gene anchor to spatial nodes:

$$\pi_{ki}^{t,(l)} = \mathrm{softmax}_i\left(\frac{S_{ki}^{(l)}}{\tau}\right). \tag{48}$$

The reverse-direction loss is:

$$\mathcal{L}_t^{(l)} = (1-\alpha) \sum_{k=1}^{G} q_k \sum_{i=1}^{N^{(l)}} D_{ki}^{(l)} \pi_{ki}^{t,(l)}. \tag{49}$$

This direction encourages each gene anchor to be supported by spatial nodes, rather than concentrating the alignment on a small set of frequent responses. The parameter $\alpha \in [0,1]$ balances the two transport directions. The BCT loss at scale l is:

$$\mathcal{L}_{\mathrm{BCT}}^{(l)} = \mathcal{L}_s^{(l)} + \mathcal{L}_t^{(l)}. \tag{50}$$

The multi-scale BCT regularizer is:

$$\mathcal{L}_{\mathrm{BCT}} = \frac{1}{L} \sum_{l=1}^{L} \mathcal{L}_{\mathrm{BCT}}^{(l)}. \tag{51}$$

Regression supervision is applied at the finest scale:

$$\mathcal{L}_{\mathrm{reg}} = \mathrm{MSE}(\hat{\mathbf{Y}}^{(L)}, \mathbf{Y}). \tag{52}$$

The final objective is:

$$\mathcal{L} = \mathcal{L}_{\mathrm{reg}} + \lambda_{\mathrm{BCT}} \mathcal{L}_{\mathrm{BCT}}, \tag{53}$$

where λ_{BCT} controls the strength of transport regularization.

C Additional Experiments

This section provides additional experiments to further characterize Carta and its key design choices. We first extend the empirical comparison with recent context-aware histology-to-ST methods and stronger prediction heads built on frozen pathology features. We then evaluate the field-reconstruction formulation through prediction at held-out spatial coordinates, spatial field diagnostics, and gene sparsity-wise analysis. Next, we examine the regional-to-local reconstruction design, including the scale schedule, cross-scale state propagation, cross-scale fusion, and the coarsest-scale initial state. We further analyze BCT through hyperparameter sensitivity, gene-anchor construction, and the relationship between gene-anchor geometry and gene co-expression. Finally, we report computational efficiency. Together, these experiments complement the main results and provide a more complete characterization of Carta's reconstruction behavior, design choices, and computational cost.

Table 6: **Comparison with recent methods that incorporate broader tissue context.** PCC is reported for each organ across HEST and STImage. Subscripts indicate the standard deviation across multiple random seeds. The best results are highlighted in bold, and the second-best results are underlined.

Dataset	Organ	TRIPLEX	FmH2ST	MERGE	Carta
HEST	CCRCC	$\underline{0.284}_{.004}$	$0.229_{.008}$	$0.262_{.009}$	$\mathbf{0.335}_{.007}$
	COAD	$0.302_{.002}$	$0.267_{.006}$	$\underline{0.307}_{.007}$	$\mathbf{0.363}_{.008}$
	HCC	$0.056_{.005}$	$0.081_{.004}$	$\underline{0.097}_{.002}$	$\mathbf{0.132}_{.006}$
	IDC	$\mathbf{0.599}_{.003}$	$0.541_{.011}$	$0.586_{.006}$	$\underline{0.595}_{.009}$
	LUNG	$0.605_{.002}$	$0.562_{.008}$	$\underline{0.608}_{.003}$	$\mathbf{0.621}_{.004}$
	LYMPH	$\underline{0.301}_{.002}$	$0.264_{.004}$	$0.282_{.010}$	$\mathbf{0.315}_{.005}$
	PAAD	$\underline{0.487}_{.005}$	$0.445_{.009}$	$0.458_{.006}$	$\mathbf{0.516}_{.010}$
	PRAD	$\underline{0.398}_{.006}$	$0.351_{.003}$	$0.387_{.002}$	$\mathbf{0.422}_{.004}$
	READ	$0.203_{.011}$	$0.192_{.014}$	$\underline{0.235}_{.016}$	$\mathbf{0.245}_{.033}$
	SKCM	$0.694_{.006}$	$0.656_{.003}$	$\underline{0.696}_{.004}$	$\mathbf{0.703}_{.006}$
	Avg	$\underline{0.393}$	0.359	0.392	$\mathbf{0.425}$
STImage	Brain	$\underline{0.551}_{.007}$	$0.536_{.010}$	$0.546_{.003}$	$\mathbf{0.586}_{.016}$
	Breast	$0.544_{.014}$	$0.509_{.002}$	$\underline{0.553}_{.009}$	$\mathbf{0.595}_{.013}$
	Heart	$\underline{0.739}_{.005}$	$0.697_{.008}$	$0.736_{.002}$	$\mathbf{0.763}_{.006}$
	Liver	$\underline{0.598}_{.016}$	$0.574_{.019}$	$0.595_{.013}$	$\mathbf{0.659}_{.022}$
	Pancreas	$\underline{0.797}_{.009}$	$0.752_{.026}$	$0.795_{.017}$	$\mathbf{0.806}_{.017}$
	Skin	$\underline{0.499}_{.006}$	$0.450_{.009}$	$0.486_{.004}$	$\mathbf{0.509}_{.005}$
	Avg	$\underline{0.621}$	0.586	0.619	$\mathbf{0.653}$

C.1 Comparison with Methods Incorporating Broader Tissue Context

The main comparison covers baselines spanning pathology foundation-model features, image–expression alignment, task-specific regression, and generative modeling. We further compare with three recent methods that incorporate broader tissue context: TRIPLEX (Chung et al., 2024), FmH2ST (Wang et al., 2025), and MERGE (Ganguly et al., 2025). TRIPLEX integrates multi-resolution local and global context, FmH2ST combines slice-level context with spatial and feature graphs, and MERGE models local and long-range interactions through a hierarchical graph. Carta instead formulates prediction as spatial field reconstruction, where a coordinate-derived regional-to-local hierarchy progressively reconstructs fine-scale expression states from coarse-scale expression priors.

As shown in Table 6, Carta achieves the highest average PCC on both HEST and STImage, reaching 0.425 and 0.653, compared with 0.393/0.621 for TRIPLEX, 0.359/0.586 for FmH2ST, and 0.392/0.619 for MERGE. Carta obtains the highest PCC on 15 of the 16 individual tasks, with TRIPLEX marginally higher only on IDC (0.599 versus 0.595). These results show that the field-reconstruction formulation provides effective prediction improvements compared with existing context-aware histology-to-ST approaches.

C.2 Ablation on Pathology Features

Table 7 compares different prediction heads applied to the same frozen pathology features from Ciga, Gigapath, and UNI. We evaluate Random Forest, linear regression, ridge regression, and a two-layer MLP while keeping the pretrained feature extractors fixed. Hyperparameters are selected on HEST following the same cross-validation and model-selection protocol as in the main experiments.

Across all three pathology encoders, the feature-only predictors remain below Carta. The best average PCC achieved using frozen Ciga, Gigapath, and UNI features is 0.303, 0.353, and 0.356, respectively, compared with 0.363, 0.412, and 0.425 for Carta. This corresponds to gains of 0.060, 0.059, and 0.069 PCC. These results indicate that pathology features alone do not fully explain

Table 7: **Comparison of prediction heads on frozen pathology foundation-model features.** Average PCC on HEST is reported for Ciga, Gigapath, and UNI.

Predictor	Ciga	Gigapath	UNI
Random Forest	0.281	0.327	0.332
Linear regression	0.249	0.232	0.269
Ridge regression	0.294	0.346	0.342
2-layer MLP	0.303	0.353	0.356
Carta	**0.363**	**0.412**	**0.425**

the prediction performance, and that the spatial expression modeling in Carta provides additional predictive information.

C.3 Field Reconstruction from Partial Spatial Support

To evaluate Carta under incomplete spatial support, we randomly hold out 10%, 25%, or 50% of measured coordinates from the finest-scale support, coarse-scale binning, and local-attention neighborhoods. These locations do not participate in hierarchy construction, and their expression labels are used only for evaluation. At inference, each held-out location is predicted from its H&E patch together with the expression context reconstructed from the retained coordinates. This setting tests whether Carta can recover omitted spatial expression from a partially observed tissue layout.

Table 8: **Prediction under partial spatial support.** PCC is reported when different ratios of measured coordinates are excluded from hierarchy construction and evaluated as held-out locations. All methods are evaluated on the same held-out coordinates, with results averaged over five fixed masking seeds.

Method	10%	25%	50%
UNI+RF	0.3320	0.3312	0.3316
kNN-Pred	0.3373	0.3320	0.3205
STFlow	<u>0.3838</u>	<u>0.3834</u>	<u>0.3836</u>
Carta	**0.4101**	**0.4093**	**0.4080**

As shown in Table 8, Carta achieves the highest PCC at all three held-out ratios, outperforming STFlow by 0.0255 PCC on average. Its performance remains highly stable as the spatial support becomes sparser, decreasing only from 0.4101 to 0.4080 when the held-out ratio increases from 10% to 50%. UNI+RF, which predicts expression from local histology alone, and kNN-Pred, which interpolates from retained spatial locations, remain substantially weaker. Carta also consistently outperforms STFlow, which introduces local spatial interactions only within its histology-conditioned expression generation process. These results indicate that Carta can reconstruct useful spatial expression context from the retained coordinates and use it to support prediction at omitted locations.

Table 9: **Spatial field diagnostics on HEST.** We report regional error, variance consistency, and recovery of spatially selective activation.

Method	Region MAE ↓	Var. Log Error ↓	Top-10 IoU ↑	Top-10 Recall ↑
STFlow	0.524	2.965	0.176	0.284
Carta w/o BCT	0.513	2.823	0.180	0.282
Carta	**0.487**	**2.651**	**0.191**	**0.293**

C.4 Spatial Field Reconstruction Diagnostics

Diagnostics beyond gene-wise correlation. Since PCC mainly measures gene-wise correlation, we further introduce several spatial field diagnostics to evaluate whether the reconstructed expression

maps preserve regional structure and localized activation patterns. *Region MAE* first partitions each tissue section into spatial regions using K-means on normalized coordinates and compares the average predicted and measured expression within each region; it measures whether the regional expression intensity is correctly reconstructed. *Var. Log Error* compares the variance of predicted and measured regional expression in log space; a lower value indicates that the model better preserves field-level variability and is less prone to over-smoothing or noisy over-dispersion. *Top-10 IoU* computes the intersection-over-union between the predicted and ground-truth top-10% high-expression locations, measuring the spatial overlap of localized activation regions. *Top-10 Recall* measures the fraction of ground-truth top-10% high-expression locations covered by the predicted top-10% regions, reflecting whether the model misses true localized activation. All metrics are first computed within each test slide and then aggregated at the organ level.

Results on field-level structure. Table 9 shows that Carta improves over STFlow on all four spatial field diagnostics. Carta reduces Region MAE from 0.524 to 0.487, indicating better reconstruction of regional expression intensity. It also lowers Var. Log Error from 2.965 to 2.651, suggesting that the predicted field better preserves expression variability rather than collapsing to an over-smoothed map. For spatially selective activation, Carta improves both Top-10 IoU and Top-10 Recall, showing better recovery of localized activated regions. Carta w/o BCT consistently lies between STFlow and the full model on most metrics, indicating that BCT further shapes the spatial distribution of the reconstructed field. These results support our formulation of spatial gene expression prediction as field reconstruction rather than independent coordinate-level regression.

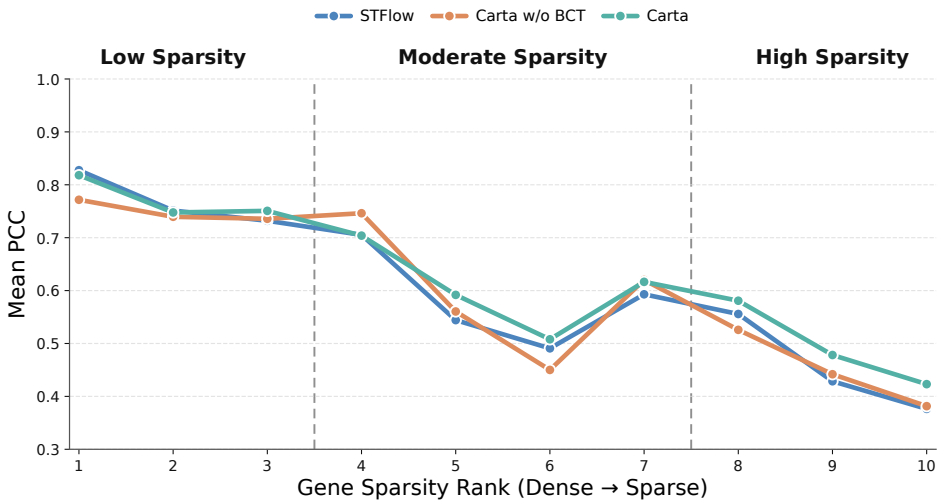

Figure 4: **Gene sparsity-wise PCC on HEST-LUNG.** We use HEST-LUNG because its selected genes exhibit a clear and broad sparsity spectrum, making the dense-to-sparse comparison more informative. Genes are ranked by the fraction of spots with zero expression and grouped into ten bins from dense to sparse; curves show the mean PCC within each bin.

C.5 Gene sparsity-wise comparison

Figure 4 compares gene-wise performance across ten sparsity-ranked bins on HEST-LUNG, from dense to sparse genes. All methods decline as genes become sparser, indicating that low-activation genes are generally harder to recover from histology. Carta achieves the highest mean PCC in most bins, while Carta w/o BCT remains competitive and typically ranks near the top. The gap between the two variants is small on dense genes but becomes clearer on medium-to-sparse bins, where the full Carta more consistently maintains higher PCC. This pattern suggests that the regional-to-local reconstruction provides the main gain across sparsity levels, while BCT further improves the recovery of weaker and more spatially selective expression patterns.

Table 10: **Ablation study on multi-scale schedule design.** We report organ-wise PCC on HEST under different multi-scale schedules. N denotes the native spot-level scale of each tissue section.

Strategy	Schedule	CCRCC	COAD	HCC	IDC	LUNG	LYMPH	PAAD	PRAD	READ	SKCM	Avg.
Very sparse	$\mathcal{S}_{\mathrm{sp}}$	0.328	0.347	0.120	0.581	0.613	0.303	0.503	0.406	0.243	0.691	0.414
w/o early coarse	$\mathcal{S}_{\mathrm{wc}}$	0.327	0.339	0.118	0.577	0.610	0.308	0.507	0.399	0.237	0.685	0.411
Finer late	$\mathcal{S}_{\mathrm{fine}}$	0.339	**0.369**	0.130	0.594	**0.624**	**0.318**	**0.518**	0.420	0.258	**0.710**	**0.428**
Larger pre-native	$\mathcal{S}_{\mathrm{large}}$	**0.340**	0.357	0.128	**0.596**	0.622	0.310	0.514	0.419	**0.260**	0.695	0.424
Default	$\mathcal{S}_{\mathrm{def}}$	0.335	0.363	**0.132**	0.595	0.621	0.315	0.516	**0.422**	0.245	0.703	0.425

Schedule definitions. $\mathcal{S}_{\mathrm{sp}} = \{1^2, 8^2, 20^2, N\}$; $\mathcal{S}_{\mathrm{wc}} = \{8^2, 12^2, 16^2, 20^2, N\}$; $\mathcal{S}_{\mathrm{def}} = \{1^2, 2^2, 4^2, 8^2, 12^2, 16^2, 20^2, N\}$; $\mathcal{S}_{\mathrm{fine}} = \{1^2, 2^2, 4^2, 8^2, 10^2, 12^2, 14^2, 16^2, 18^2, 20^2, N\}$; $\mathcal{S}_{\mathrm{large}} = \{1^2, 2^2, 4^2, 8^2, 12^2, 16^2, 20^2, 28^2, N\}$.

C.6 Scale Schedule Design

Table 10 studies how the choice of scale schedule affects field reconstruction. The two reduced schedules perform worse: the sparse schedule $\mathcal{S}_{\mathrm{sp}}$ reaches an average PCC of 0.414, and the schedule without early coarse stages $\mathcal{S}_{\mathrm{wc}}$ reaches 0.411. These drops indicate that Carta benefits from both early regional states and intermediate transitions between regional and local predictions. Adding denser late-stage scales improves the average PCC to 0.428 for $\mathcal{S}_{\mathrm{fine}}$, showing that refinement near the measured coordinates is useful. The larger pre-final schedule $\mathcal{S}_{\mathrm{large}}$ reaches 0.424, suggesting that adding a single larger intermediate layout is less effective than maintaining a balanced sequence of regional and near-coordinate scales. The default schedule $\mathcal{S}_{\mathrm{def}}$ obtains 0.425 average PCC, within 0.003 of $\mathcal{S}_{\mathrm{fine}}$, while using a more compact scale sequence. We therefore use $\mathcal{S}_{\mathrm{def}}$ in the main experiments.

Table 11: **Organ-wise ablation study of cross-scale state propagation on HEST.** We report PCC and MSE for each organ using UNI as the default pathology foundation model. All variants use the same regional-to-local hierarchy, scale schedule, refinement blocks, BCT configuration, and training protocol. The four expression-state variants differ only in the cross-scale transfer mechanism. The latent-state control instead propagates a dimension-matched latent state derived from the same intermediate prediction feature, while retaining the distance-weighted transfer rule. Best and second-best results are highlighted in bold and underlined, respectively.

Metric	State	Transfer	Organ										Avg.
			CCRCC	COAD	HCC	IDC	LUNG	LYMPH	PAAD	PRAD	READ	SKCM	
PCC ↑	Expression	Global broadcast	0.301	0.342	0.117	0.597	0.613	0.311	0.512	0.417	0.228	0.702	0.414
	Expression	Learned cross-scale attention	0.312	0.357	0.122	0.603	0.619	0.310	0.514	0.411	0.235	0.708	0.419
	Expression	Nearest parent	0.327	0.348	0.124	**0.610**	**0.622**	0.308	0.513	0.414	0.239	0.704	0.421
	Latent	Distance-weighted	0.322	0.353	0.130	0.591	0.620	0.301	0.511	0.419	0.236	**0.709**	0.419
	Expression	**Distance-weighted**	**0.335**	**0.363**	**0.132**	0.595	0.621	**0.315**	**0.516**	**0.422**	**0.245**	0.703	**0.425**
MSE ↓	Expression	Global broadcast	**0.483**	**1.838**	0.418	2.752	2.040	0.533	1.316	0.549	**0.297**	1.967	1.219
	Expression	Learned cross-scale attention	0.493	1.873	0.416	2.629	1.967	0.568	**1.311**	0.554	0.298	1.975	1.208
	Expression	Nearest parent	0.501	1.866	**0.408**	2.669	2.021	0.531	1.339	0.553	0.298	1.875	1.206
	Latent	Distance-weighted	0.500	1.928	0.448	**2.619**	**1.880**	0.620	1.347	**0.547**	0.300	1.881	1.207
	Expression	**Distance-weighted**	0.508	1.853	0.415	2.656	2.018	**0.528**	1.346	0.550	0.305	**1.872**	**1.205**

C.7 Cross-scale State Propagation

Table 11 compares different cross-scale state transfer strategies under the same regional-to-local hierarchy and training configuration. Global broadcast averages the coarse-scale expression states and assigns the same condition to all finer nodes; nearest-parent transfer uses the expression state of the spatially closest coarse node; learned cross-scale attention computes content-dependent aggregation from neighboring coarse nodes; and distance-weighted propagation interpolates neighboring coarse-scale states according to their spatial distances. Among these expression-state variants, distance-weighted propagation achieves the best overall performance, with an average PCC of 0.4247 and an average MSE of 1.2051, and obtains the highest PCC on 7 of the 10 organs.

We further compare the propagated state itself while fixing the transfer rule to distance weighting. Replacing the predicted expression state with a dimension-matched latent state reduces the average PCC from 0.4247 to 0.4192, while the average MSE remains similar (1.2051 vs. 1.2070). Expression-

state propagation achieves higher PCC on 9 of the 10 organs, indicating that predicted expression states provide a more effective cross-scale condition for regional-to-local refinement.

Table 12: **Organ-wise ablation study of cross-scale fusion strategies on HEST.** We report PCC and MSE for each organ using UNI as the default pathology foundation model. All variants use the same training configuration, multi-scale reconstruction schedule, and BCT configuration, except for the way propagated expression states are fused with current-scale histology features. Best and second-best results are highlighted in bold and underlined, respectively.

Metric	Fusion strategy	Organ										Avg.
		CCRCC	COAD	HCC	IDC	LUNG	LYMPH	PAAD	PRAD	READ	SKCM	
PCC ↑	Concat + Linear	0.315	**0.364**	0.119	0.591	0.620	0.310	0.506	0.418	0.239	0.695	0.418
	Linear + Add	0.309	0.360	0.121	0.590	0.618	0.306	0.505	0.420	0.234	0.700	0.416
	Cross-attention	**0.335**	0.363	**0.132**	**0.595**	**0.621**	**0.315**	**0.516**	**0.422**	**0.245**	**0.703**	**0.425**
MSE ↓	Concat + Linear	0.516	**1.828**	**0.403**	2.689	**2.016**	**0.492**	1.382	**0.538**	**0.301**	1.925	1.209
	Linear + Add	0.514	1.837	0.404	2.676	**2.016**	0.542	1.380	0.540	0.304	1.906	1.212
	Cross-attention	**0.508**	1.853	0.415	**2.656**	2.018	0.528	**1.346**	0.550	0.305	**1.872**	**1.205**

C.8 Cross-scale Fusion Strategy

Table 12 compares different strategies for fusing propagated coarse-scale expression states with current-scale histology features. The concat + linear strategy directly concatenates the two representations and projects them back to the hidden space, while the linear + add strategy first maps the expression state into the histology feature space and then performs residual addition. Although these two simple fusion strategies can inject coarse-scale expression information, they treat the propagated expression state as a fixed feature-level condition and do not explicitly model local spatial correspondence during fusion.

In contrast, the default cross-attention fusion uses current-scale histology features as queries and propagated expression states as keys and values, with geometric edge features incorporated into the attention computation. This allows each node to selectively aggregate expression conditions from its local spatial neighborhood in a content-adaptive and geometry-aware manner. As shown in Table 12, cross-attention achieves the best average PCC of 0.425, outperforming concat + linear fusion (0.418) and linear + add fusion (0.416). It also obtains the lowest average MSE of 1.205.

The organ-wise results further show that cross-attention brings consistent gains in PCC across most organs, achieving the best result on nine out of ten organs. This suggests that adaptive cross-scale conditioning is more effective for preserving relative spatial expression variation than simple feature concatenation or addition. Meanwhile, concat + linear and linear + add obtain competitive MSE on some organs, indicating that simple fusion may still reduce point-wise errors in certain cases. However, their lower PCC suggests weaker recovery of spatial expression patterns. Overall, these results support our design choice of using local geometry-aware cross-attention for cross-scale expression fusion in Carta.

Table 13: **Organ-wise ablation study of the initial expression state on HEST.** We report PCC and MSE for each organ using UNI as the default pathology foundation model. All variants use the same multi-scale reconstruction schedule and BCT configuration. They differ only in whether a small stochastic perturbation is injected into the learnable coarsest-scale expression state during training. Evaluation is performed with deterministic initialization. Best results are highlighted in bold, and the default setting is shaded.

Metric	Initial state during training	Organ										Avg.
		CCRCC	COAD	HCC	IDC	LUNG	LYMPH	PAAD	PRAD	READ	SKCM	
PCC ↑	Learnable state	0.317	**0.367**	0.129	0.590	0.618	0.311	0.506	0.414	0.238	0.701	0.419
	Learnable state + perturb	**0.335**	0.363	**0.132**	**0.595**	**0.621**	**0.315**	**0.516**	**0.422**	**0.245**	**0.703**	**0.425**
MSE ↓	Learnable state	0.517	1.864	**0.410**	2.671	2.022	**0.514**	1.349	**0.537**	**0.296**	1.993	1.217
	Learnable state + perturb	**0.508**	**1.853**	0.415	**2.656**	**2.018**	0.528	**1.346**	0.550	0.305	**1.872**	**1.205**

C.9 Training-time Perturbation of Initial State

Table 13 studies the effect of training-time perturbation on the coarsest-scale initial expression state. Since Carta is formulated as deterministic field reconstruction at evaluation, both variants use the same deterministic initialization at test time. The only difference is whether a small stochastic perturbation is added to the learnable initial state during training.

Adding this perturbation increases the average PCC from 0.419 to 0.425 and decreases the average MSE from 1.217 to 1.205. This suggests that training-time perturbation provides useful regularization for the coarse-to-fine reconstruction path and prevents the model from relying too strongly on a fixed initial state. The organ-wise results show that the perturbation improves most tasks, while the deterministic learnable state remains competitive on a few organs. Since both variants use deterministic initialization at evaluation, this comparison isolates the effect of training-time perturbation. In Carta, we therefore use the perturbed learnable state during training and deterministic mean initialization during evaluation.

C.10 BCT Hyperparameter Sensitivity

Table 14 studies the sensitivity of BCT to its key hyperparameters on HEST. Overall, Carta remains stable under different BCT configurations, while the default setting consistently achieves the best average PCC and MSE. When $\lambda_{\mathrm{BCT}} = 0$, the model degenerates to the variant without BCT, leading to a clear drop in PCC from 0.425 to 0.413 and an increase in MSE from 1.205 to 1.225. This confirms the effectiveness of BCT beyond point-wise regression. Increasing λ_{BCT} to 0.01 also degrades performance, suggesting that an overly strong BCT constraint may interfere with coordinate-level expression fitting.

Table 14: **Sensitivity analysis of BCT hyperparameters.** The default setting is highlighted.

λ_{BCT}	PCC↑	MSE↓
0	0.413	1.225
0.001	**0.425**	**1.205**
0.01	0.417	1.220

(a) λ_{BCT}

τ	PCC↑	MSE↓
0.5	0.419	1.217
1	**0.425**	**1.205**
2	0.421	1.215

(b) Temperature τ

α	PCC↑	MSE↓
0	0.420	1.213
0.5	**0.425**	**1.205**
1	0.421	1.211

(c) Balance α

β	PCC↑	MSE↓
0	0.422	1.212
0.05	**0.425**	**1.205**
0.1	0.423	1.210

(d) Momentum β

For the temperature τ, the best performance is obtained at $\tau = 1$. A smaller temperature makes the anchor assignment sharper, while a larger temperature produces smoother assignments; both variants slightly reduce performance. For the balance coefficient α, using both transport directions with $\alpha = 0.5$ outperforms using either direction alone, indicating that bidirectional conditional transport is beneficial for aligning node representations and gene anchors. Finally, the momentum coefficient $\beta = 0.05$ gives the best results among the tested values. The setting $\beta = 0$ disables the momentum update of the gene-anchor prior; since the prior is initialized uniformly, this setting effectively reduces BCT to the variant without a dynamic prior. Compared with $\beta = 0$ or $\beta = 0.1$, the moderate update rate $\beta = 0.05$ provides a more stable dynamic prior for gene-anchor assignment.

Table 15: **Organ-wise ablation study of gene-anchor design in BCT on HEST.** We report PCC and MSE for each organ using UNI as the default pathology foundation model. All variants use the same training configuration and BCT hyperparameters, except for the source of gene anchors. Best and second-best results are highlighted in bold and underlined, respectively.

Metric	BCT variant	Organ										Avg.
		CCRCC	COAD	HCC	IDC	LUNG	LYMPH	PAAD	PRAD	READ	SKCM	
PCC ↑	w/o BCT	0.316	0.352	0.119	0.576	0.612	0.308	**0.517**	0.412	0.238	0.683	0.413
	Separate anchors	0.330	**0.366**	0.118	0.585	0.618	0.310	0.505	0.421	**0.248**	0.701	0.420
	Prediction-head anchors	**0.335**	0.363	**0.132**	**0.595**	**0.621**	**0.315**	0.516	**0.422**	0.245	**0.703**	**0.425**
MSE ↓	w/o BCT	**0.498**	**1.822**	0.420	2.760	**1.987**	0.556	**1.294**	0.566	0.304	2.040	1.225
	Separate anchors	0.499	1.837	**0.403**	2.666	2.040	**0.517**	1.345	0.568	**0.284**	1.957	1.212
	Prediction-head anchors	0.508	1.853	0.415	**2.656**	2.018	0.528	1.346	**0.550**	0.305	**1.872**	**1.205**

C.11 Gene-anchor Design in BCT

Table 15 presents an organ-wise comparison of different gene-anchor designs in BCT. Compared with the variant without BCT, using separate learnable anchors improves the average PCC from 0.413 to 0.420 and reduces the average MSE from 1.225 to 1.212, indicating that an explicit gene-anchor space provides useful structure beyond point-wise regression. However, the default design that uses the prediction-head weights as gene anchors achieves the best overall performance, further improving the average PCC to 0.425 and reducing the average MSE to 1.205. This suggests that the latent directions directly responsible for gene-wise expression prediction provide a more consistent anchor space than separately learned anchors.

The organ-wise results also show that lower MSE does not always correspond to higher PCC. For example, in CCRCC, COAD, LUNG, and PAAD, the variant without BCT obtains competitive or even lower MSE, but its PCC is clearly lower than that of the BCT-based variants. This indicates that point-wise MSE alone cannot fully supervise the reconstruction of spatial expression patterns. While MSE mainly measures absolute prediction error, PCC better reflects whether the predicted expression preserves the relative spatial variation across spots. Therefore, the improvement brought by BCT is not merely a reduction of point-wise error, but a better preservation of expression structures and gene-wise variation patterns.

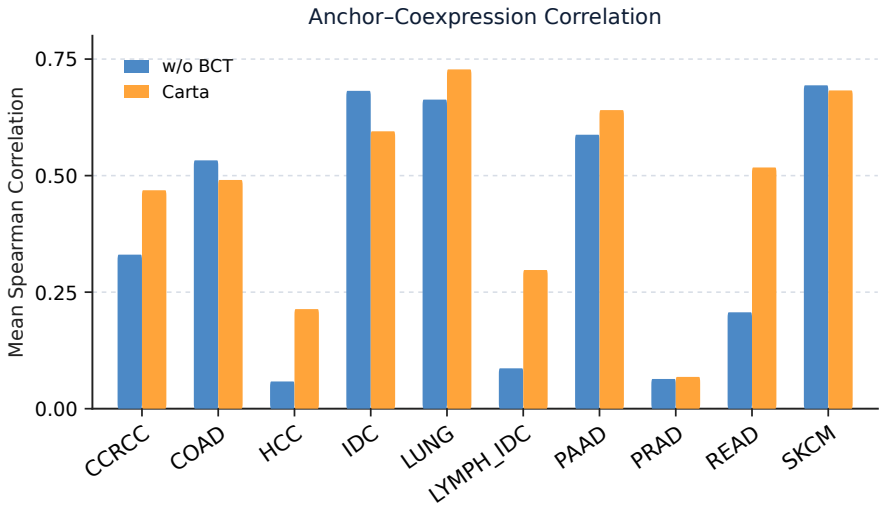

Figure 5: **Association between gene-anchor geometry and gene co-expression across cancer types.** For each patient-disjoint HEST fold, we calculate the Spearman correlation between pairwise cosine similarities of the learned gene anchors and gene co-expression computed exclusively from the corresponding training spots. The bars report the mean correlation across folds belonging to the same cancer type for Carta and its variant without BCT.

C.12 Relation to Gene Co-expression

As shown in Figure 5, the learned gene-anchor geometry exhibits positive mean correlations with gene co-expression across all evaluated cancer types. The association is particularly strong for LUNG, SKCM, and PAAD, while relatively weaker correlations are observed for PRAD, HCC, and LYMPH_IDC. Compared with the variant without BCT, Carta achieves higher mean correlations in seven of the ten cancer types, with particularly clear improvements on CCRCC, HCC, LYMPH_IDC, and READ. Overall, the mean correlation increases from 0.39 without BCT to 0.47 with BCT. These results suggest that the expression-supervised gene-anchor space used by BCT reflects gene co-expression structure, despite substantial variation in correlation strength across cancer types. This observation is consistent with the gene-aware design of BCT, where each gene is represented by a dedicated anchor and the bidirectional assignment encourages each gene anchor to receive support from spatial nodes rather than concentrating the matching on a small set of frequently activated gene responses.

Table 16: **Efficiency comparison on HEST.** We report trainable parameters, inference time normalized to 10,000 spots, and average PCC for representative neural methods. Feature extraction time is excluded when applicable.

Method	Prediction paradigm	Computational cost		Performance
		Trainable Params. ↓	Time / 10K spots ↓	Avg. PCC ↑
STNet	Spot-wise neural regression	**0.102M**	0.121s	0.317
HisToGene	Slide-level Transformer regression	169.075M	0.397s	0.213
BLEEP	Contrastive retrieval-based prediction	0.672M	**0.036s**	0.350
Stem	Diffusion-based generation	7.474M	318.311s	0.359
STFlow	Flow matching generation	1.148M	0.508s	0.401
Carta	**Scale-wise expression-field reconstruction**	4.526M	0.230s	**0.425**

C.13 Efficiency Analysis

Before analyzing the results, we clarify the evaluation protocol for efficiency comparison. Trainable parameters only include modules updated during training under each method's actual training configuration. For runtime measurement, we exclude image or pathology feature extraction when applicable and report the inference time normalized to 10,000 spots. For generative baselines, Stem is evaluated with 1000 DDPM denoising steps and one generated sample per spot, while STFlow uses 5 flow-matching refinement steps. Thus, the reported runtime mainly reflects the cost of the prediction module rather than the frozen histology encoder.

Table 16 compares the computational cost and predictive accuracy of representative neural methods. Lightweight methods such as STNet and BLEEP have small trainable modules and fast inference, but their average PCC remains limited, suggesting that simple spot-wise regression or retrieval-based prediction is insufficient for capturing spatial expression patterns. Stem improves the prediction accuracy by introducing diffusion-based generation, but its 1000-step DDPM sampling leads to a much higher inference cost. STFlow largely alleviates this issue with 5-step flow-matching refinement, reducing the runtime from 318.311s to 0.508s while further improving the average PCC. Carta takes another step by reconstructing the expression field in a scale-wise manner. Compared with STFlow, Carta increases the average PCC from 0.401 to 0.425 and reduces the normalized inference time from 0.508s to 0.230s. This indicates that Carta does not simply trade efficiency for accuracy, but achieves stronger prediction with a moderate computational budget.

D Qualitative and Further Analysis

This section provides additional qualitative analysis of Carta. We first visualize more biomarker expression maps with additional baseline methods, complementing the main-text visualization and biomarker-level quantitative results. We then show the intermediate coarse-to-fine reconstruction process of Carta, which illustrates how regional expression priors are gradually refined into native-resolution expression maps. These visualizations provide direct evidence that Carta preserves both localized activation and low-expression background, supporting the formulation of histology-based gene expression prediction as spatial field reconstruction.

D.1 Additional Biomarker Expression Maps

In this section, we provide additional qualitative comparisons on four representative biomarkers, including KRTDAP, PTPRC, MET, and UBE2C. Compared with Figure 3 in the main text, we include more baseline methods and additional biomarker genes for a more comprehensive supplementary comparison. These visualizations complement the quantitative results in Table 2 by showing how different methods reconstruct spatial expression patterns at the tissue-section level.

As shown in Figures 6–9, several baselines recover part of the global expression trend, especially for biomarkers with broad activation regions, but their predictions are less consistent across genes. For the sparse KRTDAP pattern, some methods miss boundary-localized activation or introduce signals in unrelated tissue areas. For PTPRC, MET, and UBE2C, the shared tissue outline but different ground-truth expression fields make it possible to examine whether a model simply follows tissue

morphology or truly adapts to gene-specific spatial patterns. In these cases, some baselines produce overly broad activation maps or smooth out the contrast between high- and low-expression regions.

Carta gives more stable predictions across these settings. It better preserves the boundary-associated activation for KRTDAP and adapts to distinct biomarker-specific patterns for PTPRC, MET, and UBE2C under the same morphology. Although some baselines also produce reasonable maps in certain cases, Carta more consistently matches the measured expression fields in both spatial localization and expression contrast. These observations are consistent with the quantitative biomarker results and further support the benefit of scale-wise expression field reconstruction.

D.2 Visualization of Coarse-to-Fine Reconstruction

Figures 10–12 visualize the coarse-to-fine reconstruction process of Carta for representative biomarkers. At the earliest scales, only a small number of regional nodes are available, so the maps mainly encode coarse expression priors rather than detailed spatial patterns. As the scale increases, the predicted expression field gradually recovers the tissue outline and forms gene-specific activation regions. For ERBB2 and PTPRC, which show broad expression patterns, the intermediate scales first capture the global high- and low-expression trends and then refine local variations at later stages. For KRTDAP, whose expression is more localized, high-expression regions emerge progressively near the boundary-associated areas instead of appearing as isolated noisy spots. The final native-resolution maps further sharpen these structures and better align with the measured expression fields.

These visualizations illustrate the role of scale-wise autoregression in Carta. Coarse scales provide regional expression context, while finer scales refine the prediction using local histology and spatial neighborhood information. This process differs from direct spot-wise regression, where each measured location is fitted independently and intermediate field structure is not explicitly modeled. The observed refinement behavior supports our formulation of histology-based gene expression prediction as morphology-conditioned expression field reconstruction.

E Limitations and Broader Impact

Limitations. Although Carta consistently improves spatial gene expression prediction across two benchmarks, several limitations remain.

First, our current evaluation focuses on task-specific top-50 highly variable genes rather than full-transcriptome prediction. Extending Carta to substantially larger gene sets would require further study of scalability and gene-level generalization.

Second, Carta uses a coordinate-derived hierarchy together with fixed interpolation between scales. While this design is effective for the assay layouts considered in our experiments, different spatial geometries, resolutions, or cross-platform settings may require adapting the hierarchy construction and interpolation scheme.

Third, Carta relies on frozen pathology foundation-model features to represent histology patches. Although these features provide strong morphology-aware representations, prediction quality can still depend on the domain coverage and representation quality of the underlying pathology encoder.

Finally, the scale-wise autoregressive reconstruction process introduces additional computation compared with direct single-scale prediction. In practice, Carta remains efficient because the hierarchy is compact and the pathology encoder is frozen, but very large spatial layouts may still increase memory and inference costs.

Broader impact. Carta may help reduce the cost of exploratory spatial molecular profiling by estimating spatial gene expression from routinely available histology images. This could support hypothesis generation and help researchers select samples for follow-up spatial transcriptomics experiments. However, the predictions are computational estimates rather than measured molecular profiles. If used without biological validation, incorrect predictions may mislead downstream interpretation, especially in clinically sensitive settings. Therefore, Carta should be used as a research tool for exploratory analysis rather than as a replacement for molecular assays or clinical decision making.

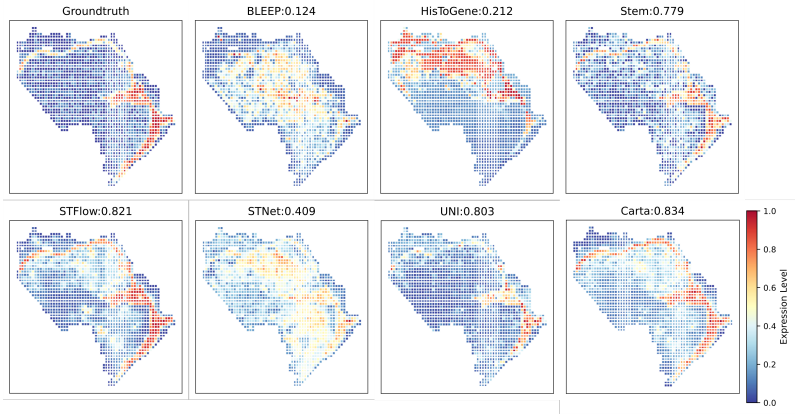

Figure 6: Qualitative comparison on KRTDAP expression prediction at TENX115.

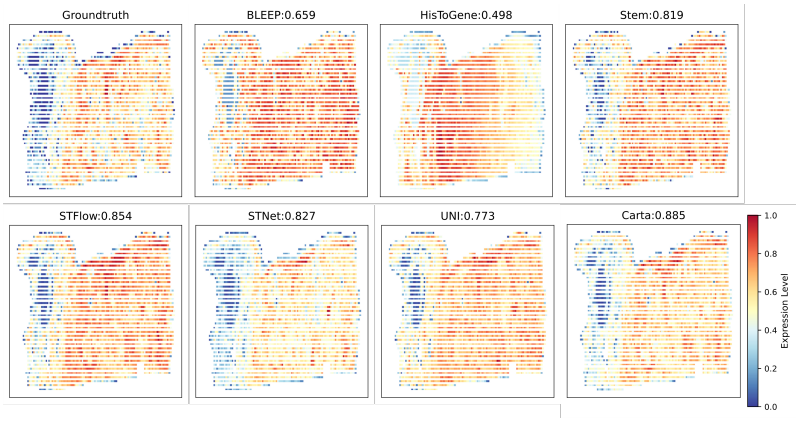

Figure 7: Qualitative comparison on PTPRC expression prediction at TENX141.

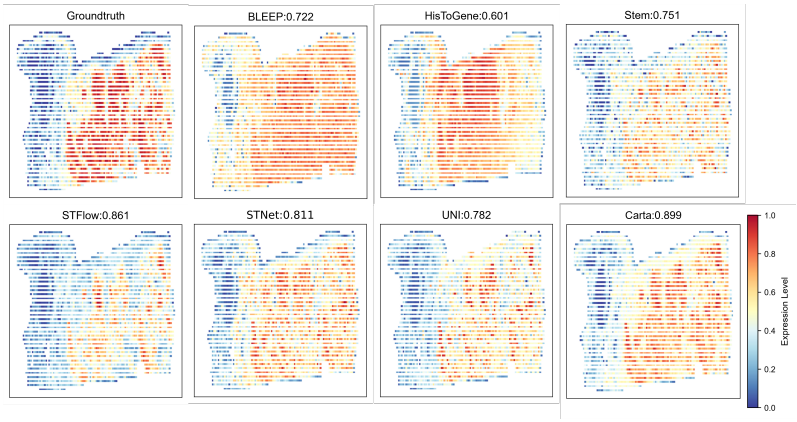

Figure 8: Qualitative comparison on MET expression prediction at TENX141.

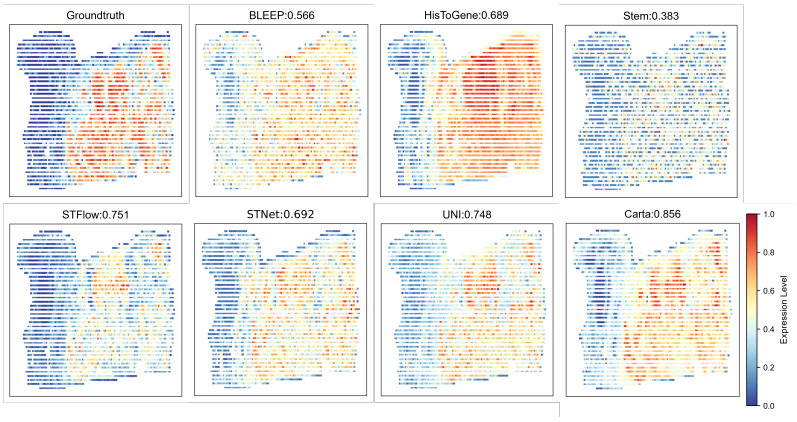

Figure 9: Qualitative comparison on UBE2C expression prediction at TENX141.

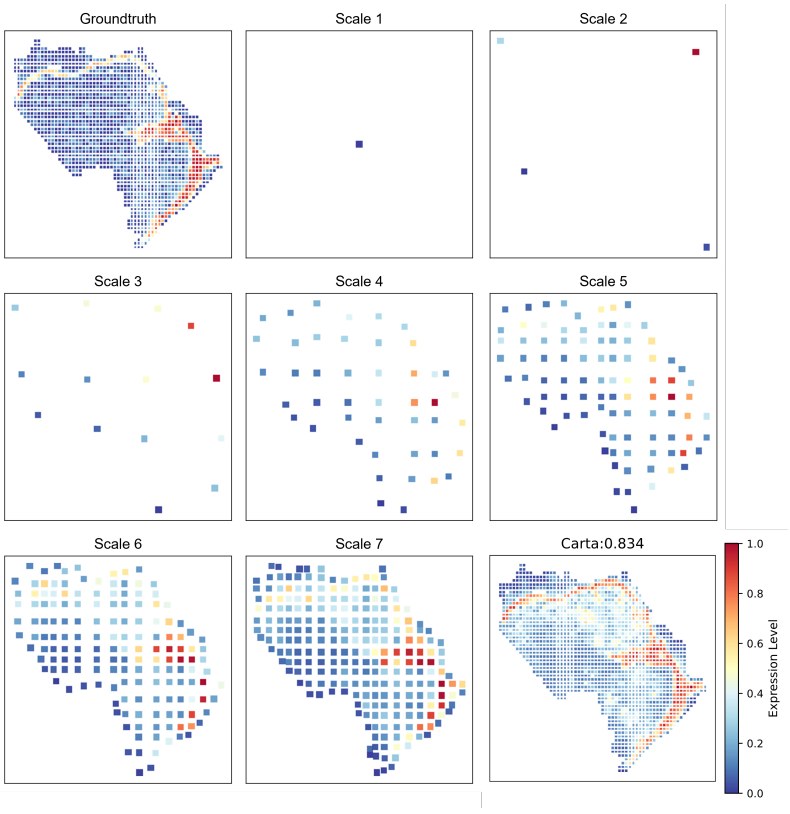

Figure 10: Coarse-to-fine reconstruction of KRTDAP expression maps by Carta at TENX115.

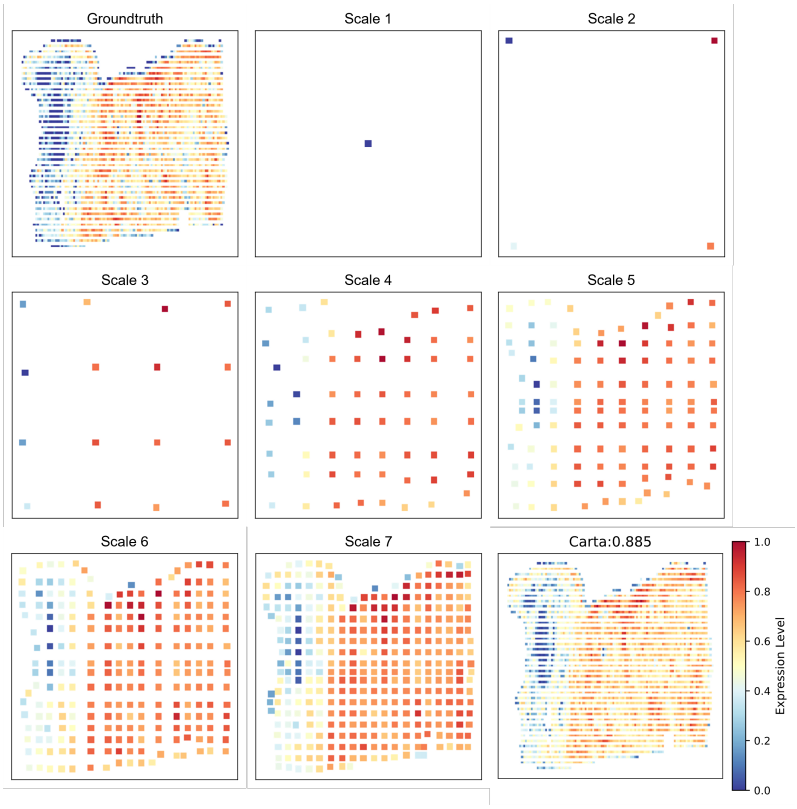

Figure 11: Coarse-to-fine reconstruction of PTPRC expression maps by Carta at TENX141.

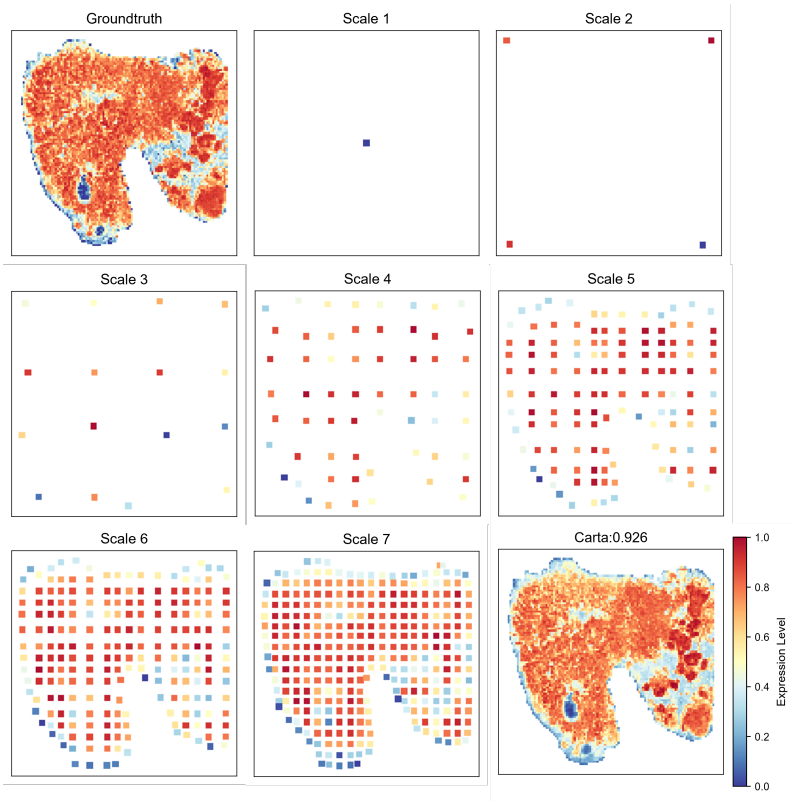

Figure 12: Coarse-to-fine reconstruction of ERBB2 expression maps by Carta at TENX95.