KNOWLEDGE-GUIDED SEMANTIC CONFUSION-AWARE AND RELIABLE SPATIAL GRADIENT LEARNING FOR CELL ABUNDANCE PREDICTION FROM H&E IMAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate prediction of the abundance and spatial distribution of different cell types within tissues is essential for characterizing tissue heterogeneity and disease microenvironments. Existing two-stage methods are susceptible to error propagation from intermediate gene expression prediction, whereas one-stage methods that directly predict cell abundance from H&E images still suffer from insufficient modeling of cell-type confusion and inadequate preservation of local spatial structures. To address these issues, we propose a Knowledge-Guided Semantic Confusion-Aware and Reliable Spatial Gradient Learning framework for cell abundance prediction from H&E images. Specifically, we develop a knowledge-guided cell semantic confusion-aware relative abundance learning mechanism to reduce abundance confusion among biologically related cell types. We further propose a reliable region-aware spatial gradient consistency constraint to improve the spatial consistency of local cell abundance variations. Experiments on multiple public spatial transcriptomics datasets demonstrate that the proposed method effectively improves fine-grained cell abundance prediction and enhances the recovery of spatial cell distributions.

1 INTRODUCTION

The composition and spatial distribution of different cell types within tissues are fundamental to understanding tissue function, intercellular communication, and disease progression. Spatial transcriptomics enables gene expression profiling while preserving tissue architecture (Ståhl et al., 2016) and, when integrated with single-cell transcriptomic reference data, can resolve cellular compositions across spatial locations (Li et al., 2022; Ma & Zhou, 2022; Kleshchevnikov et al., 2022), providing an important tool for studying tissue heterogeneity and cellular microenvironments. However, spatial transcriptomics experiments are costly and involve complex workflows, which limits their application in large-scale studies and routine clinical settings. In contrast, hematoxylin and eosin (H&E)-stained images are inexpensive, readily accessible, and widely available in clinical practice. Leveraging these advantages, recent studies have explored predicting spatial molecular and cellular information directly from H&E images and achieved encouraging results (He et al., 2020; Zeng et al., 2022; Yan et al., 2025; Zhao et al., 2026), making H&E-based spatial cellular inference an increasingly important direction in computational pathology and spatial biology.

Existing methods mainly follow a two-stage prediction strategy: spatial gene expression is first predicted from H&E images (He et al., 2020; Monjo et al., 2022; Zeng et al., 2022; Jia et al., 2024), and cell-type abundances at different locations are then estimated through cell deconvolution (Li et al., 2022; Ma & Zhou, 2022). However, errors in gene expression prediction may propagate to the subsequent deconvolution process, and the intermediate task is not fully aligned with the ultimate goal of cell abundance prediction. To reduce errors introduced by this intermediate step, recent studies have begun to explore one-stage approaches that directly predict fine-grained cell abundances from H&E images (Yan et al., 2025; Zhao et al., 2025; 2026), allowing models to learn a direct mapping between tissue morphology and cellular composition.

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.

Despite avoiding progressive error propagation, one-stage approaches still face two key challenges in fine-grained cell abundance prediction. First, existing methods typically treat different cell types as parallel prediction targets and lack explicit modeling of potential confusion relationships among cell types. Cell types with similar lineages, functions, or tissue contexts may exhibit similar morphological characteristics, which can lead to abundance misallocation. Second, spot-level supervision cannot explicitly constrain local spatial variations in cell abundance, potentially resulting in inaccurate localization of enriched regions, blurred tissue boundaries, and noisy background predictions. In contrast, simple neighborhood smoothing may obscure genuine regional boundaries. Therefore, simultaneously improving fine-grained cell-type discrimination and local spatial structure recovery remains a key challenge.

To address these issues, we propose a Knowledge-Guided Semantic Confusion-Aware and Reliable Spatial Gradient Learning method for directly predicting cell abundance from H&E images. To reduce potential confusion among cell types, we develop a knowledge-guided cell semantic confusion-aware relative abundance learning mechanism (KGSC). KGSC constructs biologically grounded cell-type semantic representations from single-cell reference data and combines semantic similarity with abundance correlations in the training data to identify potentially confusable cell-type pairs. Relative abundance constraints are then imposed to correct overly similar predictions and incorrect abundance ordering, thereby reducing abundance misallocation among fine-grained cell types. Furthermore, to preserve local spatial structures, we design a reliable region-aware spatial gradient consistency constraint (RSGC). RSGC identifies cell-type-specific enriched regions according to their abundance distributions and models spatial variations within enriched regions and across regional boundaries. The supervision strength is adaptively adjusted according to the reliability of the ground-truth spatial gradients, encouraging predicted abundance variations to remain consistent with the corresponding ground-truth spatial changes. In this way, RSGC suppresses background noise while preserving the continuity of enriched regions and the integrity of their boundaries.

The main contributions of this work are summarized as follows:

- We propose KGSC to improve fine-grained cell-type discrimination through biological semantic modeling, confusion-pair mining, and relative abundance correction.
- We develop RSGC to improve the local spatial structural accuracy of predicted cell abundance maps through region-aware modeling and reliable gradient supervision.
- We conduct systematic experiments on multiple public spatial transcriptomics datasets to validate the effectiveness of the proposed method for fine-grained cell abundance prediction and spatial distribution recovery.

2 Related Work

2.1 Histopathology Image-Based Prediction of Spatial Cellular Composition

Early methods for inferring spatial cellular composition from H&E images typically followed a two-stage pipeline: spatial gene expression was first predicted from histopathology images and then combined with single-cell reference data for cell deconvolution. For gene expression prediction, ST-Net used convolutional neural networks to regress expression from image patches (He et al., 2020), while HisToGene (Pang et al., 2021), Hist2ST (Zeng et al., 2022), THItoGene (Jia et al., 2024), and SEPAL (Mejia et al., 2023) incorporated Transformers, graph neural networks, or attention mechanisms to capture spatial context. BLEEP (Xie et al., 2023) and mclSTExp (Min et al., 2024) further employed contrastive learning to align histology and transcriptomic representations, while ST-Align introduced multi-scale image–gene alignment to capture both local and broader tissue contexts. Cell2location (Kleshchevnikov et al., 2022), RCTD (Cable et al., 2022), Stereoscope (Andersson et al., 2020), SPOTlight (Elosua-Bayes et al., 2021), and CARD (Ma & Zhou, 2022) represent widely used approaches for inferring cellular composition from spatial expression profiles with single-cell references.

Although effective, two-stage methods depend heavily on the accuracy of intermediate gene expression prediction. Prediction errors may propagate to the deconvolution stage, while the intermediate expression objective is not fully aligned with the final goal of cell abundance estimation.

Recent studies have therefore explored direct image-based prediction of cell abundance or spatial cellular distributions. Hist2Cell directly predicts spot-level fine-grained cell abundance by jointly modeling morphology, local neighborhoods, and long-range tissue context (Zhao et al., 2026). HPCell further estimates cell-type abundance directly from histology images by integrating pathology foundation features, hypergraph-based spatial modeling, and Transformer-based long-range dependencies (Zhao et al., 2025). CUCA integrates morphological and molecular information during training through cross-modal representation learning to improve fine-grained cell abundance prediction (Yan et al., 2025). In parallel, STASCAN predicts fine-resolution cellular distributions in captured and unmeasured tissue regions (Wu et al., 2024), whereas HistoCell infers super-resolution cell spatial profiles at the single-nucleus level from histology images. These studies have promoted a shift from indirect two-stage pipelines toward direct image-based inference.

However, existing spot-level one-stage methods mainly supervise the absolute abundance of individual cell types and rarely model potential confusion among fine-grained cell types. This limits their ability to correct abundance misallocation and relative ordering errors among confusable cell types. To address this issue, we identify potential confusion pairs using cell-type semantics and abundance correlations, and introduce relative abundance learning to improve fine-grained cell-type discrimination.

2.2 Spatial Context Modeling and Spatial Distribution Preservation

Cellular composition in tissues exhibits clear spatial organization. To capture spatial context, existing methods commonly employ graph neural networks, Transformers, and multi-scale feature learning. Graph-based models such as Hist2ST and SEPAL aggregate local tissue information through spatial adjacency (Zeng et al., 2022; Mejia et al., 2023), while Transformer-based approaches model longer-range dependencies across spatial locations (Pang et al., 2021; Zeng et al., 2022; Jia et al., 2024). More recent methods further integrate local neighborhoods with broader tissue context or multi-scale representations to capture spatial information at different levels (Zhao et al., 2026; 2025). Some studies additionally infer cellular distributions in unmeasured regions or at the single-nucleus level to increase the spatial resolution of cellular maps (Wu et al., 2024; Zhang et al., 2025).

Despite these advances, spatial context modeling alone does not ensure accurate recovery of local cell abundance variations. Existing spot-level methods mainly rely on point-wise abundance supervision (Zhao et al., 2026; Yan et al., 2025; Zhao et al., 2025) and lack explicit constraints on abundance changes between neighboring locations, which may lead to displaced enriched regions, distorted local gradients, and blurred boundaries. Simple neighborhood smoothing may also suppress genuine transitions between enriched regions and background.

Moreover, cell types differ in abundance scales and spatial patterns, making uniform neighborhood constraints unsuitable for different spatial regions. To address this limitation, we explicitly model abundance variations within enriched regions and across their boundaries using region-aware spatial gradient supervision, thereby preserving local continuity while maintaining genuine regional boundaries.

3 Methodology

As shown in Figure 1, we propose a Knowledge-Guided Semantic Confusion-Aware and Reliable Spatial Gradient Learning framework for direct spot-level cell abundance prediction from H&E images. The model integrates spot-level, local, and global tissue information, while KGSC and RSGC reduce abundance misallocation among potentially confusable cell types and preserve local spatial variations, respectively.

3.1 Multi-level Tissue Representation and Cell Abundance Prediction

Given H&E image patches and their spatial adjacency, ResNet18 first extracts morphological features F_{s}, as shown in Figure 1. A linear projection produces the spot-level representation F_{spot}, while GATv2 aggregates spatial neighborhood information to obtain the local representation F_{local}. A Vision Transformer further models long-range spatial dependencies from F_{local} to generate the global representation F_{global}. The three representations are equally fused as

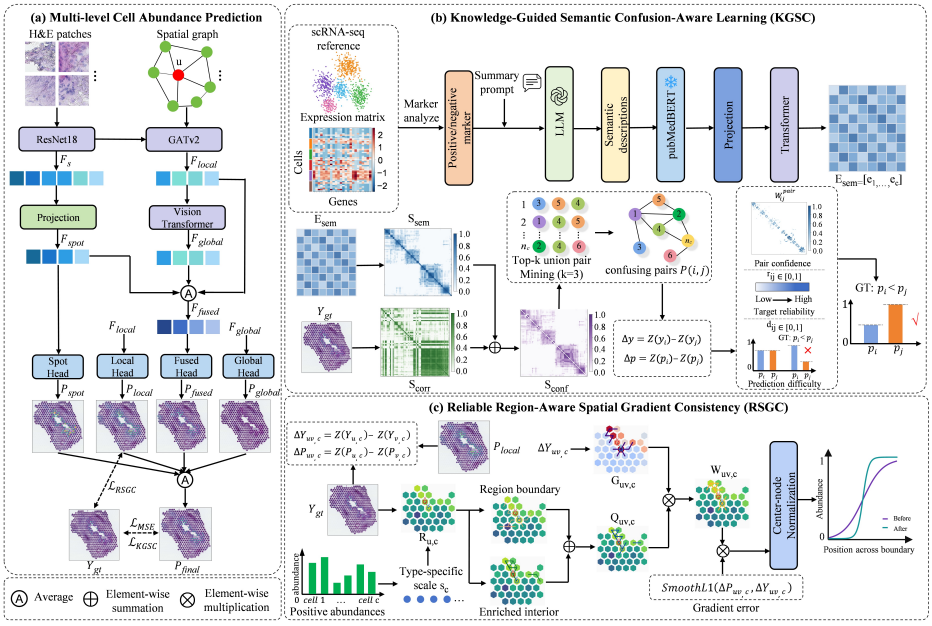

Figure 1: Overview of the proposed framework with KGSC and RSGC.

$$F_{\text{fused}} = \frac{F_{\text{spot}} + F_{\text{local}} + F_{\text{global}}}{3}. \tag{1}$$

These representations yield the spot-level, local, global, and fused predictions P_{spot}, P_{local}, P_{global}, and P_{fused}, respectively. The final cell abundance prediction is obtained by equal averaging:

$$P_{\text{final}} = \frac{P_{\text{spot}} + P_{\text{local}} + P_{\text{global}} + P_{\text{fused}}}{4}. \tag{2}$$

P_{final} serves as the final abundance prediction and is supervised by the basic regression objective and KGSC during training, whereas RSGC is applied only to P_{local} for local spatial gradient supervision.

3.2 Knowledge-Guided Cell Semantic Confusion-Aware Relative Abundance Learning

Existing spot-level regression methods mainly supervise the absolute abundance of each cell type independently, without explicitly accounting for potential confusion among biologically related cell types. We therefore propose KGSC, which identifies potentially confusable cell-type pairs and constrains their relative abundance relationships.

We first extract positive and negative marker evidence from cell-type-annotated scRNA-seq reference data. The marker evidence and cell-type information are then provided to a large language model (LLM) to generate structured biological descriptions covering six semantic fields: cell lineage, positive marker evidence, negative marker evidence, biological function, phenotype, and spatial context. A frozen PubMedBERT followed by semantic interaction modeling is subsequently used to obtain the representation e_c of each cell type. The semantic similarity between cell types i and j is defined as $S_{\text{sem}}(i, j) = \max(0, \cos(e_i, e_j))$. To account for dataset-specific confusion patterns, it is further calibrated using the nonnegative Pearson abundance correlation S_{corr} computed from the training set:

$$S_{\text{conf}}(i, j) = S_{\text{sem}}(i, j) \left[(1 - \alpha) + \alpha S_{\text{corr}}(i, j)\right], \tag{3}$$

where α controls the contribution of abundance correlation. For each cell type, the top-k related cell types are selected, and the deduplicated union of these pairs forms the confusion set $\mathcal{P}_{\mathrm{conf}}$. Each selected pair is further assigned a confidence weight according to its normalized confusion score.

To directly supervise relative abundance relationships, the ground-truth and predicted abundances are standardized using training-set statistics. For a confusion pair (i, j) at spot u, their relative abundance differences are defined as

$$\Delta y_{u,ij} = z_{u,i} - z_{u,j}, \qquad \Delta p_{u,ij} = \hat{z}_{u,i} - \hat{z}_{u,j}. \tag{4}$$

The supervision strength jointly considers pair confidence, target reliability, and prediction difficulty:

$$w_{u,ij}^{\mathrm{KGSC}} = w_{ij}^{\mathrm{pair}} r_{u,ij} d_{u,ij}, \tag{5}$$

where $r_{u,ij}$ increases with the magnitude of the ground-truth relative difference, while $d_{u,ij}$ emphasizes overly similar predictions and incorrect relative ordering. The KGSC loss is defined as

$$\mathcal{L}_{\mathrm{KGSC}} = \frac{\sum_{u \in \mathcal{V}_{\mathrm{center}}} \sum_{(i,j) \in \mathcal{P}_{\mathrm{conf}}} w_{u,ij}^{\mathrm{KGSC}} \, \mathrm{SmoothL1}\left(\Delta p_{u,ij}, \Delta y_{u,ij}\right)}{\sum_{u \in \mathcal{V}_{\mathrm{center}}} \sum_{(i,j) \in \mathcal{P}_{\mathrm{conf}}} w_{u,ij}^{\mathrm{KGSC}} + \epsilon}. \tag{6}$$

KGSC therefore corrects the relative abundance relationships of potentially confusable cell types rather than forcing semantically related cell types to have similar absolute abundances, thereby reducing abundance misallocation and relative ordering errors.

3.3 Reliable Region-Aware Spatial Gradient Consistency Constraint

Although GATv2 aggregates spatial neighborhood information, it does not directly ensure accurate recovery of local cell abundance variations. We therefore propose RSGC to explicitly supervise spatial variations of the local branch P_{local}.

For cell type c, let $y_{u,c}$ denote its ground-truth abundance at spot u. A type-specific scale s_c is estimated from the positive training abundances, and the soft enrichment score is defined as

$$R_{u,c} = \frac{y_{u,c}^{+}}{y_{u,c}^{+} + s_c + \epsilon}, \qquad y_{u,c}^{+} = \max(y_{u,c}, 0). \tag{7}$$

For neighboring spots u and v, enriched-region interiors and boundaries are jointly modeled using

$$Q_{uv,c} = \mathrm{clip}\left[\min(R_{u,c}, R_{v,c}) + \beta_b |R_{u,c} - R_{v,c}|, 0, 1\right], \tag{8}$$

where β_b controls the contribution of the enrichment–background boundary term. The two terms emphasize enriched-region interiors and enrichment–background boundaries, respectively.

The ground-truth abundance and local prediction are then standardized for each cell type. For a spatial edge (u, v), their gradients are defined as

$$\Delta Y_{uv,c} = z_{u,c} - z_{v,c}, \qquad \Delta P_{uv,c} = \hat{z}_{u,c}^{\mathrm{local}} - \hat{z}_{v,c}^{\mathrm{local}}. \tag{9}$$

A gradient reliability term $G_{uv,c}$ is derived from the magnitude of the ground-truth gradient, and the final spatial weight is given by

$$W_{uv,c} = Q_{uv,c} G_{uv,c}, \tag{10}$$

which emphasizes reliable spatial variations within enriched regions and across their boundaries while reducing the influence of weak gradients.

For each center spot u, we define the accumulated spatial weight as $D_u = \sum_{v\in\mathcal{N}(u)} \sum_{c=1}^{C} W_{uv,c}$, and retain center spots with valid spatial supervision as $\mathcal{V}^* = \{u \in \mathcal{V}_{\text{center}} \mid D_u > 10^{-8}\}$. The RSGC loss matches predicted and ground-truth gradients using a weighted SmoothL1 loss with center-node normalization:

$$\mathcal{L}_{\text{RSGC}} = \frac{1}{|\mathcal{V}^*|} \sum_{u\in\mathcal{V}^*} \frac{\sum_{v\in\mathcal{N}(u)} \sum_{c=1}^{C} W_{uv,c}\, \text{SmoothL1}\left(\Delta P_{uv,c}, \Delta Y_{uv,c}\right)}{D_u}. \tag{11}$$

Here, $\mathcal{N}(u)$ denotes the spatial neighbors of center spot u. The center-node normalization reduces sensitivity to variations in the number of valid spatial edges associated with different center spots. Rather than simply smoothing neighboring predictions, RSGC explicitly encourages the predicted local abundance variations to follow reliable ground-truth spatial gradients within enriched regions and across their boundaries.

3.4 Overall Loss Function

Let Y denote the ground-truth cell abundance matrix. The basic regression loss is defined as $\mathcal{L}_{\text{MSE}} = \text{MSE}(P_{\text{final}}, Y)$. The overall objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MSE}} + \lambda_{\text{KGSC}}(t)\mathcal{L}_{\text{KGSC}} + \lambda_{\text{RSGC}}(t)\mathcal{L}_{\text{RSGC}}. \tag{12}$$

$\mathcal{L}_{\text{MSE}}$ and $\mathcal{L}_{\text{KGSC}}$ supervise P_{final}, whereas $\mathcal{L}_{\text{RSGC}}$ is applied only to P_{local}. A linear warm-up strategy gradually increases the weights of the two auxiliary losses, allowing the model to first learn stable abundance predictions before introducing cell-type relation and local spatial structure supervision.

Cell semantic knowledge and both auxiliary constraints are used only during training. At inference, only H&E image patches and their spatial adjacency are required.

4 Experiments

4.1 Datasets

We evaluated the proposed method on three public spatial transcriptomics datasets: Human Lung, HER2ST, and STNet. The Human Lung dataset (Madissoon et al., 2023) contains 11 lung tissue sections from four donors, comprising 20,770 spatial spots across the trachea, bronchi, and different lung parenchymal regions. It includes 80 fine-grained cell types and exhibits substantial variation in tissue morphology and cellular composition across regions. The HER2ST (Andersson et al., 2021) dataset consists of 36 tissue sections and 13,620 spatial spots from eight patients with HER2-positive breast cancer. It covers different patients and tumor regions with considerable morphological and microenvironmental heterogeneity. Using the corresponding single-cell reference data, we obtained spot-level abundances for 39 cell types. The STNet dataset (He et al., 2020) contains H&E images and matched spatial transcriptomics data from 23 breast cancer patients, with a total of 30,655 spatial spots and pronounced inter-patient heterogeneity. As in HER2ST, spatial abundances of 39 cell types were used as prediction targets.

All three datasets were processed using a unified pipeline (Zhao et al., 2026). Data splitting was performed at the tissue-section level, with separate sections used for training and testing, thereby preventing spots from the same section from appearing in both sets.

4.2 Implementation Details and Evaluation Metrics

All experiments were implemented using PyTorch and PyTorch Geometric and conducted on a single NVIDIA GeForce RTX 3090 GPU. An ImageNet-pretrained ResNet18 was used as the image encoder, GATv2 was employed to aggregate local spatial neighborhood information, and a Vision

Table 1: Performance comparison of different methods on the Human Lung, HER2ST, and STNet datasets. Results are reported as mean ± standard deviation over three random seeds.

Method	PCC ↑			SCC ↑		
	Lung	HER2ST	STNet	Lung	HER2ST	STNet
ST-Net	0.281 ± 0.003	0.468 ± 0.019	0.263 ± 0.001	0.240 ± 0.002	0.515 ± 0.005	0.294 ± 0.007
DeepSpaCE	0.060 ± 0.004	0.435 ± 0.010	0.254 ± 0.045	0.042 ± 0.008	0.522 ± 0.005	0.267 ± 0.037
Hist2ST	0.079 ± 0.025	0.282 ± 0.022	0.134 ± 0.034	0.077 ± 0.013	0.371 ± 0.024	0.132 ± 0.044
THItoGene	0.112 ± 0.021	0.237 ± 0.025	0.112 ± 0.061	0.177 ± 0.022	0.295 ± 0.013	0.107 ± 0.065
Hist2Cell	0.334 ± 0.010	0.542 ± 0.015	0.260 ± 0.028	0.321 ± 0.007	0.494 ± 0.013	0.288 ± 0.034
Ours	**0.345** ± **0.004**	**0.551** ± **0.020**	**0.271** ± **0.005**	**0.355** ± **0.013**	**0.524** ± **0.024**	**0.302** ± **0.008**

Table 2: Ablation study evaluating the effectiveness of KGSC and RSGC on the lung, HER2ST, and STNet datasets.

KGSC	RSGC	PCC			SCC		
		Lung	HER2ST	STNet	Lung	HER2ST	STNet
×	×	0.326	0.552	0.268	0.307	0.509	0.295
✓	×	0.337	0.555	0.275	0.352	0.510	0.307
×	✓	0.338	0.561	0.277	0.343	0.523	0.304
✓	✓	**0.341**	**0.568**	**0.283**	**0.366**	**0.529**	**0.310**

Transformer was used to model long-range tissue context. The graph network sampled 2-hop spatial neighborhoods.

The model was trained using the Adam optimizer with an initial learning rate and weight decay of 1×10^{-4}, together with a CosineAnnealing learning-rate schedule. In KGSC, the top-k potentially confusable cell types were selected for each cell type. Both KGSC and RSGC were gradually introduced into the training objective using linear warm-up. Cell semantic knowledge and the two auxiliary constraints were used only during training. During inference, only H&E image patches and their spatial adjacency relationships were required.

We evaluated cell abundance prediction using the Pearson correlation coefficient (PCC) and Spearman correlation coefficient (SCC). PCC measures the linear correlation between predicted and ground-truth abundances, whereas SCC evaluates the consistency of their abundance rankings. Higher values indicate better performance.

4.3 Comparison with Representative Methods

4.3.1 Quantitative Comparison

We compared the proposed method with ST-Net (He et al., 2020), DeepSpaCE (Monjo et al., 2022), Hist2ST (Zeng et al., 2022), THItoGene (Jia et al., 2024), and Hist2Cell (Zhao et al., 2026) under the same data splits and cell abundance prediction settings. Table 1 reports the mean and standard deviation of PCC and SCC over three independent random seeds.

As shown in Table 1, our method achieves the highest mean PCC and SCC on all three datasets. These results demonstrate the consistent effectiveness of our method across different datasets, indicating that jointly modeling potentially confusable cell-type relationships and local spatial variations improves fine-grained cell abundance prediction.

4.3.2 Qualitative Comparison of Spatial Distributions

To further evaluate spatial distribution recovery, we selected four representative cell types from a Human Lung test section and visualized their ground-truth spatial distributions together with the predictions of the comparison methods and the proposed method.

As shown in Figure 2, some comparison methods exhibit displaced enrichment regions, scattered local responses, or nonspecific predictions in low-abundance background areas. In contrast, the pro-

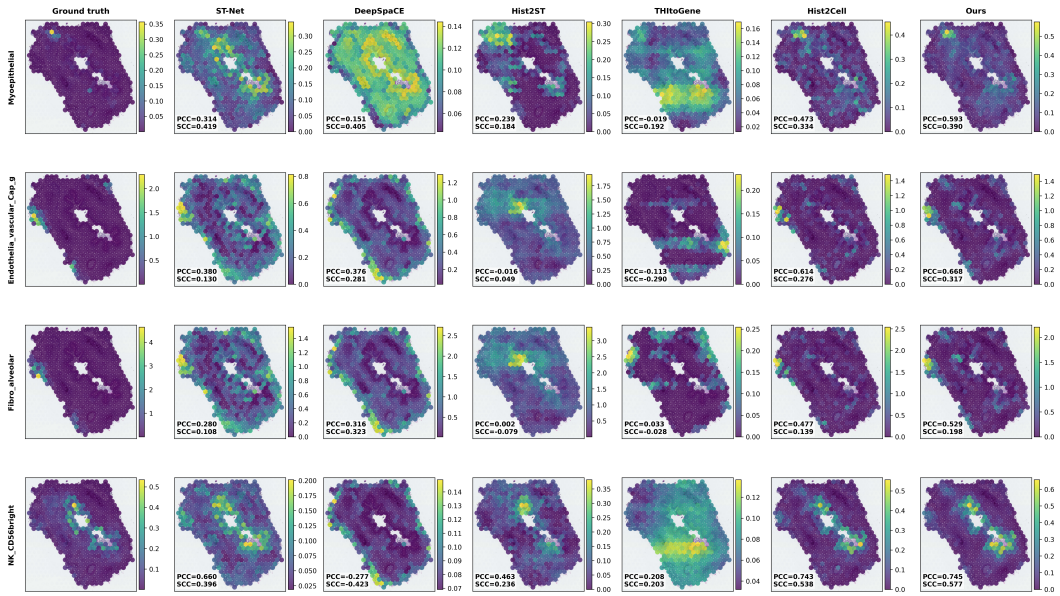

Figure 2: Qualitative comparison of spatial cell abundance distributions on the Human Lung dataset.

posed method produces enrichment regions that are more consistent with the ground truth and better preserves local high-abundance regions and their surrounding spatial variations. These qualitative results are consistent with the quantitative comparison, demonstrating that the proposed method improves both cell abundance prediction and spatial distribution recovery.

4.4 Ablation Study and Analysis

We conduct ablation experiments on KGSC and RSGC under the same experimental settings, as shown in Table 2. Both KGSC and RSGC improve model performance when introduced individually. KGSC reduces errors among potentially confusable cell types through semantic-confusion-aware relative abundance supervision, while RSGC improves the modeling of local spatial abundance variations. Combining the two modules leads to further performance gains, suggesting that cell-type relation modeling and local spatial structure supervision provide complementary benefits for cell abundance prediction.

4.5 Analysis of the KGSC Mechanism

To further analyze the effect of KGSC, we compared the KGSC-only model with the baseline Hist2Cell. As shown in Figure 3A, KGSC improves the PCC of multiple cell types, indicating enhanced cell-type abundance prediction. More pronounced improvements are observed in SCC (Figure 3B), suggesting that KGSC effectively improves the rank consistency of predicted cell abundances.

We further evaluated the relative-abundance differences of the semantic-confusion cell pairs used by KGSC. As shown in Figure 3C, most confusion pairs exhibit improved SCC between the predicted and ground-truth relative-abundance differences. These results indicate that KGSC mainly enhances cell-type abundance ranking and improves relative-abundance consistency among potentially confusable cell types.

4.6 Spatial Gradient Analysis of RSGC

To investigate the effect of RSGC on local spatial structure, we compared the RSGC-only model with the baseline Hist2Cell. Since RSGC directly supervises the local graph branch P_{local}, the corresponding local predictions were used for analysis.

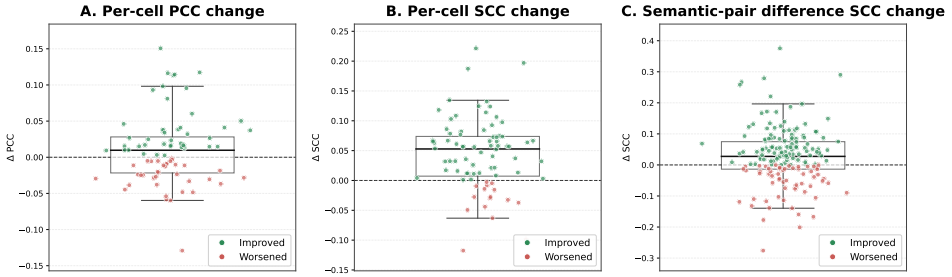

Figure 3: Effect of KGSC on cell-type prediction and semantic-confusion pair relative-abundance consistency.

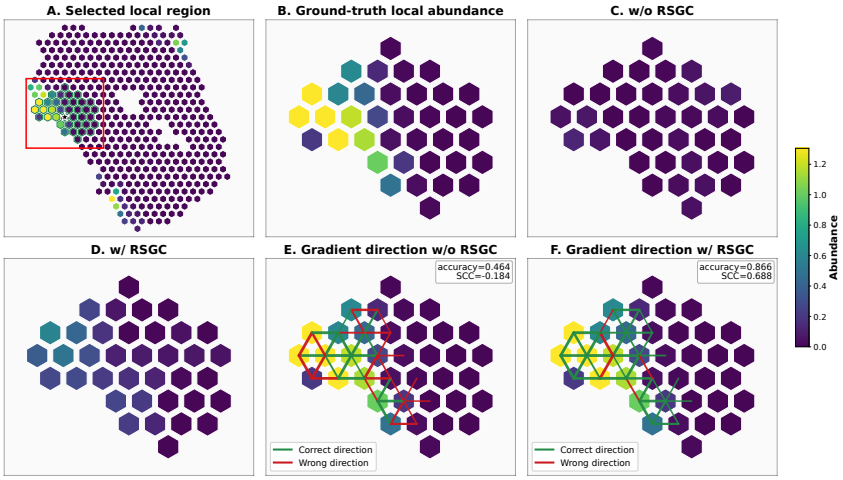

Figure 4: Analysis of RSGC on local abundance recovery and spatial gradient consistency.

As shown in Figure 4A–D, a representative region containing both cell-enriched and low-abundance areas was selected from a test section. Compared with the baseline, RSGC produces an abundance pattern that better agrees with the ground truth, with a clearer transition between enriched and background regions.

Figure 4E and Figure 4F further compare the local gradient directions along spatial adjacency edges. Green and red edges denote gradient directions that are consistent and inconsistent with the ground truth, respectively. RSGC yields more correctly aligned gradients, together with higher gradient-direction accuracy and Spearman correlation. These results indicate that RSGC improves local spatial-gradient preservation rather than simply smoothing neighboring predictions, thereby better recovering spatial abundance variations and their boundaries.

5 Conclusion

In this work, we proposed a knowledge-guided semantic confusion-aware and reliable spatial gradient learning method for fine-grained cell abundance prediction from H&E images. KGSC reduces abundance misallocation and relative ordering errors among potentially confusable cell types, while RSGC improves the recovery of local spatial abundance variations within enriched regions and across their boundaries. Experiments on multiple public spatial transcriptomics datasets demonstrate that the proposed method consistently improves cell abundance prediction and spatial distribution recovery.

AI USE STATEMENT

In this work, an LLM was used as part of the proposed pipeline of cell semantic knowledge construction to generate structured biological descriptors from marker-gene evidence and cell-type information. Specifically, GPT-5.5 was used in our implementation, and the generated descriptors were subsequently encoded by a frozen PubMedBERT model for semantic representation learning. Outside this methodological use, generative AI tools were used solely to assist with the polishing and refinement of the writing in this paper, including improving grammatical fluency, sentence structure, and overall clarity of the manuscript. All research ideation, methodological development, experimental design and execution, result analysis, and technical conclusions remain the work of the authors. The authors reviewed all AI-assisted content and take full responsibility for the final content presented in this work.

REFERENCES

Alma Andersson, Joseph Bergenstråhle, Michaela Asp, Ludvig Bergenstråhle, Aleksandra Jurek, José Fernández Navarro, and Joakim Lundeberg. Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Communications biology*, 3(1):565, 2020.

Alma Andersson, Ludvig Larsson, Linnea Stenbeck, Fredrik Salmén, Anna Ehinger, Sunny Z Wu, Ghamdan Al-Eryani, Daniel Roden, Alex Swarbrick, Åke Borg, et al. Spatial deconvolution of her2-positive breast cancer delineates tumor-associated cell type interactions. *Nature communications*, 12(1):6012, 2021.

Dylan M Cable, Evan Murray, Luli S Zou, Aleksandrina Goeva, Evan Z Macosko, Fei Chen, and Rafael A Irizarry. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nature biotechnology*, 40(4):517–526, 2022.

Marc Elosua-Bayes, Paula Nieto, Elisabetta Mereu, Ivo Gut, and Holger Heyn. Spotlight: seeded nmf regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic acids research*, 49(9):e50–e50, 2021.

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.

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.

Vitalii Kleshchevnikov, Artem Shmatko, Emma Dann, Alexander Aivazidis, Hamish W King, Tong Li, Rasa Elmentaite, Artem Lomakin, Veronika Kedlian, Adam Gayoso, et al. Cell2location maps fine-grained cell types in spatial transcriptomics. *Nature biotechnology*, 40(5):661–671, 2022.

Bin Li, Wen Zhang, Chuang Guo, Hao Xu, Longfei Li, Minghao Fang, Yinlei Hu, Xinye Zhang, Xinfeng Yao, Meifang Tang, et al. Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. *Nature methods*, 19(6):662–670, 2022.

Ying Ma and Xiang Zhou. Spatially informed cell-type deconvolution for spatial transcriptomics. *Nature biotechnology*, 40(9):1349–1359, 2022.

Elo Madissoon, Amanda J Oliver, Vitalii Kleshchevnikov, Anna Wilbrey-Clark, Krzysztof Polanski, Nathan Richoz, Ana Ribeiro Orsi, Lira Mamanova, Liam Bolt, Rasa Elmentaite, et al. A spatially resolved atlas of the human lung characterizes a gland-associated immune niche. *Nature genetics*, 55(1):66–77, 2023.

Gabriel Mejia, Paula Cárdenas, Daniela Ruiz, Angela Castillo, and Pablo Arbeláez. Sepal: spatial gene expression prediction from local graphs. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 2286–2295. IEEE, 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.

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.

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.

Ying Wu, Jia-Yi Zhou, Bofei Yao, Guanshen Cui, Yong-Liang Zhao, Chun-Chun Gao, Ying Yang, Shihua Zhang, and Yun-Gui Yang. Stascan deciphers fine-resolution cell distribution maps in spatial transcriptomics by deep learning. *Genome Biology*, 25(1):278, 2024.

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.

Chaoyang Yan, Zhihan Ruan, Songkang Chen, Yichen Pan, Xue Han, Yuanyu Li, and Jian Liu. Predicting fine-grained cell types from histology images through cross-modal learning in spatial transcriptomics. *Bioinformatics*, 41(Supplement_1):i115–i124, 2025.

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.

Peng Zhang, Chaofei Gao, Zhuoyu Zhang, Zhiyuan Yuan, Qian Zhang, Ping Zhang, Shiyu Du, Weixun Zhou, Yan Li, and Shao Li. Systematic inference of super-resolution cell spatial profiles from histology images. *Nature communications*, 16(1):1838, 2025.

Weiqin Zhao, Zhuo Liang, Xianjie Huang, Yuanhua Huang, and Lequan Yu. Hist2cell: deciphering fine-grained cellular architectures from histology images. *Cell Genomics*, 6(3), 2026.

Yongkang Zhao, Youyang Li, Weijiang Yu, Hongyu Zhang, Zheng Wang, Yuedong Yang, and Yuansong Zeng. Accurately predicting cell type abundance from spatial histology image through hpcell. *Interdisciplinary Sciences: Computational Life Sciences*, pp. 1–14, 2025.

A Additional Details of the Proposed Method

A.1 Cell Semantic Knowledge Construction

Cell-type semantic knowledge was constructed from the corresponding annotated single-cell reference data. During marker extraction, cell types containing fewer than 30 cells were discarded, while at most 2000 cells were sampled for each cell type. Genes expressed in fewer than 20 cells were removed. Expression values were normalized to a total count of 10,000 followed by log transformation. Up to 50 positive and 30 negative marker genes were retained for each cell type, using a minimum expression fraction of 0.1 and an absolute $\log_2$ fold-change threshold of 0.25. During descriptor generation, up to 30 positive and 15 negative markers were provided to an LLM together with cell-type metadata, including the cell-type name, species, tissue context, disease context, and the number of sampled reference cells. GPT-5.5 was used in our implementation to generate structured biological descriptions for each cell type. The resulting descriptions contained six semantic fields used for embedding: lineage, positive marker evidence, negative marker evidence, biological function, phenotype, and spatial context.

Table 3: Unified LLM prompt framework for cell semantic descriptor generation.

Component	Specification
Role definition	Biomedical scientist specializing in single-cell transcriptomics and tissue biology.
Task specification	Generate concise and structured biological descriptions for individual cell types from the supplied reference evidence.
Reference evidence	Cell-type identity, species, tissue, disease context, number of sampled reference cells, positive marker genes, and negative marker genes.
Generation constraints	Preserve cell-type identity; use only supplied marker evidence; avoid invented marker genes, unsupported disease associations, and overly specific spatial localization; generate concise, structured, and conservative biological descriptions.
Structured output	Lineage, positive marker evidence, negative marker evidence, biological function, phenotype, and spatial context.

LLM prompt framework. As summarized in Table 3, a unified prompt framework was used to generate cell-type semantic descriptions across all datasets. The framework consists of five components: role definition, task specification, reference evidence, generation constraints, and structured output specification. Dataset-specific information is supplied through placeholders for cell type, species, tissue, disease context, sampled reference-cell number, and positive and negative marker genes. The general prompt structure, evidence-grounding principles, and output schema are shared across datasets, while dataset-specific contextual information is adapted to the corresponding reference data. The LLM is instructed to generate conservative biological descriptions grounded in the supplied evidence and to avoid unsupported marker genes, disease associations, or overly specific spatial localization. The resulting structured descriptions contain six biological semantic fields used for subsequent semantic encoding: lineage, positive marker evidence, negative marker evidence, biological function, phenotype, and spatial context.

Field-wise semantic encoding. Each semantic field was independently encoded using a frozen PubMedBERT encoder. Let $\mathbf{H}_{c,f}^{(\ell)} \in \mathbb{R}^{T_{c,f} \times d_B}$ denote the hidden representations produced at layer ℓ for cell type c and semantic field f, where $T_{c,f}$ denotes the token length and $d_B = 768$. The representations from the last four PubMedBERT layers were first averaged:

$$\bar{\mathbf{H}}_{c,f} = \frac{1}{4} \sum_{\ell=L-3}^{L} \mathbf{H}_{c,f}^{(\ell)}. \tag{13}$$

Mean pooling was then performed over valid tokens, followed by ℓ_2 normalization:

$$\mathbf{h}_{c,f} = \mathrm{Norm}_2 \left(\frac{\sum_{t=1}^{T_{c,f}} m_{c,f,t} \bar{\mathbf{H}}_{c,f,t}}{\sum_{t=1}^{T_{c,f}} m_{c,f,t}} \right), \tag{14}$$

where $m_{c,f,t} \in \{0, 1\}$ denotes the attention mask and $\mathrm{Norm}_2(\mathbf{x}) = \mathbf{x}/\|\mathbf{x}\|_2$ for a nonzero vector.

To reduce field-shared directions across different cell types, field-wise centering was subsequently performed. For each semantic field f, we first compute

$$\mathbf{c}_{c,f} = \mathbf{h}_{c,f} - \frac{1}{C} \sum_{c'=1}^{C} \mathbf{h}_{c',f}, \tag{15}$$

and obtain the centered representation as

$$\tilde{\mathbf{h}}_{c,f} = \mathrm{Norm}_2 \left(\mathbf{c}_{c,f} \right), \tag{16}$$

where C denotes the number of cell types. Near-zero vectors are handled using the numerical fallback implemented in the semantic encoder.

The centered 768-dimensional representations were then mapped to a 256-dimensional semantic space using a PCA-initialized residual projection with a bottleneck dimension of 128. A one-layer Transformer with four attention heads, a feed-forward dimension of 512, and a dropout rate

of 0.1 was subsequently used to model cross-field semantic interactions. Reliability-aware common/specific field selection was then used to aggregate the contextualized field representations. The resulting common semantic representation was retained as the cell-type embedding used by KGSC to construct semantic-confusion relationships, whereas the specific representation was used only during semantic representation learning.

A.2 KNOWLEDGE-GUIDED SEMANTIC CONFUSION-AWARE LEARNING

The main text defines the semantic similarity, abundance-correlation calibration, semantic-confusion score, confusion-pair set, relative abundance differences, and the overall KGSC objective. Here, we provide the implementation details of the standardization, pair confidence, target reliability, and prediction-difficulty terms.

Training-statistic standardization. All abundance statistics used by KGSC were computed exclusively from the training set. Let μ_c and σ_c denote the training-set mean and raw standard deviation of cell type c, respectively. To avoid unstable standardization for cell types with extremely small variance, the standard-deviation floor used in the implementation is

$$\sigma_{\text{floor}} = \begin{cases} \max\left(0.1\,\text{Median}\left(\{\sigma_c \mid \sigma_c > 0\}\right), 10^{-6}\right), & \text{if at least one } \sigma_c > 0, \\ 10^{-6}, & \text{otherwise.} \end{cases} \tag{17}$$

The effective standard deviation is therefore

$$\bar{\sigma}_c = \max\left(\sigma_c, \sigma_{\text{floor}}\right). \tag{18}$$

Using the notation of the main text, the standardized ground-truth and predicted abundances are

$$z_{u,c} = \frac{y_{u,c} - \mu_c}{\bar{\sigma}_c}, \qquad \hat{z}_{u,c} = \frac{\hat{y}_{u,c} - \mu_c}{\bar{\sigma}_c}. \tag{19}$$

Accordingly, for each semantic-confusion pair $(i, j) \in \mathcal{P}_{\text{conf}}$, the relative abundance differences are

$$\Delta y_{u,ij} = z_{u,i} - z_{u,j}, \qquad \Delta p_{u,ij} = \hat{z}_{u,i} - \hat{z}_{u,j}, \tag{20}$$

which are identical to the relative-difference definitions used in the main text.

Confusion-pair confidence. As described in the main text, the top-k confusion candidates are independently selected for each cell type and their union is deduplicated to form $\mathcal{P}_{\text{conf}}$. The implementation further assigns each retained pair a normalized confidence weight:

$$w_{ij}^{\text{pair}} = \left(\frac{S_{\text{conf}}(i,j)}{\max\limits_{(p,q)\in\mathcal{P}_{\text{conf}}} S_{\text{conf}}(p,q)} \right)^2. \tag{21}$$

A numerical floor of 10^{-6} is applied to the denominator in the implementation. The squared normalization preserves the balanced per-cell top-k coverage while emphasizing pairs with stronger semantic-confusion evidence.

Target-difference reliability. The reliability of the ground-truth relative abundance difference is defined as

$$r_{u,ij} = \sigma\left(\frac{|\Delta y_{u,ij}| - m_d}{T_d} \right), \tag{22}$$

where m_d and T_d denote the target-difference margin and temperature, respectively. Thus, pairs with very small ground-truth abundance differences receive weaker relative-order supervision, whereas pairs with clearer differences receive larger reliability weights.

Prediction-difficulty weighting. Two complementary terms are used to identify difficult predictions. The first term emphasizes near-tie predictions:

$$d_{u,ij}^{\mathrm{tie}} = \exp\left(-\frac{|\Delta p_{u,ij}|}{T_p}\right), \tag{23}$$

where T_p is the near-tie temperature.

The second term emphasizes incorrect or insufficiently confident relative ordering:

$$d_{u,ij}^{\mathrm{ord}} = \sigma\left(\frac{m_s - \Delta p_{u,ij}\Delta y_{u,ij}}{T_s}\right), \tag{24}$$

where m_s and T_s denote the relative-order margin and temperature, respectively. In the implementation, $\Delta p_{u,ij}$ is detached when computing both $d_{u,ij}^{\mathrm{tie}}$ and $d_{u,ij}^{\mathrm{ord}}$, so the prediction-difficulty weights act only as adaptive supervision coefficients and do not introduce additional gradient paths.

The final prediction-difficulty weight is

$$d_{u,ij} = \max\left(d_{u,ij}^{\mathrm{tie}}, d_{u,ij}^{\mathrm{ord}}\right). \tag{25}$$

This construction prevents confidently incorrect relative ordering from being treated as an easy case. In particular, when the predicted and ground-truth relative differences have opposite signs,

$$\Delta p_{u,ij}\Delta y_{u,ij} \le 0. \tag{26}$$

Because $m_s > 0$ and $T_s > 0$, it follows that

$$\frac{m_s - \Delta p_{u,ij}\Delta y_{u,ij}}{T_s} > 0, \tag{27}$$

and therefore

$$d_{u,ij}^{\mathrm{ord}} > \frac{1}{2}. \tag{28}$$

Consequently,

$$d_{u,ij} \ge d_{u,ij}^{\mathrm{ord}} > \frac{1}{2}. \tag{29}$$

Hence, a prediction with reversed relative ordering cannot receive an arbitrarily small difficulty weight merely because the magnitude of its predicted difference is large.

Adaptive weighting and robust relative-abundance penalty. Following the definition in the main text, the complete adaptive weight is

$$w_{u,ij}^{\mathrm{KGSC}} = w_{ij}^{\mathrm{pair}} r_{u,ij} d_{u,ij}. \tag{30}$$

For completeness, let

$$e_{u,ij} = \Delta p_{u,ij} - \Delta y_{u,ij}. \tag{31}$$

The SmoothL1 penalty used by KGSC is

$$\ell_\beta(e) = \begin{cases} \dfrac{e^2}{2\beta}, & |e| < \beta, \\ |e| - \dfrac{\beta}{2}, & |e| \ge \beta. \end{cases} \tag{32}$$

Substituting the pair confidence, target reliability, and prediction difficulty defined above into $w_{u,ij}^{\mathrm{KGSC}}$ gives the weighted relative-abundance constraint defined in the main text. The weighting terms are treated as adaptive supervision coefficients rather than additional prediction targets, so KGSC reallocates optimization emphasis toward reliable and difficult semantic-confusion pairs while retaining the robust SmoothL1 regression objective.

A.3 Reliable Region-Aware Spatial Gradient Consistency

The main text defines the soft enrichment score $R_{u,c}$, the region-aware edge weight $Q_{uv,c}$, the standardized spatial gradients, the gradient reliability, and the RSGC objective. Here, we provide additional implementation details and analyze several properties of these components.

Cell-type-specific enrichment scale. For cell type c, define the positive training-abundance set as

$$\mathcal{Y}_c^+ = \{y_{u,c} \mid y_{u,c} > 0\}, \qquad n_c^+ = |\mathcal{Y}_c^+|. \tag{33}$$

The cell-type-specific scale used in the soft enrichment score is estimated exclusively from training data:

$$s_c = \begin{cases} \mathrm{Median}\left(\mathcal{Y}_c^+\right), & n_c^+ \geq n_{\min}, \\ \mathrm{Mean}\left(\mathcal{Y}_c^+\right), & 0 < n_c^+ < n_{\min}, \\ \mathrm{Median}\left(\bigcup_{k=1}^{C} \mathcal{Y}_k^+\right), & n_c^+ = 0, \end{cases} \tag{34}$$

where $n_{\min} = 8$.

The implementation additionally enforces

$$s_c \leftarrow \max\left(s_c, 10^{-4}\right). \tag{35}$$

Using the same notation as in the main text, the soft enrichment score is

$$R_{u,c} = \frac{y_{u,c}^+}{y_{u,c}^+ + s_c + \epsilon}, \qquad y_{u,c}^+ = \max(y_{u,c}, 0). \tag{36}$$

Region-aware spatial weighting. For a valid spatial edge (u, v), the region-aware weight follows the same definition as in the main text:

$$Q_{uv,c} = \mathrm{clip}\left[\min\left(R_{u,c}, R_{v,c}\right) + \beta_b |R_{u,c} - R_{v,c}|, 0, 1\right]. \tag{37}$$

The first term emphasizes edges whose two endpoints both lie in enriched regions, whereas the second term retains supervision around enrichment boundaries. This behavior can be illustrated using several special cases. If $R_{u,c} = R_{v,c} = r$, then

$$Q_{uv,c} = r. \tag{38}$$

Thus, enriched-region interior edges receive weights proportional to their enrichment strength. If both endpoints correspond to low-abundance background,

$$R_{u,c} \approx R_{v,c} \approx 0, \tag{39}$$

then

$$Q_{uv,c} \approx 0. \tag{40}$$

In contrast, for an idealized enrichment–background boundary with $R_{u,c} \approx 1$ and $R_{v,c} \approx 0$,

$$Q_{uv,c} \approx \beta_b. \tag{41}$$

Therefore, the region-aware weighting does not reduce RSGC to conventional neighborhood smoothing: low-information background edges are suppressed, while genuine enrichment boundaries remain explicitly supervised.

Valid spatial edges. RSGC is evaluated on the local GAT prediction branch. Under neighborhood sampling, not every edge in the sampled subgraph contributes to the RSGC objective. Consistent with the implementation, we define the valid directed edge set as

$$\mathcal{E}^* = \{(u, v) \in \mathcal{E}_{\mathrm{sub}} \mid u \in \mathcal{V}_{\mathrm{center}},\ u \neq v\}, \tag{42}$$

where $\mathcal{V}_{\mathrm{center}}$ denotes the center nodes of the sampled mini-batch and self-loops are excluded. Hence, RSGC constrains local gradients originating from center spots while neighboring sampled nodes provide the spatial reference required to form the gradient.

Reliable spatial-gradient weighting. Using the same training-set standardization as above, the ground-truth and local-branch spatial gradients for $(u, v) \in \mathcal{E}^*$ are defined, consistently with the main text, as

$$\Delta Y_{uv,c} = z_{u,c} - z_{v,c}, \qquad \Delta P_{uv,c} = \hat{z}^{\mathrm{local}}_{u,c} - \hat{z}^{\mathrm{local}}_{v,c}. \tag{43}$$

The gradient reliability is

$$G_{uv,c} = g_0 + (1 - g_0)\sigma\left(\frac{|\Delta Y_{uv,c}| - m_g}{T_g}\right), \tag{44}$$

where g_0 is the gradient-reliability floor, m_g is the gradient margin, and T_g is the corresponding temperature.

Let

$$a_{uv,c} = \frac{|\Delta Y_{uv,c}| - m_g}{T_g}. \tag{45}$$

Because $0 \le g_0 < 1$ and $T_g > 0$,

$$\frac{\partial G_{uv,c}}{\partial |\Delta Y_{uv,c}|} = \frac{1 - g_0}{T_g}\sigma(a_{uv,c})\left[1 - \sigma(a_{uv,c})\right] > 0, \tag{46}$$

the reliability weight increases monotonically with the magnitude of the ground-truth spatial gradient. Strong local abundance changes therefore receive larger reliability weights, whereas weak gradients are downweighted but are not completely discarded because $g_0 > 0$.

As defined in the main text, the final RSGC weight is

$$W_{uv,c} = Q_{uv,c} G_{uv,c}. \tag{47}$$

Center-node-normalized gradient objective. For each center node u, we define the accumulated RSGC weight as

$$D_u = \sum_{v:(u,v)\in\mathcal{E}^*} \sum_{c=1}^{C} W_{uv,c}. \tag{48}$$

The implementation retains center nodes whose accumulated weight exceeds the numerical threshold:

$$\mathcal{V}^* = \left\{u \in \mathcal{V}_{\mathrm{center}} \mid D_u > 10^{-8}\right\}. \tag{49}$$

For each $u \in \mathcal{V}^*$, the center-node-normalized gradient loss is defined as

$$\mathcal{L}^{\mathrm{grad}}_u = \frac{\displaystyle\sum_{v:(u,v)\in\mathcal{E}^*} \sum_{c=1}^{C} W_{uv,c}\ell_\beta\left(\Delta P_{uv,c} - \Delta Y_{uv,c}\right)}{D_u}. \tag{50}$$

The final RSGC objective is therefore

$$\mathcal{L}_{\mathrm{RSGC}} = \frac{1}{|\mathcal{V}^*|} \sum_{u\in\mathcal{V}^*} \mathcal{L}^{\mathrm{grad}}_u, \tag{51}$$

which corresponds to the center-node-normalized objective used in the implementation.

This normalization reduces the sensitivity of RSGC to variations in the number of valid sampled edges associated with different center nodes. To illustrate this property, let

$$\ell_{uv,c} = \ell_\beta\left(\Delta P_{uv,c} - \Delta Y_{uv,c}\right). \tag{52}$$

If every weighted edge contribution associated with a center node were duplicated m times, then

$$\frac{m\sum_{v,c} W_{uv,c}\ell_{uv,c}}{m\sum_{v,c} W_{uv,c}} = \frac{\sum_{v,c} W_{uv,c}\ell_{uv,c}}{\sum_{v,c} W_{uv,c}}. \tag{53}$$

Thus, a center spot does not receive a larger contribution to RSGC merely because more valid spatial edges are associated with it.

Algorithm 1 Training Procedure of the Proposed Framework

Require: H&E image patches $\mathcal{X}$; spatial graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$; cell-abundance labels $\mathbf{Y}$; annotated single-cell reference data $\mathcal{D}_{\mathrm{sc}}$
Ensure: Trained cell-abundance prediction model f_θ

Stage 1: Cell Semantic Knowledge Construction
Extract positive and negative marker evidence from $\mathcal{D}_{\mathrm{sc}}$
Generate six-field semantic descriptions for each cell type using an LLM
Encode each semantic field using frozen PubMedBERT
Apply field-wise normalization, centering, projection, and cross-field semantic modeling
Obtain common cell-type semantic embeddings $\{\mathbf{e}_c\}_{c=1}^{C}$
Stage 2: Training Statistics and Confusion-Pair Construction
Compute training-set abundance statistics $\{\mu_c, \bar{\sigma}_c\}_{c=1}^{C}$
Compute cell-type-specific enrichment scales $\{s_c\}_{c=1}^{C}$
Compute semantic similarities $S_{\mathrm{sem}}(i, j)$
Compute nonnegative training-abundance correlations $S_{\mathrm{corr}}(i, j)$
Compute semantic-confusion scores $S_{\mathrm{conf}}(i, j)$
Select the top-k candidates for each cell type and construct $\mathcal{P}_{\mathrm{conf}}$
Compute pair-confidence weights $\{w_{ij}^{\mathrm{pair}}\}_{(i,j)\in\mathcal{P}_{\mathrm{conf}}}$
Stage 3: Joint Model Training
for $t = 1$ to T **do**
    Compute $\lambda_{\mathrm{KGSC}}(t)$ and $\lambda_{\mathrm{RSGC}}(t)$ using linear warm-up
    for each sampled 2-hop spatial subgraph **do**
        Obtain sampled nodes, center nodes $\mathcal{V}_{\mathrm{center}}$, sampled edges $\mathcal{E}_{\mathrm{sub}}$, and labels $\mathbf{Y}_{\mathrm{sub}}$
        $(\mathbf{P}_{\mathrm{spot}}, \mathbf{P}_{\mathrm{local}}, \mathbf{P}_{\mathrm{global}}, \mathbf{P}_{\mathrm{fused}}) \leftarrow f_\theta(\mathcal{X}_{\mathrm{sub}}, \mathcal{E}_{\mathrm{sub}})$
        $\mathbf{P}_{\mathrm{final}} \leftarrow \frac{1}{4}\left(\mathbf{P}_{\mathrm{spot}} + \mathbf{P}_{\mathrm{local}} + \mathbf{P}_{\mathrm{global}} + \mathbf{P}_{\mathrm{fused}}\right)$
        $\mathcal{L}_{\mathrm{MSE}} \leftarrow \mathrm{MSE}(\mathbf{P}_{\mathrm{final}}, \mathbf{Y}_{\mathrm{sub}})$
        Extract center-node final predictions and corresponding labels
        Compute $\Delta y_{u,ij}$ and $\Delta p_{u,ij}$ for $(i, j) \in \mathcal{P}_{\mathrm{conf}}$
        Compute $r_{u,ij}$, $d_{u,ij}$, and $w_{u,ij}^{\mathrm{KGSC}}$
        Compute $\mathcal{L}_{\mathrm{KGSC}}$ from center-node final predictions
        $\mathcal{E}^* \leftarrow \{(u, v) \in \mathcal{E}_{\mathrm{sub}} \mid u \in \mathcal{V}_{\mathrm{center}},\ u \neq v\}$
        Compute $R_{u,c}$ and $Q_{uv,c}$ for $(u, v) \in \mathcal{E}^*$
        Compute $\Delta Y_{uv,c}$ from labels and $\Delta P_{uv,c}$ from $\mathbf{P}_{\mathrm{local}}$
        Compute $G_{uv,c}$ and $W_{uv,c} = Q_{uv,c} G_{uv,c}$
        Compute center-node-normalized $\mathcal{L}_{\mathrm{RSGC}}$
        $\mathcal{L}_{\mathrm{total}} \leftarrow \mathcal{L}_{\mathrm{MSE}} + \lambda_{\mathrm{KGSC}}(t)\mathcal{L}_{\mathrm{KGSC}} + \lambda_{\mathrm{RSGC}}(t)\mathcal{L}_{\mathrm{RSGC}}$
        Update θ by back-propagation
    end for
end for
return f_θ

A.4 Training and Implementation Details

Algorithm 1 summarizes the complete training procedure, including offline cell semantic knowledge construction, semantic-confusion pair construction, branch-specific supervision, and joint optimization with KGSC and RSGC.

Training configuration. For the main quantitative comparison in Table 1, the models were independently trained using three random seeds, 3407, 3050, and 3840, and the results are reported as mean and standard deviation across the three runs. Unless otherwise specified, the ablation and mechanism analyses were conducted using seed 3407. The spatial graph was sampled using 2-hop neighborhoods, with all available neighbors retained at each hop. The subgraph batch size was set to 16. The model was trained for 25 epochs using Adam with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-4}. A cosine annealing schedule was adopted with $T_{\max} = 20$ and a minimum learning rate of 1×10^{-5}.

The four prediction heads were averaged to obtain the final prediction used by the basic abundance-regression objective. In the sampled subgraph, the MSE loss was computed from the final prediction, KGSC was evaluated using the center-node final predictions, and RSGC was applied only to the local GAT prediction branch through the center-originated valid spatial edges defined above. This branch-decoupled supervision allows KGSC to correct semantic-confusion errors in the final abundance prediction while RSGC directly regularizes the branch responsible for local spatial propagation.

Auxiliary-loss warm-up. Both auxiliary constraints were gradually introduced during the first five training epochs. Let $t = 1, 2, \ldots, T$ denote the one-based training epoch and let $T_w = 5$. The epoch-dependent KGSC coefficient is

$$\lambda_{\mathrm{KGSC}}(t) = \lambda_{\mathrm{KGSC}} \min\left(1, \frac{t}{T_w}\right), \tag{54}$$

and the RSGC coefficient is

$$\lambda_{\mathrm{RSGC}}(t) = \lambda_{\mathrm{RSGC}} \min\left(1, \frac{t}{T_w}\right). \tag{55}$$

Using the same total-objective definition as in the main text, training minimizes

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{MSE}} + \lambda_{\mathrm{KGSC}}(t)\mathcal{L}_{\mathrm{KGSC}} + \lambda_{\mathrm{RSGC}}(t)\mathcal{L}_{\mathrm{RSGC}}. \tag{56}$$

KGSC configuration. For each cell type, the top $k = 3$ candidate cell types were selected according to the semantic-confusion score, and the deduplicated union of the resulting pairs was used as the confusion set. The contribution of the training-set abundance correlation was controlled by $\alpha = 0.30$. The KGSC loss weight was set to $\lambda_{\mathrm{KGSC}} = 0.01$. The target-difference margin and temperature were set to $m_d = 0.15$ and $T_d = 0.10$, respectively. The prediction near-tie temperature was set to $T_p = 0.25$, while the relative-order margin and temperature were set to $m_s = 0.05$ and $T_s = 0.25$, respectively. The SmoothL1 parameter was set to $\beta = 0.10$.

RSGC configuration. The RSGC loss weight was set to $\lambda_{\mathrm{RSGC}} = 0.005$. For enrichment-scale estimation, $n_{\min} = 8$ positive observations were required before using the median positive abundance; otherwise, the positive mean was used, while cell types without positive training observations used the global median of positive training abundances. The minimum enrichment scale was 10^{-4}. The boundary coefficient was set to $\beta_b = 0.75$. The gradient-reliability floor, margin, and temperature were set to $g_0 = 0.25$, $m_g = 0.15$, and $T_g = 0.10$, respectively. The SmoothL1 parameter was set to $\beta = 0.10$.

Cell semantic knowledge, the semantic-confusion pair set, training-set reliability statistics, KGSC, and RSGC are required only during training. At inference time, the trained model requires only the H&E image patches and their spatial adjacency relationships, so the proposed auxiliary learning strategy introduces no additional semantic-knowledge input at inference.

B Quantitative Analysis of RSGC

To further evaluate the effectiveness of RSGC in preserving local spatial abundance variations, we compared the RSGC-only model with the baseline model on the Human Lung dataset under the same experimental settings. Since RSGC directly constrains the local GAT prediction branch, the corresponding local predictions were used for evaluation. Gradient MAE, Direction Accuracy, and Gradient Spearman correlation were employed to assess local spatial gradient recovery.

As shown in Table 4, introducing RSGC consistently improves local spatial gradient recovery, with modest gains in Gradient MAE and substantial improvements in Direction Accuracy and Gradient Spearman correlation. These results indicate that RSGC improves the consistency between predicted and ground-truth local abundance variations, particularly in terms of gradient direction and relative variation patterns, further demonstrating its effectiveness in preserving reliable local spatial structures.

Table 4: Quantitative comparison of local spatial gradient recovery on the Human Lung dataset.

Method	**Grad. MAE** $\downarrow$	**Direction Acc.** $\uparrow$	**Gradient Spearman** $\rho \uparrow$
Hist2cell	0.7483 ± 0.0504	0.4767 ± 0.0045	-0.0423 ± 0.0100
+RSGC	$\mathbf{0.7460 \pm 0.0504}$	$\mathbf{0.5781 \pm 0.0039}$	$\mathbf{0.1572 \pm 0.0077}$

Table 5: Ablation of semantic knowledge construction in KGSC on the Human Lung dataset.

Method	PCC $\uparrow$	SCC $\uparrow$
Baseline	0.326	0.307
KGSC w/ Raw Evidence	0.331	0.340
KGSC w/ Structured Semantics	**0.337**	**0.352**

C Additional Analysis of KGSC

C.1 Ablation of Semantic Knowledge Construction

We compare the full KGSC with a variant that uses raw reference evidence directly, without constructing structured biological descriptions. In this variant, cell-type identity, positive and negative marker genes, and tissue context are encoded directly, while the remaining KGSC pipeline, including semantic similarity computation, abundance-correlation calibration, confusion-pair selection, and relative-abundance supervision, is kept unchanged. The results are reported in Table 5.

The raw-evidence variant improves PCC and SCC from 0.326 and 0.307 to 0.331 and 0.340, respectively. Using structured biological semantics further increases them to 0.337 and 0.352. Since the downstream KGSC components are identical in the two variants, the additional improvement is attributable to the structured semantic representation built from the reference evidence.

C.2 Analysis of Semantic-Confusion Pairs

To further analyze the cell-type relationships identified by KGSC, we examined the semantic-confusion pairs constructed on the Human Lung training set. KGSC identified 171 confusion pairs among 80 cell types. Each cell type participated in at least three pairs, with an average of 4.28 pairs per cell type, indicating that the selected pairs were not concentrated on only a small subset of cell types.

Across all selected pairs, the average semantic similarity and positive abundance correlation were 0.661 and 0.605, respectively. In addition, 83.6% of the selected pairs belonged to the same coarse cell lineage, compared with only 12.3% among all possible cell-type pairs.

Table 6 lists the top-10 semantic-confusion pairs ranked by the confusion score S_{conf}. The selected pairs are dominated by biologically related cell types, particularly closely related T-cell states and subtypes, including multiple CD8 effector-memory and tissue-resident memory populations as well as CD4 effector-memory and tissue-resident memory cells. Other highly ranked pairs include memory and naive B cells, AT1 and AT2 alveolar epithelial cells, capillary endothelial subtypes, and pulmonary and arterial smooth-muscle subtypes. These results further support the effectiveness of the semantic-confusion pair construction strategy in identifying cell-type relationships that are suitable for relative abundance supervision.

To further examine whether the selected semantic-confusion pairs are indeed more prone to prediction confusion, we compared their relative-abundance ordering errors with those of the remaining cell-type pairs under the baseline model. As shown in Table 7, the selected pairs exhibit a significantly higher ordering error than the other pairs (0.298 and 0.209, respectively, $p < 10^{-11}$), indicating that the pairs identified by KGSC are more susceptible to relative-abundance ordering confusion under the baseline model.We further evaluated the same selected pairs before and after introducing KGSC. The relative-difference PCC increases from 0.199 to 0.211 ($p = 0.0025$), while the relative-difference SCC increases from 0.232 to 0.257 ($p < 10^{-6}$). These results show that KGSC not only identifies cell-type pairs that are more prone to relative-abundance confusion, but also improves the consistency of their predicted relative-abundance differences through targeted pairwise supervision.

Table 6: Top-10 semantic-confusion pairs ranked by the confusion score S_{conf} on the Human Lung training set.

Cell type i	**Cell type** j	S_{sem}	S_{corr}	S_{conf}	w_{pair}
CD8_EM	CD8_TRM_EM	0.961	0.990	0.958	1.000
CD8_TRM_EM	CD8_TRM	0.965	0.962	0.954	0.992
CD8_EM	CD8_TRM	0.944	0.958	0.932	0.947
B_memory	B_naive	0.939	0.930	0.919	0.921
Muscle_smooth_pulmonary	Muscle_smooth_syst_arterial	0.922	0.977	0.915	0.913
CD4_EM_Effector	CD4_TRM	0.914	0.988	0.910	0.903
AT1	AT2	0.913	0.932	0.895	0.872
CD8_EM_EMRA	CD8_TRM_EM	0.902	0.965	0.892	0.867
Endothelia_vascular_Cap_a	Endothelia_vascular_Cap_g	0.892	0.993	0.890	0.863
CD4_TRM	CD8_TRM	0.901	0.949	0.887	0.857

Table 7: Quantitative validation of KGSC on the Human Lung dataset.

Pair Set	Model	Ordering Error ↓	Relative-diff. PCC ↑	Relative-diff. SCC ↑
Other pairs	Baseline	0.209	–	–
Selected pairs	Baseline	$0.298^{\dagger}$	0.199	0.232
Selected pairs	+KGSC	–	$\mathbf{0.211}^{*}$	$\mathbf{0.257}^{\ddagger}$

$^{\dagger}p < 10^{-11}$ for Selected and Other pairs under the baseline model; $^{*}p = 0.0025$ and $^{\ddagger}p < 10^{-6}$ for Baseline and +KGSC on the selected pairs.

D Additional Spatial Structure Analysis on the Human Lung Dataset

To further evaluate whether the proposed method preserves the intrinsic spatial organization of different cell types, we conducted additional spatial autocorrelation, inter-cell-type spatial association, and cell-proportion consistency analyses on the Human Lung dataset.

As shown in Figure 5A, we first examined the relationship between the ground-truth spatial autocorrelation of each cell type, measured by Moran's I, and its prediction accuracy. For each cell type, prediction accuracy is quantified by the PCC between the predicted and ground-truth abundances across spots. We then computed the Pearson correlation across cell types between Moran's I and the corresponding per-cell-type PCC values. The resulting correlations are 0.65 and 0.73 for the two representative test sections, respectively. On both sections, cell types with stronger ground-truth spatial autocorrelation generally tend to exhibit higher prediction accuracy, indicating a clear association between intrinsic cell-type spatial organization and abundance prediction performance.

We further evaluated whether different methods recover the spatial association patterns among cell types. Figure 5B summarizes the results across all held-out test sections. For each test section, we computed the bivariate Moran's R matrix from the ground-truth cell abundances and compared it with the corresponding matrix obtained from each method. Pearson correlation and cosine similarity are calculated between the predicted and ground-truth bivariate Moran's R values to quantify the agreement in inter-cell-type spatial association patterns. Our method achieves the highest average Pearson correlation and cosine similarity among the compared methods, indicating a closer recovery of the ground-truth inter-cell-type spatial association structure.

Finally, Figure 5C compares the cell-proportion consistency of Ours and Hist2Cell for four representative cell types on test section WSA_LngSP9258463. For each cell type, the inferred cell proportions are compared with the corresponding Cell2location reference proportions across spatial spots. Pearson correlation measures the agreement between the inferred and reference abundances, whereas spatial JSD quantifies the discrepancy between their normalized spatial abundance distributions. Across all four representative cell types, our method consistently achieves higher Pearson correlation and lower spatial JSD than Hist2Cell, indicating improved agreement with the reference cell abundances together with more faithful recovery of their spatial distributions.

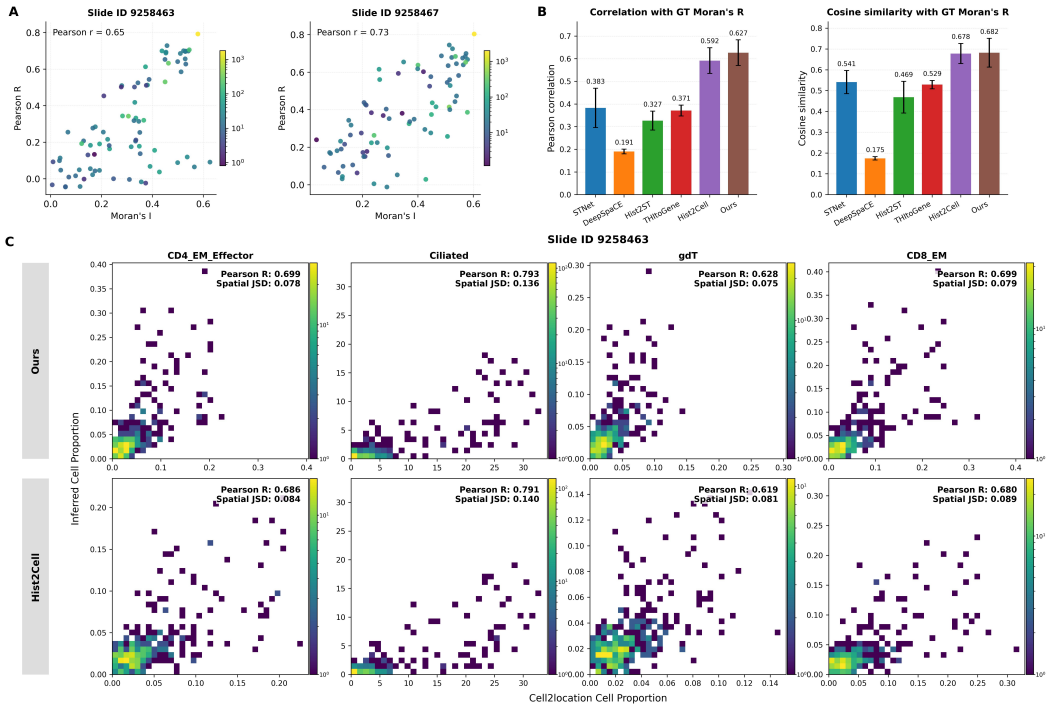

Figure 5: **Additional spatial structure analysis on the Human Lung dataset.** **(A)** Relationship between ground-truth Moran's I and cell-type prediction accuracy measured by Pearson correlation on two representative test sections. Each point represents one cell type, and the color indicates its total ground-truth abundance. **(B)** Recovery of inter-cell-type spatial association patterns by different methods across all held-out test sections. Pearson correlation and cosine similarity are computed between the predicted and ground-truth bivariate Moran's R values and averaged across test sections. Error bars denote the standard error across test sections. **(C)** Cell-proportion consistency comparison between Ours and Hist2Cell for four representative cell types on test section WSA_LngSP9258463. The inferred cell proportions are compared with the corresponding Cell2location reference proportions. Pearson correlation and spatial Jensen–Shannon divergence (JSD) are reported for each method and cell type.

E Cross-Dataset Spatial Structure Analysis on HER2ST and STNet

To further examine whether the spatial structure recovery ability of the proposed method generalizes beyond the Human Lung dataset, we conducted complementary analyses on the HER2ST and STNet datasets.

Figures 6A and 6B analyze the relationship between cell-pair abundance and inter-cell-type spatial association on HER2ST and STNet, respectively. For each cell-type pair, we used the log-transformed mean ground-truth cell-pair density and the corresponding average bivariate Moran's R. The distributions obtained from our predictions exhibit overall patterns similar to those observed in the ground truth on both datasets, indicating that the proposed method preserves the relationship between cell-pair abundance and spatial association.

Figures 6C and 6D further present representative spatial abundance maps on HER2ST and STNet. Three cell types are visualized for each dataset, with the ground truth, Hist2Cell prediction, and our prediction shown side by side. For all displayed examples, our method achieves higher PCC and SCC than Hist2Cell and produces spatial abundance patterns that are more consistent with the corresponding ground-truth distributions. These results further demonstrate the ability of the proposed method to preserve spatial abundance structures across different datasets.

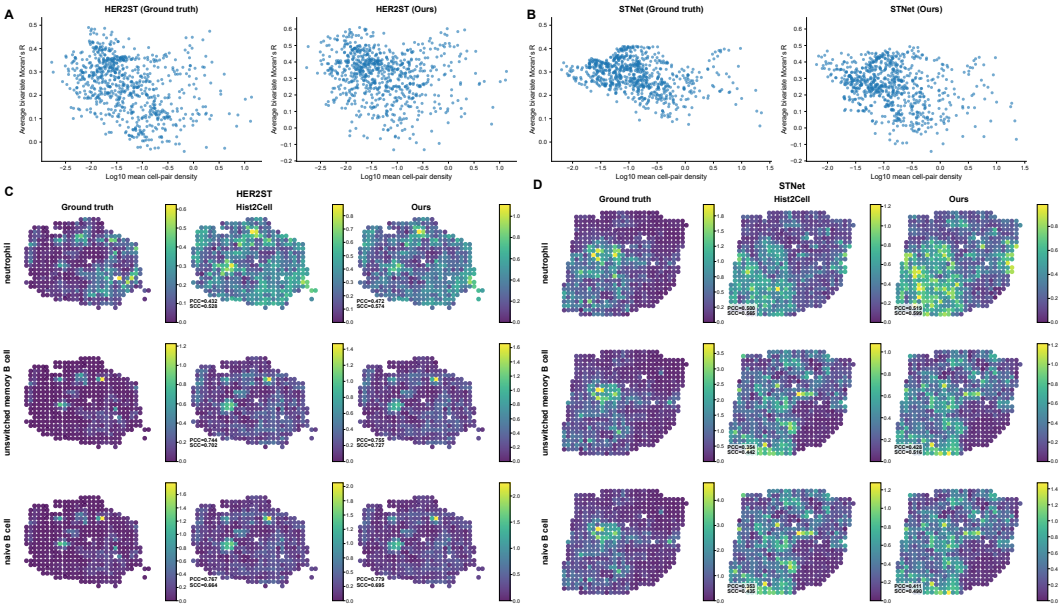

Figure 6: **Cross-dataset spatial structure analysis on HER2ST and STNet.** **(A)** Relationship between the log-transformed mean cell-pair density and average bivariate Moran's R on HER2ST for the ground truth and our predictions. **(B)** Corresponding analysis on STNet. Each point in (A–B) represents one cell-type pair. **(C)** Representative spatial abundance maps on HER2ST for neutrophils, unswitched memory B cells, and naive B cells. Ground truth, Hist2Cell predictions, and our predictions are shown from left to right. **(D)** Corresponding spatial abundance maps on STNet. PCC and SCC are reported for the predicted maps. Color scales are determined independently for each abundance map.

To further compare cell-type-level abundance consistency across different methods, Figures 7 and 8 present the predicted cell proportions of six methods against the corresponding Cell2location-based reference proportions for four representative cell types on HER2ST and STNet, respectively. On HER2ST, our method achieves the highest Pearson correlation for all four representative cell types and the lowest or tied-lowest Spatial JSD. On STNet, our method achieves both the highest Pearson correlation and the lowest Spatial JSD for all four representative cell types. These results further demonstrate that the proposed method can accurately recover cell-type-specific abundance patterns while preserving their spatial distributions across different datasets.

F Per-Cell-Type Performance Analysis Across Datasets

To provide supplementary results for individual cell types and further compare the fine-grained prediction performance of different methods, we conducted a cell-type-level evaluation on the Human Lung, HER2ST, and STNet datasets. For each cell type, all test spots across the corresponding test sections were pooled, and the PCC between the predicted and ground-truth abundances was computed for each method. As shown in Figures 9, 10, and 11, the proposed method maintains competitive per-cell-type PCC across diverse cell types on all three datasets.

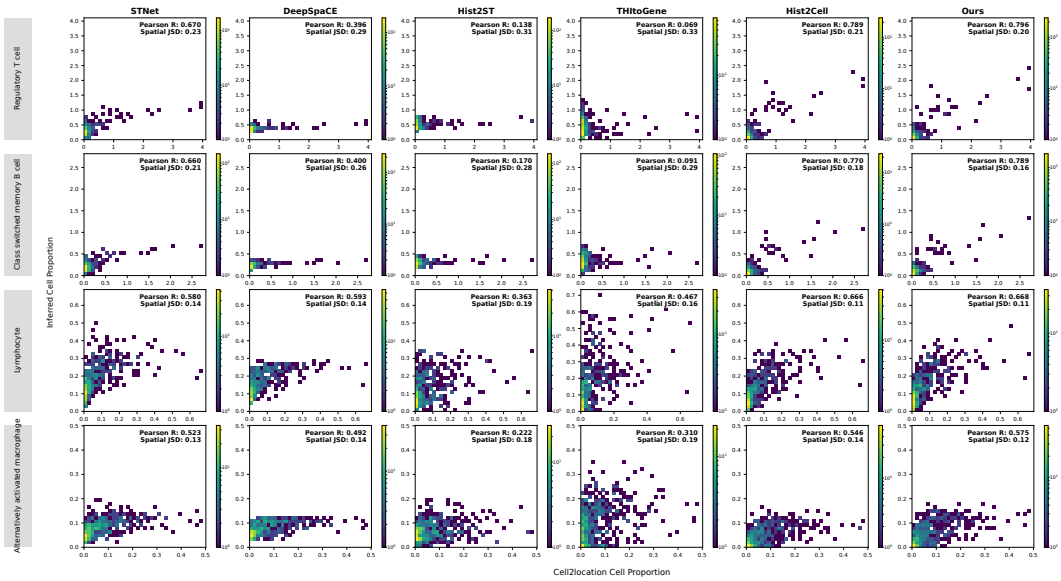

Figure 7: **Cell-type-level abundance consistency comparison on the HER2ST dataset.** Rows correspond to four representative cell types, and columns correspond to ST-Net, DeepSpaCE, Hist2ST, THItoGene, Hist2Cell, and Ours. Each panel compares the Cell2location-based reference cell proportions with the inferred cell proportions across spots. Pearson correlation and Spatial JSD are reported for each method and cell type.

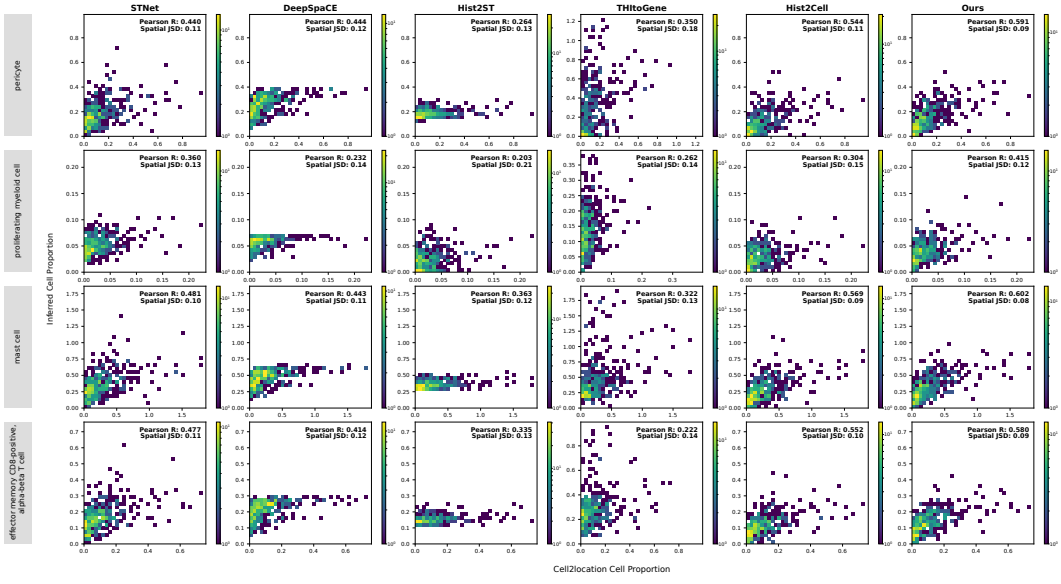

Figure 8: **Cell-type-level abundance consistency comparison on the STNet dataset.** Rows correspond to four representative cell types, and columns correspond to ST-Net, DeepSpaCE, Hist2ST, THItoGene, Hist2Cell, and Ours. Each panel compares the Cell2location-based reference cell proportions with the inferred cell proportions across spots. Pearson correlation and Spatial JSD are reported for each method and cell type.

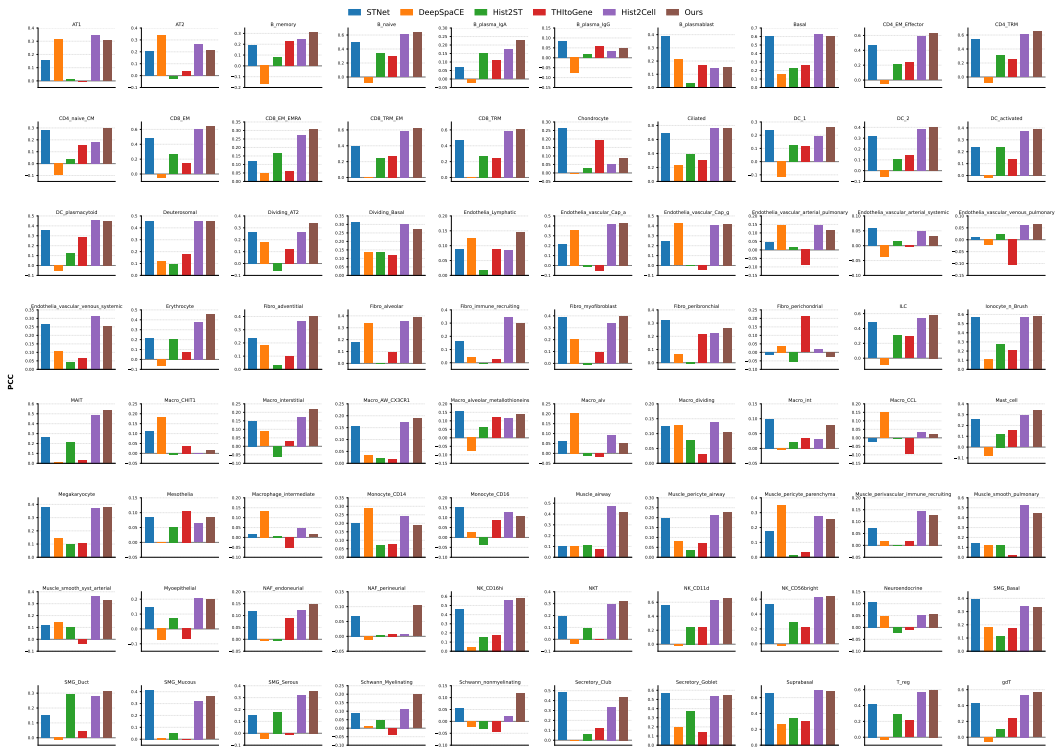

Figure 9: **Per-cell-type PCC comparison of all methods on the Human Lung dataset.** Bar plots representing the PCC values of ST-Net, DeepSpaCE, Hist2ST, THItoGene, Hist2Cell, and Ours for each cell type on the Human Lung dataset.

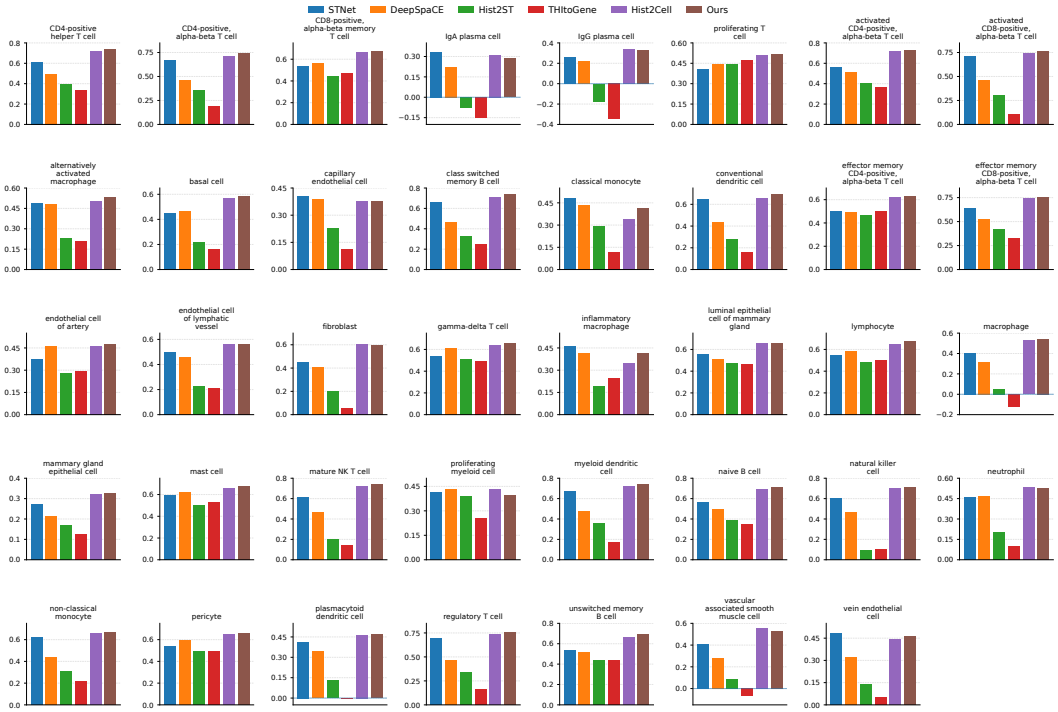

Figure 10: **Per-cell-type PCC comparison of all methods on the HER2ST dataset.** Bar plots representing the PCC values of ST-Net, DeepSpaCE, Hist2ST, THItoGene, Hist2Cell, and Ours for each cell type on the HER2ST dataset.

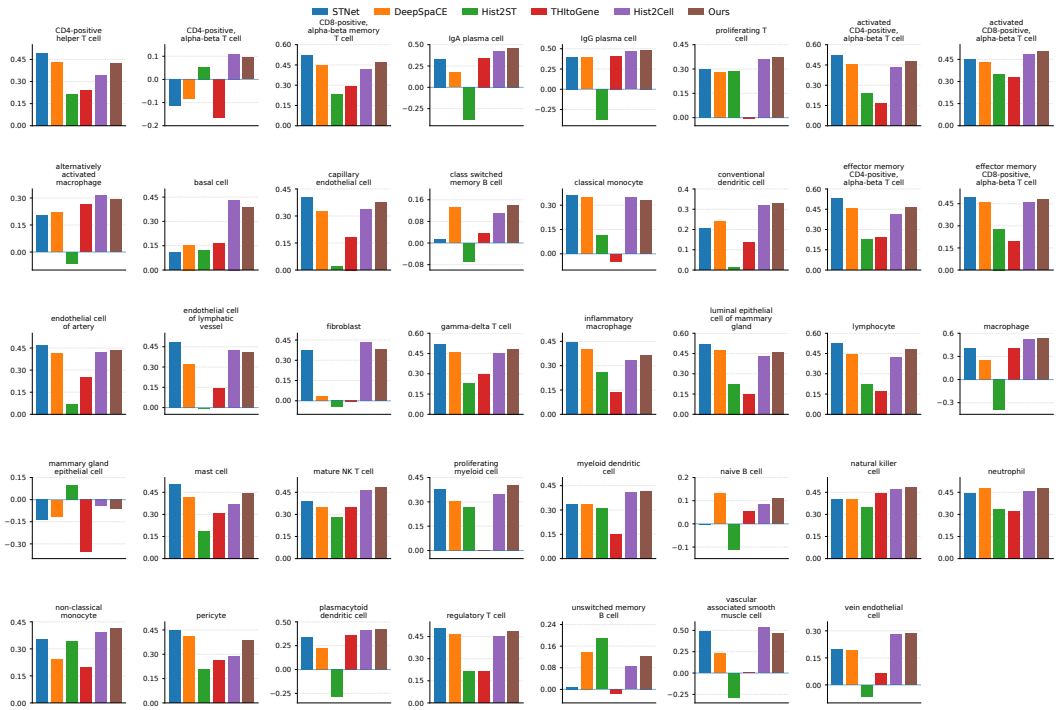

Figure 11: **Per-cell-type PCC comparison of all methods on the STNet dataset.** Bar plots representing the PCC values of ST-Net, DeepSpaCE, Hist2ST, THItoGene, Hist2Cell, and Ours for each cell type on the STNet dataset.