SPARSE: SEMANTIC-PRESERVING ADAPTER FOR ROBUST SPATIAL EMBEDDINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial transcriptomic representations should incorporate local tissue context without erasing expression-derived cell-type structure. We introduce **SPARSE**, a parameter-efficient hierarchical adapter for frozen transcriptomic embeddings. Within each centered spatial neighborhood, SPARSE builds a multi-resolution token hierarchy through feature-guided token merging with coordinate propagation, spatial rotary position encoding, cross-resolution attention, and top-down fusion. Masked expression reconstruction encourages retention of molecular information. With scGPT as the main backbone, SPARSE improves spatial-domain identification on two held-out, coordinate-defined human-brain regions and the held-out final stage of a mouse-embryo dataset. Across four transcriptomic backbones, it retains or improves cell-type separability on a held-out mouse-heart slice. On a held-out mouse whole-brain section, the same fixed embeddings support cell-type label transfer, neighborhood-composition prediction, and cell-wise tissue-region classification. Controlled comparisons show gains beyond parameter-matched adapters and graph-Laplacian regularization. With approximately 5M trainable adapter parameters, SPARSE complements dedicated spatial pretraining.

1 INTRODUCTION

Spatial transcriptomics extends single-cell RNA sequencing by preserving the physical locations of cells or capture spots within tissue sections (Marx, 2021; Moses & Pachter, 2022), enabling joint analysis of molecular states, cellular microenvironments, tissue regions, and spatial patterns (Liao et al., 2021). An effective embedding must capture expression-derived cell identity alongside local neighborhoods, regional boundaries, and multi-scale organization (Long et al., 2024). Yet adjacent cells can be molecularly distinct, while similar cells may occupy different anatomical contexts. The challenge is to integrate molecular and spatial information without losing cellular discrimination or tissue organization.

Existing approaches use spatial graphs (Hu et al., 2021; Dong & Zhang, 2022), neighborhood-aware learning (Zhao et al., 2021; Xu et al., 2024), contrastive objectives (Long et al., 2023), cross-dataset integration (Zhou et al., 2023), masked modeling (Tejada-Lapuerta et al., 2025), and large-scale spatial pretraining (Blampey et al., 2025). Despite this progress, dataset-specific methods typically learn representations anew for each target dataset rather than adapting a fixed transcriptomic representation, while dedicated spatial foundation models rely on large, harmonized corpora. Spatial profiling continues to scale (Hu et al., 2026), but datasets differ in platform, resolution, gene coverage, tissue quality, sequencing depth, and annotations, complicating both harmonization and transfer (Han et al., 2025; Zhuang, 2021; Bressan et al., 2023). Training or continually pretraining large spatial models also incurs substantial computational cost. Together, these approaches leave a gap: learning transferable spatial context on top of frozen transcriptomic embeddings without refitting representations for each target sample or pretraining on a large, harmonized spatial corpus.

A complementary route is to organize existing features across resolutions, as explored in computer vision. Feature pyramids construct multi-resolution representations (Lin et al., 2017); ViTDet constructs them from a pretrained single-scale backbone (Li et al., 2022); and TCFormer forms a feature-guided token hierarchy rather than relying on fixed-grid pooling (Zeng et al., 2024). Together, these works suggest that feature learning and scale construction can be designed separately.

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.

Pretrained transcriptomic models encode reusable information about cell identity, transcriptional state, and gene-expression programs (Yang et al., 2022; Theodoris et al., 2023; Cui et al., 2024; Hao et al., 2024; Rosen et al., 2024). For our goal, coordinate concatenation provides location and neighborhood smoothing provides local continuity, but neither explicitly organizes these embeddings across resolutions; smoothing may also blur molecularly distinct cells (Long et al., 2023; Dong & Zhang, 2022). Inspired by TCFormer's feature-guided token hierarchy, we introduce **SPARSE** (**S**emantic-**P**reserving **A**dapter for **R**obust **S**patial **E**mbeddings), a parameter-efficient spatial adapter for frozen transcriptomic embeddings (Chen et al., 2022). To adapt this hierarchy to spatial transcriptomics, SPARSE constructs it within each centered cell neighborhood, propagates coordinates through token merging, applies spatial RoPE in attention, and fuses multi-resolution context back to cell-level outputs. Here, scale denotes token granularity, not an expanding physical radius. Masked expression reconstruction encourages retention of molecular information. We define *semantic preservation* as retaining expression-derived, cell-type-discriminative structure, evaluated by cell-type separability on a held-out slice after adaptation on the remaining slices. Here, robustness refers to generalization to tissue samples or developmental stages excluded from adaptation.

Our main contributions are as follows:

- We formulate spatial transcriptomic representation learning as parameter-efficient adaptation of frozen expression-derived embeddings. With approximately 5M trainable adapter parameters, SPARSE learns the missing spatial context while retaining or improving mean held-out-slice cell-type accuracy and weighted F1 across four transcriptomic backbones.
- We introduce a hierarchical spatial adapter that combines centered spatial neighborhoods, feature-guided token merging with coordinate propagation, spatial positional encoding, cross-resolution attention, and top-down aggregation. The same fixed embeddings support cell-type label transfer, neighborhood-composition prediction, and cell-wise tissue-region classification on a held-out mouse whole-brain section.
- We evaluate transfer to held-out spatial regions and a developmental stage. On two coordinate-defined human-brain regions, SPARSE achieves NMI/ARI scores of $57.53\%/47.03\%$ and $38.92\%/24.50\%$, respectively. On the final mouse-embryo stage, held out from adaptation, it achieves $65.77\%/49.81\%$, outperforming all reported baselines on both metrics.

2 Related Work

Transcriptomic foundation models such as scBERT, Geneformer, scGPT, and scFoundation learn reusable expression-derived cell representations but do not explicitly model spatial organization (Yang et al., 2022; Theodoris et al., 2023; Cui et al., 2024; Hao et al., 2024). Dataset-specific spatial methods, including BayesSpace, SpaGCN, STAGATE, GraphST, and SCAN-IT, combine expression with spatial priors or neighborhood graphs, typically fitting representations for each target dataset (Zhao et al., 2021; Hu et al., 2021; Dong & Zhang, 2022; Long et al., 2023; Cang et al., 2021). Spatial foundation models such as Nicheformer, Novae, and scGPT-spatial instead learn from larger spatial corpora (Tejada-Lapuerta et al., 2025; Blampey et al., 2025; Wang et al., 2025).

SPARSE occupies a complementary position: it adapts a frozen transcriptomic representation using a parameter-efficient spatial module, rather than fitting a full representation model to each evaluation sample or continually pretraining the backbone. Its hierarchical neighborhood modeling also differs from coordinate concatenation and local graph-smoothing objectives, which do not by themselves model multiple token resolutions and may blur distinct neighboring cell states (Chen et al., 2022; Belkin et al., 2006; Cai et al., 2011).

3 Approach

3.1 Overview and Problem Formulation

Given a spatial transcriptomic sample with gene-expression matrix $X \in \mathbb{R}^{N\times G}$ and two-dimensional coordinates $P = [p_1, \ldots, p_N]^\top \in \mathbb{R}^{N\times 2}$, we first identify each target's $K = 199$ nearest cells by Euclidean coordinate distance:

$$d_{ij} = \|p_i - p_j\|_2, \qquad (j_{i,1}, \ldots, j_{i,K}) = \operatorname{argsort}^{(K)}_{j\neq i}\, d_{ij}. \tag{1}$$

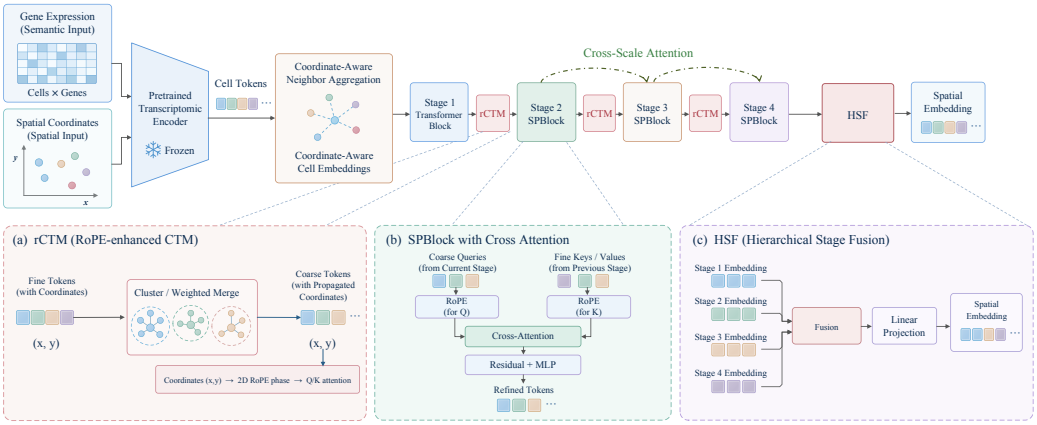

Figure 1: Overview of SPARSE. The upper pipeline shows the frozen transcriptomic encoder, coordinate-aware neighbor aggregation, four processing stages linked by rCTM, and HSF producing spatial embeddings. (a) In rCTM, fine tokens are clustered and merged with learned weights while their coordinates are propagated; the coordinates then provide 2D RoPE phases for attention queries and keys. (b) SPBlock applies RoPE to coarse queries and fine keys before cross-attention, residual updates, and MLP refinement. (c) HSF fuses Stage 1–4 features through top-down cross-attention and projects the refined output to spatial embeddings.

The superscript (K) denotes the first K indices in ascending distance order. This coordinate-based selection defines the local field for the spatial adapter; it does not feed a neighborhood into the transcriptomic backbone. Applied to expression alone, the frozen pretrained backbone F_θ produces

$$H^0 = F_\theta(X) = [h_1^0, \ldots, h_N^0]^\top \in \mathbb{R}^{N \times d_0}. \tag{2}$$

The embeddings H^0 are computed before spatial adaptation, and the backbone parameters θ remain fixed throughout training.

SPARSE processes each target cell or spot together with its selected spatial neighborhood, prepending the target to the ordered neighbor indices:

$$\mathcal{N}_i = (i, j_{i,1}, \ldots, j_{i,K}). \tag{3}$$

The target itself is excluded when selecting the K neighbors. With $M = K + 1$, the same ordered neighborhood is used to index both the transcriptomic embeddings and spatial coordinates:

$$\begin{aligned} E_i &= H^0[\mathcal{N}_i] \in \mathbb{R}^{M \times d_0}, \\ P_i &= P[\mathcal{N}_i] \in \mathbb{R}^{M \times 2}. \end{aligned} \tag{4}$$

Thus, each target is represented by a 200-token neighborhood.

3.2 Hierarchical Spatial Adapter

The local embeddings are first projected into the adapter token space and processed at the full token resolution:

$$T_i^1 = \mathrm{Stage}_1\left(\mathrm{Proj}(E_i), P_i\right), \qquad P_i^1 = P_i. \tag{5}$$

Here, Stage_1 applies spatially aware self-attention to the projected tokens without changing their token resolution.

SPARSE contains four hierarchical stages with feature dimensions $(64, 128, 320, 512)$. rCTM reduces the number of tokens by a factor of 0.5 between adjacent stages. For the default 200-token neighborhood, the four stages contain $(200, 100, 50, 25)$ tokens, respectively. The main configuration uses one attention block at each stage. At stages 2–4, rCTM forms coarser tokens before SPBlock attends to the preceding resolution.

3.3 TOKEN MERGING (RCTM)

RoPE-enhanced CTM (rCTM) constructs the token hierarchy between adjacent stages. The spatial KNN neighborhood first restricts the input to a local spatial field. Within this field, rCTM uses token features rather than coordinate distance to determine which cells or spots should be grouped. This avoids treating all spatially adjacent but molecularly distinct cells as equivalent.

For clarity, we omit the neighborhood index i in the following module-level definitions. Let $T^s = \{t_j^s\}_{j=1}^{M_s}$ and $P^s = \{p_j^s\}_{j=1}^{M_s}$ denote the token features and coordinates at stage s. rCTM first projects each token to the feature dimension of the next stage:

$$\bar{t}_j^s = \mathrm{LN}\left(t_j^s W_c^s\right). \tag{6}$$

It then predicts a scalar importance score and its corresponding positive weight:

$$r_j^s = (w_s)^\top \bar{t}_j^s, \qquad \omega_j^s = \exp(r_j^s). \tag{7}$$

DPC-KNN clustering (Du et al., 2016) is performed in the projected token-feature space. The feature distance between tokens j and ℓ is

$$D_{j\ell}^s = \frac{\left\|\bar{t}_j^s - \bar{t}_\ell^s\right\|_2}{\sqrt{d_{s+1}}}. \tag{8}$$

The learned scores r_j^s are incorporated into the density estimates used by DPC-KNN. After selecting M_{s+1} representative centers, each fine-scale token is assigned to its nearest center in feature space, producing clusters $\{C_m^s\}_{m=1}^{M_{s+1}}$. Center selection and cluster assignment are non-differentiable routing decisions computed without gradient tracking. The score predictor still receives gradients through the weighted-merge and attention-bias paths.

Within each cluster, the normalized token importance is

$$\alpha_j^s = \frac{\omega_j^s}{\sum_{\ell \in C_m^s} \omega_\ell^s}, \qquad j \in C_m^s. \tag{9}$$

The coarse feature and its coordinate are then computed using the same aggregation weights:

$$\begin{aligned} \widetilde{t}_m^{s+1} &= \sum_{j \in C_m^s} \alpha_j^s \bar{t}_j^s, \\ p_m^{s+1} &= \sum_{j \in C_m^s} \alpha_j^s p_j^s. \end{aligned} \tag{10}$$

This produces the coarse token set $\widetilde{T}^{s+1} = \{\widetilde{t}_m^{s+1}\}_{m=1}^{M_{s+1}}$ and its associated coordinates $P^{s+1} = \{p_m^{s+1}\}_{m=1}^{M_{s+1}}$. The projected fine-scale tokens $\bar{T}^s = \{\bar{t}_j^s\}_{j=1}^{M_s}$ are retained for the following cross-scale attention.

3.4 SPATIALLY AWARE CROSS-SCALE ATTENTION

SPBlock connects the coarse tokens produced by rCTM with the preceding fine-scale representations. At stage $s+1$, the coarse tokens provide the queries, whereas the projected fine-scale tokens provide the keys and values:

$$\begin{aligned} Q^{s+1} &= \mathrm{LN}\Big(\widetilde{T}^{s+1}\Big) W_q^{s+1}, \\ K^s &= \mathrm{LN}\big(\bar{T}^s\big) W_k^{s+1}, \\ V^s &= \mathrm{LN}\big(\bar{T}^s\big) W_v^{s+1}. \end{aligned} \tag{11}$$

To encode spatial coordinates and introduce spatial relationships into attention, SPARSE applies two-dimensional spatial RoPE (Su et al., 2024). Coordinates are normalized to $[-1, 1]$ using the bounding box of the original neighborhood, and the same reference frame is retained across all stages. Spatial RoPE rotates the queries and keys according to their corresponding coordinates:

$$\begin{aligned} \widetilde{Q}^{s+1} &= \mathrm{RoPE}\left(Q^{s+1}, P^{s+1}\right), \\ \widetilde{K}^s &= \mathrm{RoPE}\left(K^s, P^s\right). \end{aligned} \tag{12}$$

The cross-scale attention weights are

$$A^{s+1} = \mathrm{softmax}\left(\frac{\widetilde{Q}^{s+1}(\widetilde{K}^{s})^{\top}}{\sqrt{d_h}} + \mathbf{1}(r^{s})^{\top}\right), \tag{13}$$

where d_h is the attention-head dimension. The term $\mathbf{1}(r^s)^\top$ broadcasts the learned fine-token scores across all coarse queries and acts as a key-confidence bias. The cross-scale output is

$$O^{s+1} = \left(A^{s+1}V^{s}\right)W_o^{s+1}. \tag{14}$$

The attention output updates the coarse tokens through a residual connection, followed by an MLP and stage-level normalization. Omitting dropout for clarity, the output of stage $s+1$ is

$$\begin{aligned} Y^{s+1} &= \widetilde{T}^{s+1} + O^{s+1}, \\ T^{s+1} &= \mathrm{LN}\left(Y^{s+1} + \mathrm{MLP}\left(\mathrm{LN}(Y^{s+1})\right)\right). \end{aligned} \tag{15}$$

Stage 1 uses the same spatially rotated attention principle in a self-attention setting, without cross-scale keys or token-score bias.

3.5 Hierarchical Stage Fusion (HSF)

For target i, the hierarchical encoder produces four stage outputs $\{T_i^1, T_i^2, T_i^3, T_i^4\}$ at progressively coarser resolutions. HSF first maps each stage output to a common 64-dimensional feature space:

$$L_i^s = \mathrm{Lateral}_s(T_i^s), \qquad s = 1, 2, 3, 4. \tag{16}$$

Top-down aggregation starts from the coarsest representation:

$$F_i^4 = L_i^4. \tag{17}$$

The coarse context is then recursively propagated to finer token resolutions:

$$F_i^s = \mathrm{Merge}_s\left(L_i^s, F_i^{s+1}\right), \qquad s = 3, 2, 1. \tag{18}$$

In each merge operation, the finer representation L_i^s provides the queries, while the coarser fused representation F_i^{s+1} provides the keys and values. The output therefore has the token resolution of the finer stage. This transfers coarser local-tissue context back to individual tokens without explicit interpolation, cluster-assignment-based upsampling, or concatenation of all stage-wise features. The fusion blocks use cross-attention without token-score bias or additional spatial RoPE, followed by residual MLP updates; Refine is a full-resolution self-attention block.

A final self-attention block refines the full-resolution representation:

$$Z_i^{\mathrm{loc}} = \mathrm{OutProj}\left(\mathrm{Refine}(F_i^1)\right). \tag{19}$$

Because the target cell occupies the first token position, its final spatial embedding is

$$z_i = Z_i^{\mathrm{loc}}[1, :]. \tag{20}$$

Applying this procedure to every target yields $Z = [z_1, \ldots, z_N]^\top \in \mathbb{R}^{N\times 64}$.

3.6 Training Objective and Parameter Efficiency

SPARSE is trained using masked expression reconstruction. For every centered neighborhood, 20% of the input tokens are randomly selected and their pretrained embeddings are replaced with a learned mask token. Their coordinates and the embeddings of the remaining unmasked cells remain available to the adapter. Reconstructing a masked cell therefore encourages the model to exploit both its spatial position and the molecular context provided by its surrounding cells.

For a mini-batch containing B centered neighborhoods, let Z_b^{loc} denote the full-resolution adapted representation of neighborhood b. A training-only decoder g_ψ reconstructs the expression profiles of its neighborhood tokens:

$$\widehat{X}_b = g_\psi\left(Z_b^{\mathrm{loc}}\right), \qquad b = 1, \ldots, B. \tag{21}$$

The reconstruction loss is evaluated only at masked token positions:

$$\mathcal{L}_{\mathrm{MCM}} = \frac{1}{B|\mathcal{M}|} \sum_{b=1}^{B} \sum_{j \in \mathcal{M}_b} \left[1 - \cos\left(\widehat{x}_{b,j}, x_{b,j}\right)\right]. \tag{22}$$

Here, B is the batch size, $\mathcal{M}_b$ is the masked-token set in neighborhood b, and $|\mathcal{M}_b| = |\mathcal{M}|$ is the fixed number of masked tokens per neighborhood. The vectors $x_{b,j}$ and $\widehat{x}_{b,j}$ denote the observed and reconstructed expression profiles of token j, respectively. The masked reconstruction loss is the sole training objective used in the main experiments.

The frozen transcriptomic embeddings provide a fixed expression-derived foundation, while reconstruction in expression space encourages the adapted representations to remain informative of cellular expression programs. The reconstruction objective does not impose cell-type labels; instead, semantic preservation is evaluated empirically by measuring whether cell-type-discriminative information is retained after spatial adaptation.

Only the hierarchical adapter, learned mask token, and reconstruction decoder are optimized. The transcriptomic backbone is excluded from the adaptation computation graph, and the decoder is discarded after training. The spatial adapter contains approximately 5M trainable parameters across the evaluated configurations, excluding the reconstruction head.

4 Experiments

4.1 Setup

Quantitative evaluations use mouse heart, human brain, mouse embryo, and mouse whole brain; human lung is analyzed qualitatively in Appendix E.1. SPARSE is adapted only on designated training samples: no held-out cells or annotations enter adaptation. Held-out labels are used only to compute spatial-domain metrics or in the explicitly defined downstream reference and probe splits. Results are mean $\pm$ sample standard deviation over seeds 0, 42, and 123, except the seed-42 brain ablation. The brain holdouts are coordinate-defined regions, not confirmed independent physical sections (Appendix B.3). Every trainable adapter uses its prespecified epoch-30 checkpoint, paired with the same downstream seed; no checkpoint is selected by NMI, ARI, accuracy, weighted F1, or test labels. Appendix A, Appendix B, Appendix C, and Appendix D specify splits, metrics, baselines, and downstream protocols.

4.2 Baselines

We compare PCA; specialized spatial methods SpaGCN and SCAN-IT; spatial foundation models Nicheformer, scGPT-spatial, and Novae; and parameter-matched scGPT adapters Linear, Transformer, and Laplacian Reg. Each frozen backbone is its own unadapted semantic-retention reference. PCA fits an expression-only transformation on each evaluation sample, whereas SpaGCN and SCAN-IT fit spatial models using its unlabeled expression and coordinates. The remaining methods do not fit task-specific representations on evaluation samples. Thus "No Adapt" in Tables 1 and 4 excludes evaluation-region, -section, or -stage adaptation, but not downstream probe fitting or possible foundation-model pretraining exposure. Held-out annotations never fit representation parameters. The parameter-matched controls test the complete SPARSE configuration under matched masking and reconstruction settings, while the ablation below examines individual modules. Appendix D.4 details these objectives, probes, and Laplacian Reg.

4.3 Held-out Region and Stage Transfer

We evaluate two coordinate-defined human-brain regions excluded from adaptation. Separately, we evaluate the final mouse-embryo stage after adaptation on earlier stages. For every method, Leiden clustering uses the same embedding-space Euclidean k-nearest-neighbor graph ($k = 15$) and resolution 0.8 (Traag et al., 2019). Held-out annotations are used only to compute adjusted Rand index (ARI) and normalized mutual information (NMI) (Yuan et al., 2024). Appendix E.2 reports results across resolutions 0.1–1.0.

Table 1: Spatial-domain NMI/ARI at Leiden resolution 0.8. No Adapt means no task-specific representation fitting on the evaluation region or stage. Underlines mark the best non-SPARSE means; gains are in percentage points. Colors do not imply significance.

Baseline		Held-out Spatial Evaluation				Held-out Stage Transfer	
		Human Brain Region 1		Human Brain Region 2		E16.5 Stage Evaluation	
Methods	**No Adapt**	**NMI % ↑**	**ARI % ↑**	**NMI % ↑**	**ARI % ↑**	**NMI % ↑**	**ARI % ↑**
Non-spatial							
PCA	✗	27.65 ± 1.37	26.20 ± 2.37	13.59 ± 1.27	8.470 ± 0.95	58.02 ± 0.34	41.29 ± 0.36
Spatial Methods							
SpaGCN	✗	24.07 ± 0.12	7.370 ± 0.31	9.110 ± 0.06	1.830 ± 0.31	59.74 ± 3.36	45.37 ± 3.85
SCAN-IT	✗	53.03 ± 0.56	28.39 ± 0.71	37.15 ± 1.77	16.26 ± 1.47	64.64 ± 0.81	42.40 ± 2.74
Spatial Foundation Models							
Novae	✓	48.23 ± 2.19	36.55 ± 1.87	27.49 ± 1.22	15.09 ± 1.07	59.81 ± 0.73	38.47 ± 2.74
Nicheformer	✓	42.88 ± 0.23	39.48 ± 0.29	26.66 ± 0.25	19.48 ± 0.33	14.82 ± 0.02	3.840 ± 0.12
scGPT-spatial	✓	36.98 ± 0.85	25.65 ± 1.42	14.16 ± 0.52	8.430 ± 0.08	50.97 ± 0.93	34.73 ± 0.95
Lightweight Adapters							
scGPT + Linear	✓	37.10 ± 0.20	3.65 ± 0.24	32.48 ± 0.10	3.42 ± 0.13	52.95 ± 1.47	21.25 ± 3.16
+ Transformer	✓	41.60 ± 3.01	16.29 ± 2.25	28.60 ± 3.70	6.12 ± 1.40	19.82 ± 3.65	6.93 ± 2.19
+ Laplacian Reg	✓	37.85 ± 4.05	5.81 ± 0.42	31.44 ± 0.49	3.33 ± 0.22	41.04 ± 1.08	7.79 ± 0.70
Ours							
scGPT + SPARSE	✓	**57.53 ± 0.58**	**47.03 ± 0.66**	**38.92 ± 3.08**	**24.50 ± 4.62**	**65.77 ± 0.97**	**49.81 ± 1.19**
SPARSE gain vs. best other		**+4.50**	**+7.55**	**+1.77**	**+5.02**	**+1.13**	**+4.44**

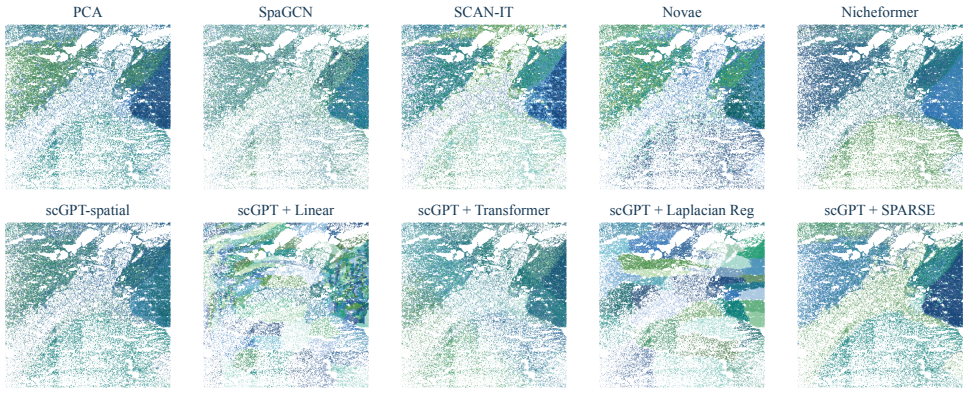

Figure 2: Spatial-domain maps for held-out human-brain region 1 (seed 42). All panels use the same cell coordinates and Leiden settings ($k = 15$, resolution 0.8). Community colors are assigned independently and do not indicate matched anatomical labels.

SPARSE ranks first in NMI and ARI on all three evaluations (Table 1): $57.53\%/47.03\%$ and $38.92\%/24.50\%$ on human-brain regions 1 and 2, and $65.77\%/49.81\%$ on the held-out mouse-embryo stage. Figure 2 shows region 1 assignments for seed 42. Under matched masking and reconstruction settings (Appendix D.4), SPARSE also exceeds the Linear, Transformer, and Laplacian Reg controls in mean NMI and ARI. This comparison supports the complete configuration; the ablation below examines components separately.

4.4 Ablation Study

The ablation uses seed 42, both held-out human-brain regions, and Leiden resolution 0.8; variants share the frozen backbone, objective, and evaluation protocol. Removing spatial-neighborhood construction has the largest effect: mean NMI/ARI drops from $48.16\%/35.65\%$ to $37.46\%/25.52\%$ (Table 2). Removing rCTM, token scoring, cross-scale attention, or HSF also lowers both metrics. Per-region results are in Appendix E.3.

Table 2: Human-brain component ablation.

	Human Brain Avg	
Model variant	**Mean NMI % ↑**	**Mean ARI % ↑**
w/o spatial neighborhood	37.46	25.52
w/o HSF	43.04	34.20
w/o rCTM	44.12	32.18
w/o cross-scale attention	44.66	32.81
w/o token score	43.31	31.79
Full scGPT+SPARSE	**48.16**	**35.65**

Among the internal removals, mean NMI falls by 3.50–5.12 points and mean ARI by 1.45–3.86 points relative to the full model. HSF removal has the largest NMI reduction, whereas token-score removal has the largest ARI reduction. Since the ablation uses one seed, these differences describe that run rather than a distribution of training outcomes.

4.5 Semantic Retention

SPARSE is adapted on 21 mouse-heart slices and tested on the held-out slice. Within it, a stratified 50% reference set supplies labels for distance-weighted 15-nearest-neighbor transfer to the other 50%. Frozen and adapted embeddings use the same seed-specific reference/query split. Feature standardization is fitted on the reference half only and then applied to the queries. Labels do not fit the representation; only downstream transfer is within-slice (Appendix C.1).

Table 3: Cell-type separability on the held-out mouse-heart slice.

Model		**Accuracy % ↑**	**F1 Weighted % ↑**
scGPT	*Frozen*	83.00 ± 0.05	77.63 ± 0.14
	+SPARSE	$\mathbf{83.91 \pm 0.20}$	$\mathbf{78.79 \pm 0.61}$
scFoundation	*Frozen*	84.28 ± 0.16	79.11 ± 0.15
	+SPARSE	$\mathbf{84.53 \pm 0.05}$	$\mathbf{79.14 \pm 0.17}$
scBERT	*Frozen*	65.73 ± 0.21	55.80 ± 0.16
	+SPARSE	$\mathbf{75.40 \pm 2.22}$	$\mathbf{69.43 \pm 2.69}$
Geneformer	*Frozen*	83.31 ± 0.15	77.58 ± 0.24
	+SPARSE	$\mathbf{84.76 \pm 0.30}$	$\mathbf{79.85 \pm 0.59}$

Table 3 gives mean ± sample standard deviation over seeds 0, 42, and 123; "Frozen" is the unadapted backbone. SPARSE retains or improves both means for all four backbones. The gain ranges from 0.25/0.03 points for scFoundation to 9.67/13.63 points for scBERT (accuracy/weighted F1). The adapted scBERT standard deviations (2.22/2.69 points) also exceed those of its frozen reference, and several three-seed confidence intervals overlap (Appendix E.5). These aggregate scores support retained mean cell-type separability, not a universal gain or proof that rare and boundary-cell distinctions are unchanged. Because reference and query cells belong to the same test slice, this experiment does not establish cross-slice annotation transfer.

4.6 Multi-Scale Evaluation

One mouse whole-brain section is excluded from adaptation. Its fixed embeddings support within-section cell-type label transfer (50%/50% reference/query), neighborhood-composition prediction (80%/20% probe split), and tissue-region classification (50%/50% split). Probes use labels only afterward; see Appendix C.

Table 4: Mouse whole-brain multi-scale results. No Adapt means no test-section representation fit (downstream probes excluded). Underlines mark best non-SPARSE means. Gains are percentage points, with MAE sign reversed; colors do not imply significance.

Baseline		Cross-slice Evaluation on Mouse Whole Brain				
		Cell-level		Niche-level	Tissue-level	
Methods	**No Adapt**	**Accuracy % ↑**	**F1 Weighted % ↑**	**Mean MAE % ↓**	**Accuracy % ↑**	**F1 Weighted % ↑**
Non-spatial						
PCA	✗	20.99 ± 0.55	15.34 ± 0.81	3.777 ± 0.110	80.05 ± 0.36	79.41 ± 0.36
Spatial Foundation Models						
Novae	✓	49.23 ± 0.32	43.38 ± 0.26	2.628 ± 0.036	84.35 ± 0.26	83.30 ± 0.21
Nicheformer	✓	56.72 ± 1.16	52.12 ± 0.42	2.952 ± 0.051	43.19 ± 0.09	38.99 ± 0.65
scGPT-spatial	✓	87.69 ± 0.13	87.03 ± 0.14	4.057 ± 0.060	72.22 ± 0.22	70.58 ± 0.26
Lightweight Adapters						
scGPT + Linear	✓	50.26 ± 1.29	44.01 ± 1.30	2.952 ± 0.051	84.86 ± 0.95	82.45 ± 1.61
+ Transformer	✓	65.53 ± 4.70	61.58 ± 4.87	2.941 ± 0.068	79.00 ± 7.05	76.47 ± 8.04
+ Laplacian Reg	✓	41.79 ± 0.29	36.43 ± 0.37	2.952 ± 0.051	83.82 ± 0.60	80.16 ± 1.17
Ours						
scGPT + SPARSE	✓	**87.68 ± 1.34**	**87.09 ± 1.42**	**2.272 ± 0.090**	**88.12 ± 0.36**	**87.35 ± 0.40**
SPARSE gain vs. best other		**-0.01**	**+0.06**	**+0.356**	**+3.26**	**+4.05**

SPARSE yields the highest cell-type weighted F1 (87.09%), near-best accuracy (within 0.01 points), lowest niche mean MAE (2.272% across 100/200/500 cells), and highest tissue-region accuracy/F1 (88.12%/87.35%; Table 4). Figure 3 shows scale and probe-learning-rate sensitivity; full statistics are in Appendix E.4. One fixed embedding thus supports several downstream granularities. Although representation adaptation excludes the test section, these supervised probes are fitted and evaluated within it; they do not establish probe transfer to another unseen section.

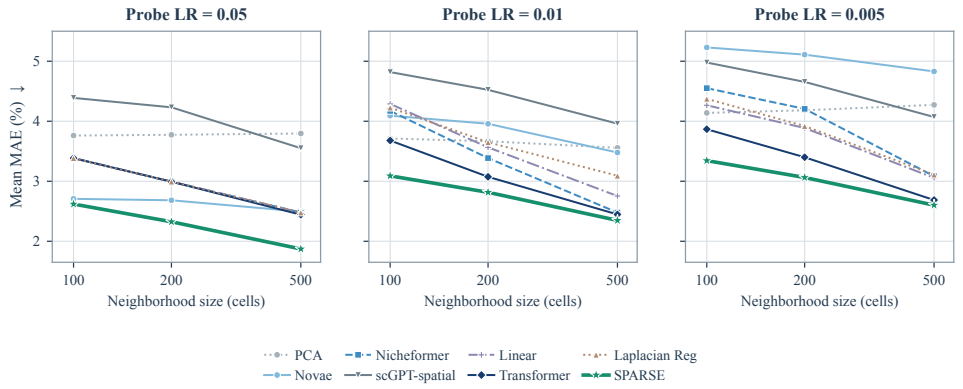

Figure 3: Mean niche-composition MAE for all eight methods on held-out mouse whole brain (three seeds; lower is better). Targets use 100, 200, or 500 cells; probe LR 0.05 is primary, with 0.01 and 0.005 for sensitivity. Full statistics are in Appendix E.4.

5 Conclusion

SPARSE adds spatial hierarchy to frozen transcriptomic embeddings with about 5M trainable parameters. It improves held-out domain identification, retains mean cell-type separability, and supports multi-granularity readouts. Its stages change token granularity within one local field, not physical radius or certified biological scale. Each dataset–backbone pair still requires separate adaptation; transfer of trained weights across datasets and validation on more independent specimens remain open. Future work should assess neighborhood-size sensitivity and computational scaling.

AI USE STATEMENT

Generative AI tools assisted with manuscript drafting, editing, translation, and LaTeX formatting. The experimental design and scientific decisions were made by the authors. The authors reviewed the AI-assisted material and take responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

The experimental protocol, data splits, preprocessing, model configurations, randomness controls, downstream evaluations, and additional results are given in Appendices A–E. The associated code is supplied as supplementary material.

REFERENCES

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

Quentin Blampey, Hakim Benkirane, Nadege Bercovici, Kevin Mulder, Gregoire Gessain, Florent Ginhoux, Fabrice Andre, and Paul-Henry Cournede. Novae: A graph-based foundation model for spatial transcriptomics data. *Nature Methods*, 22(12):2539–2550, 2025.

Dario Bressan, Giorgia Battistoni, and Gregory J. Hannon. The dawn of spatial omics. *Science*, 381 (6657):eabq4964, 2023.

Deng Cai, Xiaofei He, Jiawei Han, and Thomas S. Huang. Graph regularized nonnegative matrix factorization for data representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1548–1560, 2011.

Zixuan Cang, Xinyi Ning, Annika Nie, Min Xu, and Jing Zhang. SCAN-IT: Domain Segmentation of Spatial Transcriptomics Images by Graph Neural Network. In *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 320, 2021. doi: 10.5244/C.35.320.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. In *Advances in Neural Information Processing Systems*, volume 35, pp. 16664–16678, 2022.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: Toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21(8):1470–1480, 2024.

Kangning Dong and Shihua Zhang. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nature Communications*, 13(1):1739, 2022.

Mingjing Du, Shifei Ding, and Hongjie Jia. Study on density peaks clustering based on k-nearest neighbors and principal component analysis. *Knowledge-Based Systems*, 99:135–145, 2016.

Chuangyi Han, Senlin Lin, Zhikang Wang, Yan Cui, Qi Zou, and Zhiyuan Yuan. Reusability report: Exploring the transferability of self-supervised learning models from single-cell to spatial transcriptomics. *Nature Machine Intelligence*, 7(9):1414–1428, 2025.

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 21(8):1481–1491, 2024.

Chenlei Hu, Mehdi Borji, Giovanni J. Marrero, Vipin Kumar, Jackson A. Weir, Sachin V. Kammula, Evan Z. Macosko, and Fei Chen. Scalable spatial transcriptomics through computational array reconstruction. *Nature Biotechnology*, 44(2):215–221, 2026.

Jian Hu, Xiangjie Li, Kyle Coleman, Amelia Schroeder, Nan Ma, David J. Irwin, Edward B. Lee, Russell T. Shinohara, and Mingyao Li. Spagcn: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature Methods*, 18(11):1342–1351, 2021.

Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *Computer Vision – ECCV 2022*, pp. 280–296, 2022. doi: 10.1007/978-3-031-20077-9_17.

Jie Liao, Xiaoyan Lu, Xin Shao, Ling Zhu, and Xiaohui Fan. Uncovering an organ's molecular architecture at single-cell resolution by spatially resolved transcriptomics. *Trends in Biotechnology*, 39(1):43–58, 2021.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. doi: 10.1109/CVPR.2017.106.

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

Yahui Long, Kok Siong Ang, Raman Sethi, Sha Liao, Yang Heng, Lynn van Olst, Shuchen Ye, Chengwei Zhong, Hang Xu, Di Zhang, et al. Deciphering spatial domains from spatial multiomics with spatialglue. *Nature Methods*, 21(9):1658–1667, 2024.

Vivien Marx. Method of the year: Spatially resolved transcriptomics. *Nature Methods*, 18(1):9–14, 2021.

Lambda Moses and Lior Pachter. Museum of spatial transcriptomics. *Nature Methods*, 19(5):534–546, 2022.

Yanay Rosen, Maria Brbic, Yusuf Roohani, Kyle Swanson, Ziang Li, and Jure Leskovec. Toward universal cell embeddings: Integrating single-cell rna-seq datasets across species with saturn. *Nature Methods*, 21(8):1492–1500, 2024.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Alejandro Tejada-Lapuerta, Anna C. Schaar, Robert Gutgesell, Giovanni Palla, Lennard Halle, Mariia Minaeva, Larsen Vornholz, Leander Dony, Francesca Drummer, Till Richter, et al. Nicheformer: A foundation model for single-cell and spatial omics. *Nature Methods*, 22(12): 2525–2538, 2025. doi: 10.1038/s41592-025-02814-z.

Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, Zexian Zeng, X. Shirley Liu, et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023.

Vincent A. Traag, Ludo Waltman, and Nees Jan van Eck. From louvain to leiden: Guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233, 2019. doi: 10.1038/s41598-019-41695-z.

Chloe Wang, Haotian Cui, Andrew Zhang, Ronald Xie, Hani Goodarzi, and Bo Wang. scGPT-spatial: Continual Pretraining of Single-Cell Foundation Model for Spatial Transcriptomics. *bioRxiv*, 2025. doi: 10.1101/2025.02.05.636714. Preprint.

Hang Xu, Huazhu Fu, Yahui Long, Kok Siong Ang, Raman Sethi, Kelvin Chong, Mengwei Li, Rom Uddamvathanak, Hong Kai Lee, Jingjing Ling, et al. Unsupervised spatially embedded deep representation of spatial transcriptomics. *Genome Medicine*, 16(1):12, 2024.

Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nature Machine Intelligence*, 4(10):852–866, 2022.

Zhiyuan Yuan, Fangyuan Zhao, Senlin Lin, Yu Zhao, Jianhua Yao, Yan Cui, Xiao-Yong Zhang, and Yi Zhao. Benchmarking spatial clustering methods with spatially resolved transcriptomics data. *Nature Methods*, 21(4):712–722, 2024.

Wang Zeng, Sheng Jin, Lumin Xu, Wentao Liu, Chen Qian, Wanli Ouyang, Ping Luo, and Xiaogang Wang. TCFormer: Visual recognition via token clustering transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9521–9535, 2024. doi: 10.1109/TPAMI.2024.3425768.

Edward Zhao, Matthew R. Stone, Xing Ren, Jamie Guenthoer, Kimberly C. Smythe, Thomas Pulliam, Stephen R. Williams, Cedric R. Uytingco, Sarah E. B. Taylor, Paul Nghiem, et al. Spatial transcriptomics at subspot resolution with bayesspace. *Nature Biotechnology*, 39(11):1375–1384, 2021.

Xiang Zhou, Kangning Dong, and Shihua Zhang. Integrating spatial transcriptomics data across different conditions, technologies and developmental stages. *Nature Computational Science*, 3 (10):894–906, 2023.

Xiaowei Zhuang. Spatially resolved single-cell genomics and transcriptomics by imaging. *Nature Methods*, 18(1):18–22, 2021.

APPENDIX CONTENTS

A Datasets 14

B Data Processing and Evaluation Splits 15
B.1 Expression preprocessing and gene mapping 15
B.2 Mouse heart dataset splitting 16
B.3 Human brain dataset splitting 17
B.4 Mouse embryo dataset splitting 18
B.5 Mouse whole-brain dataset splitting 19
B.6 Human lung dataset splitting 20

C Evaluation Protocols and Metrics 20
C.1 Semantic preservation metrics 20
C.2 Spatial domain metrics 21
C.3 Neighborhood composition prediction metric 22
C.4 Disease-relevant representation analysis 22

D Models and Implementation 22
D.1 Transcriptomic Backbones 22
D.2 Spatial foundation model baselines 23
D.3 Task-Specific Spatial Baselines 23
D.4 Implementation Details 24

E Additional Experimental Results 28
E.1 Disease-Relevant Analysis 28
E.2 Resolution Sensitivity 28
E.3 Full Ablation Results 29
E.4 Neighborhood Scales 30
E.5 Confidence Intervals 30

Code availability. The complete code is provided with the supplementary material.

A DATASETS

We evaluate SPARSE on five publicly available spatial transcriptomics datasets spanning multiple organs, species, and sequencing platforms, including sequencing-based and imaging-based technologies. These datasets are all associated with published studies or large-scale atlas projects and are publicly accessible through standard repositories.

Table A1: Summary of the five spatial transcriptomics datasets used in this study. The exact preprocessed cohorts and training–test partitions used by the experiments are reported in Appendix B.

Dataset	Species	Organ	Platform	Resolution	Scale (cells/spots)	Sections / Samples	Data source
Mouse heart	Mouse	Heart	Stereo-seq (DNBSEQ)	$\sim$25 μm (bin50); nanoscale capture ($\sim$500 nm)	$\sim$330,857 bins (main study); $\sim$186k cells (GEO subset)	22 sections (E20–P14) + GEO P0/P7/P56	GEO: GSE298650; STOmics
Human brain	Human	Cortex	Stereo-seq + snRNA-seq	$\sim$500 nm (subcellular)	$>$1M spatial cells + $\sim$300k nuclei	Multiple cortical regions	CNGB: CNP0005272; Sci China Life Sci
Mouse embryo (MOSTA)	Mouse	Whole embryo	Stereo-seq	Subcellular ($\sim$500 nm–1 μm)	Millions of spots	Multiple stages (E9.5–E16.5)	STOmics MOSTA
Mouse whole brain	Mouse	Brain	MERFISH (imaging-based)	Single-cell ($\sim$ μm)	Millions of cells	Whole-brain atlas	Allen Brain Atlas
Human lung	Human	Lung	Xenium (10x Genomics)	Subcellular ($\sim$ μm)	$\sim$300k–500k cells per section	Multiple sections (normal + tumor)	10x Xenium dataset

Mouse heart dataset. The mouse heart dataset is derived from the study "Exploring the cellular and molecular basis of murine cardiac development through spatiotemporal transcriptome sequencing", with data deposited in both the Gene Expression Omnibus (GEO; accession GSE298650) and the STOmics database. This dataset was generated using the Stereo-seq (SpaTial Enhanced REsolution Omics-sequencing) platform, a high-resolution spatial transcriptomics technology based on DNA nanoball-patterned arrays and in situ RNA capture followed by reverse transcription and DNBSEQ sequencing. The dataset covers multiple developmental stages, including embryonic day 20 (E20) and postnatal days P01, P04, and P14, and contains 22 high-quality tissue sections. After spatial binning (bin size = 50, approximately 25 μm), the dataset comprises 330,857 spatially resolved cell bins annotated into 14 cell types, enabling near single-cell resolution spatial profiling of cardiac development. In addition, related releases of the same dataset series report approximately 0.186 million cells across postnatal stages (P0, P7, P56), further demonstrating its scale and consistency. Raw sequencing data ($\sim$31 GB) and processed matrices are publicly available, including spatial gene expression matrices and histological images.

Human brain dataset. The human brain dataset is obtained from a spatial transcriptomics study published in *Science China Life Sciences* and further supported by the CNGB project (accession CNP0005272). This dataset is generated using Stereo-seq technology and integrates spatial transcriptomics with single-nucleus RNA sequencing (snRNA-seq), enabling joint analysis of spatial gene expression and cell-type annotation. The dataset provides a large-scale spatial atlas of the human cortex, comprising hundreds of thousands of nuclei from snRNA-seq and over one million spatially resolved cells, with subcellular spatial resolution enabled by nanoscale capture spots ($\sim$500 nm). The use of Stereo-seq chips, typically 10 mm $\times$ 10 mm, allows dense spatial sampling with tens of thousands to hundreds of thousands of spatial bins per section, supporting detailed characterization of neuronal and glial diversity as well as laminar cortical organization.

Mouse embryo dataset. The mouse embryo dataset is obtained from the MOSTA (Mouse Organogenesis Spatial Transcriptomic Atlas) project hosted in the STOmics database, which is associated with multiple published Stereo-seq studies on embryogenesis. This dataset is generated using Stereo-seq technology and provides whole-embryo spatial transcriptomic maps across multiple developmental stages, e.g., E9.5–E16.5. It contains millions of spatial capture spots across sections

at subcellular resolution, on the order of hundreds of nanometers, enabling reconstruction of developmental trajectories, spatial lineage specification, and organogenesis processes. The sequencing workflow follows the standard Stereo-seq protocol, including in situ RNA capture on patterned arrays and high-throughput sequencing.

Mouse whole-brain dataset. The mouse whole-brain dataset is obtained from the Allen Institute's ABC Atlas, specifically the MERFISH-based imputed dataset (MERFISH-C57BL6J-638850). This dataset is generated using MERFISH (Multiplexed Error-Robust Fluorescence In Situ Hybridization), an imaging-based spatial transcriptomics technology that directly measures RNA molecules in situ through combinatorial fluorescence labeling. Unlike sequencing-based methods, MERFISH provides highly accurate spatial localization without amplification bias. The dataset covers the entire mouse brain and contains millions of cells at near single-cell resolution, with expression profiles measured for hundreds to thousands of genes, forming one of the most comprehensive spatial atlases of the mammalian brain.

Human lung dataset. The human lung dataset is obtained from the Xenium Human Lung Preview dataset released by 10x Genomics, which is accompanied by technical documentation and application reports. This dataset is generated using the Xenium in situ spatial transcriptomics platform on FFPE (formalin-fixed paraffin-embedded) tissue sections. The platform uses probe-based RNA detection combined with high-resolution imaging to achieve subcellular spatial resolution. Each tissue section contains approximately 300,000 to 500,000 cells, with gene expression measured using targeted panels covering hundreds to thousands of genes. The dataset includes both healthy lung tissue and lung adenocarcinoma samples, enabling analysis of tissue architecture, tumor microenvironment, and cellular heterogeneity.

B Data Processing and Evaluation Splits

B.1 Expression preprocessing and gene mapping

Cell and gene quality control. Cells with fewer than 10 detected genes are removed, where a gene is detected when its expression value is nonzero; cells with exactly 10 detected genes are retained. We additionally require the sample identifier, finite two-dimensional spatial coordinates, and the annotation needed to define the corresponding evaluable cohort. Evaluation annotations do not enter any expression transformation or representation-parameter update. Gene symbols are taken from the available feature annotations. Missing or empty identifiers and duplicate symbols are removed or resolved during dataset assembly when present. Symbols are converted to strings and matched case-insensitively to each backbone's vocabulary. The retained coordinates, cell order, annotations, and expression rows are propagated together so that the embeddings, neighborhoods, and reconstruction targets refer to exactly the same cells.

Expression target and HVG selection. The frozen-backbone input and the masked-expression reconstruction target follow separate preprocessing paths. The 2,000 highly variable genes (HVGs) are *not* used as a common input vocabulary for the foundation models. Instead, they define a backbone-specific reconstruction target, whereas each frozen backbone receives all genes supported by its own vocabulary or reference panel. In the quantitative experiments, the scGPT, Geneformer, and scFoundation target values are library-size normalized to 10,000 counts per cell and transformed as

$$\widetilde{x}_{ig} = \log\left(1 + \frac{10^4 x_{ig}}{\sum_{g'} x_{ig'}}\right). \tag{23}$$

scBERT uses the corresponding base-2 transformation after alignment to its reference panel. HVGs are selected once using the Seurat-v3 procedure, with the top 2,000 genes retained and no batch stratification. If fewer than 2,000 genes are available, the full panel is retained. For scGPT and scFoundation, HVGs are identified from the normalized, log-transformed expression matrix; Geneformer ranks HVGs on the count scale and transforms the selected target values afterward; scBERT selects them after reference-panel alignment and base-2 log transformation. For the 392-gene human-lung analysis, no additional HVG selection is required and the full targeted panel is used.

Backbone-specific mapping. Table B1 records the model-specific mapping and missing-gene behavior. Except for Geneformer's official symbol-to-Ensembl dictionary, no gene-alias expansion or cross-species ortholog conversion is applied.

Table B1: Backbone-specific gene mapping and expression input. The HVG panel described above is used as the reconstruction target and does not replace the backbone-specific input vocabulary.

Backbone	Gene mapping	Unsupported or missing genes	Input to the frozen backbone
scGPT	Uppercase gene symbol to the checkpoint vocabulary.	Input genes outside the vocabulary are omitted; absent vocabulary genes are not zero-filled.	Library-size normalization to 10^4 followed by the natural transformation $\log(1+x)$; supported nonzero genes are processed by the official cell-embedding tokenizer.
Geneformer	Uppercase gene symbol to Ensembl ID using the official name dictionary, followed by lookup in the token dictionary.	Unsupported and unexpressed genes are omitted; the padding token is used only to equalize sequence lengths within a batch.	Raw counts are divided by the cell total, scaled to 10^4, divided by the official gene-specific median, and ranked; no log transformation is applied to the backbone input.
scFoundation	Uppercase gene symbol to the official 19,264-gene reference order.	Unrecognized input genes are omitted; reference genes absent from the measured panel remain zero.	A fixed-length aligned count vector is normalized internally to 10^4 and transformed as $\log(1+x)$; library-size covariates are appended by the official preprocessing procedure.
scBERT	Uppercase gene symbol to the 16,906-gene Panglao reference order.	Unrecognized input genes are omitted and unmeasured reference genes remain zero; the 10-detected-gene filter is applied again after alignment.	The aligned vector is normalized to 10^4 and transformed with $\log_2(1+x)$, then clipped to $[0, 5]$ and integer encoded for the pretrained model.

B.2 Mouse heart dataset splitting

After preprocessing, the mouse-heart cohort contains 330,857 cells and 23,397 genes. SPARSE adapts on 21 sections and is evaluated directly on held-out P14_1_1; its cells and neighborhoods never enter adaptation. This section-level split avoids the spatial leakage possible with a random cell-level split.

Table B2: Mouse-heart section inventory and adaptation split.

E20 and P1 adaptation			P4 and P14 adaptation		
Section	**Cells**	**Types**	**Section**	**Cells**	**Types**
E20_1_1	9,129	14	P4_1_1	14,833	14
E20_1_2	9,331	14	P4_1_2	13,949	14
E20_1_3	9,032	14	P4_2_1	14,356	12
E20_2_1	11,058	14	P4_2_2	13,747	13
E20_2_2	11,752	14	P4_2_3	14,151	13
E20_2_3	10,194	14	P14_1_2	25,917	13
P1_1_1	8,726	13	P14_1_3	27,406	14
P1_1_2	9,321	13	P14_2_1	27,619	14
P1_2_1	9,106	14	P14_2_2	25,408	14
P1_2_2	8,885	14	P14_2_3	24,195	14
P1_2_3	8,956	14			
E20: 60,496; P1: 44,994 cells			P4: 71,036; P14: 130,545 cells		

Held out: P14_1_1 (23,786 cells; 14 cell types)
Adaptation total: 21 sections, 307,071 cells, 14 cell types
Overall: 22 sections, 330,857 cells, 14 cell types



Figure B1: Mouse-heart section geometry and adaptation split. Each panel shows one section: gray marks adaptation and teal marks held-out P14_1_1. Plotted points are a deterministic subsample of quality-controlled cells; Table B2 counts all retained cells.

B.3 Human brain dataset splitting

The human brain dataset contains 1,693,487 cells and 48,878 genes after preprocessing. Unlike datasets with explicit section identifiers, it was provided as a spatially assembled map containing multiple tissue areas. As shown in Figure B2, white gaps and discontinuities delineate spatially separated areas, but the original physical-section identifiers were unavailable. We therefore defined two held-out test regions by spatial coordinates and used all remaining cells for adaptation. These regions are not asserted to be independent physical sections.

Specifically, cells with coordinates $x \in [-36{,}000, -20{,}000]$ and $y \in [-16{,}000, 0]$ were assigned to test region 1, while cells with coordinates $x \in [4{,}000, 20{,}000]$ and $y \in [-16{,}000, 0]$ were assigned to test region 2. SPARSE and the transfer baselines were adapted without either test region and evaluated without target-region fine-tuning.

Table B3: Human-brain coordinate-defined adaptation and evaluation regions.

Split	Subset	X range	Y range	Cells	Domain types
Train	Remaining tissue	outside test regions	outside test regions	1,433,566	17
Test	region 1	[-36,000, -20,000]	[-16,000, 0]	122,325	17
Test	region 2	[4,000, 20,000]	[-16,000, 0]	137,596	16
Overall	All retained cells	–	–	1,693,487	17

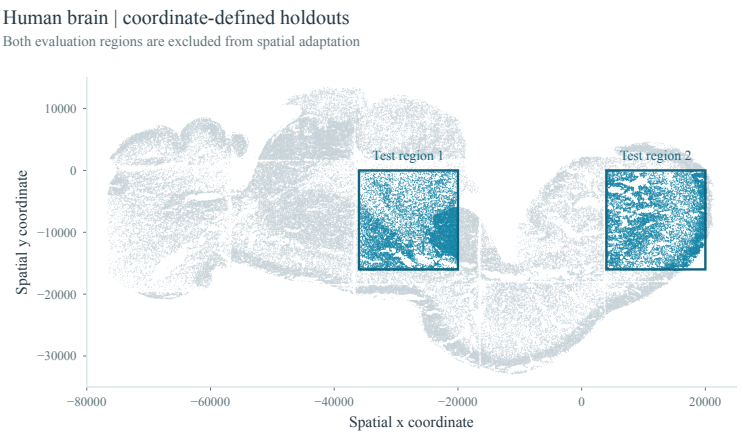

Figure B2: Human-brain coordinate-defined holdouts. Gray denotes cells available for spatial adaptation; teal denotes cells in the two withheld rectangles. The boxes correspond to the exact coordinate ranges in Table B3. Points are deterministically subsampled for display, while all retained cells are used for the reported split statistics. The original physical-section identifiers were unavailable.

B.4 Mouse embryo dataset splitting

The mouse embryo dataset contains 520,754 cells and 23,761 genes after preprocessing. It covers eight developmental timepoints from E9.5 to E16.5 and includes 45 domain annotation types. Here, E denotes embryonic day; for example, E9.5 refers to embryonic day 9.5, and E16.5 refers to embryonic day 16.5.

For this dataset, different sections correspond to different embryonic developmental timepoints. We constructed a held-out-stage transfer setting to evaluate whether the learned spatial representations can be applied to a later developmental stage. Specifically, cells from E9.5, E10.5, E11.5, E12.5, E13.5, E14.5, and E15.5 were used for adaptation, while cells from the unseen E16.5 stage were held out for testing. Unlike the human-brain experiment, which evaluates coordinate-defined spatial holdouts, this experiment evaluates transfer from earlier developmental stages to a later stage. Models were directly evaluated on the full E16.5 expression matrix and coordinates without additional representation fitting; this is not a forecast from earlier stages alone.

Table B4: Mouse-embryo stage inventory and temporal adaptation split.

Role	Stage	Cells	Domain types
Adaptation	E9.5	5,912	12
	E10.5	18,399	12
	E11.5	30,121	19
	E12.5	51,351	23
	E13.5	77,354	19
	E14.5	102,519	26
	E15.5	113,338	26
Held out	**E16.5**	**121,760**	**25**
Adaptation total	E9.5–E15.5	398,994	43
Overall	E9.5–E16.5	520,754	45

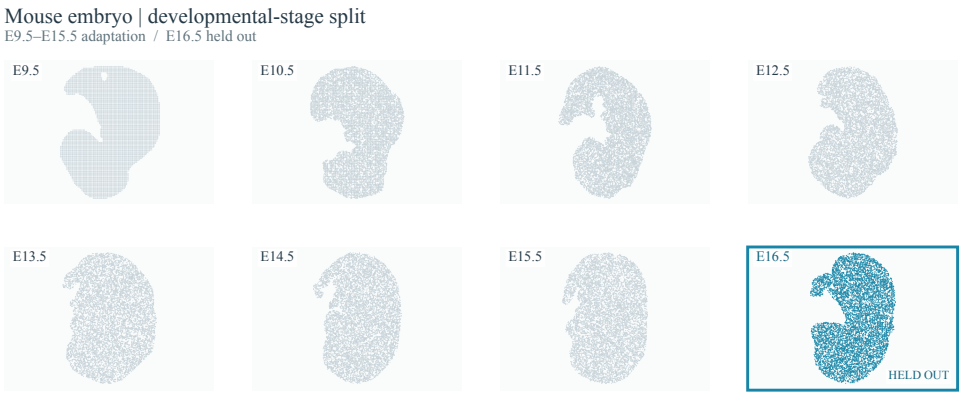

Figure B3: Mouse-embryo developmental-stage geometry. Gray denotes adaptation stages E9.5–E15.5; teal identifies the held-out E16.5 stage. E denotes embryonic day. Points are deterministically subsampled for display; the stage statistics in Table B4 use all retained cells.

B.5 Mouse whole-brain dataset splitting

The mouse whole-brain dataset contains 382,379 cells and 1,120 genes after preprocessing and section selection. Since the original dataset contains many brain sections, we selected ten developmentally similar sections for evaluation: C57BL6J-1.093 and C57BL6J-1.095 through C57BL6J-1.103.

For this dataset, we evaluated cell-type label transfer and cell-wise tissue-region classification using cell_type and tissue annotations, respectively. We adopted a cross-section evaluation setting, where section C57BL6J-1.099 was held out as the test section and the remaining nine sections were used for training. This setting evaluates whether the learned spatial representations can generalize to an unseen brain section under a similar developmental context.

Table B5: Mouse whole-brain section inventory and adaptation split.

Role	Section suffix	Cells	Cell types	Tissue labels
Adaptation	1.093	38,268	22	9
	1.095	36,706	21	9
	1.096	35,361	20	8
	1.097	34,784	20	9
	1.098	39,340	18	9
Held out	**1.099**	**39,026**	**21**	**10**
Adaptation	1.100	38,450	19	10
	1.101	38,474	20	10
	1.102	42,518	20	9
	1.103	39,452	20	9
Adaptation total	9 sections	343,353	23	11
Overall	10 sections	382,379	23	11

All section identifiers in Table B5 share the prefix C57BL6J-; only the suffix is printed to avoid repeating it in every row. Section C57BL6J-1.099 is excluded from adaptation and used for all three fixed-embedding downstream evaluations.

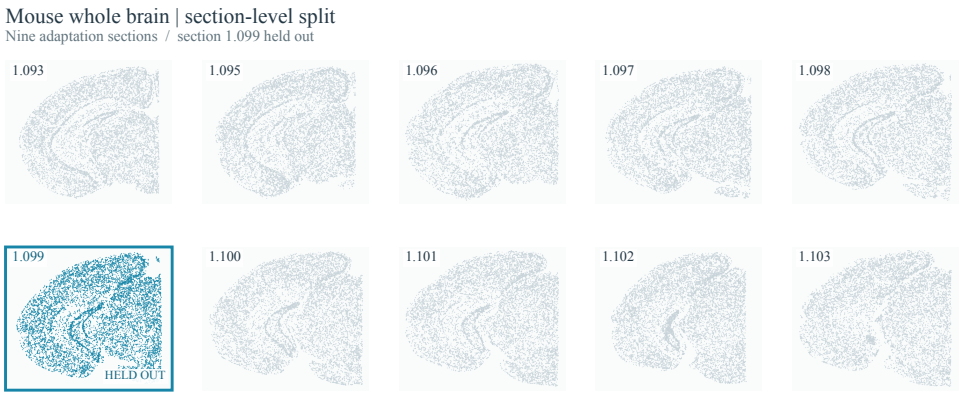

Figure B4: Mouse whole-brain section geometry and adaptation split. Gray marks nine adaptation sections; teal marks the held-out section C57BL6J-1.099. Panel labels use the C57BL6J- section suffixes listed in Table B5. Points are deterministically subsampled for display, while all retained cells contribute to the split statistics.

B.6 Human lung dataset splitting

The human lung dataset was constructed from two tissue sections, including a healthy lung section and a lung cancer section. We assigned the condition label healthy to the healthy lung section and cancer to the lung cancer section. After cleaning gene names, removing invalid genes and cells, and retaining the shared gene set between the two sections through inner-join merging, the merged dataset contains 817,443 cells and 392 genes. The dataset was randomly shuffled and split into training and test sets with a 9:1 ratio. This dataset was used for disease-relevant representation analysis based on the condition label.

Table B6: Preprocessed statistics and 90%/10% random split of the merged human lung dataset used for the qualitative condition-associated representation analysis.

Split	Cells	Genes	Healthy cells	Cancer cells	labels
Train	735,698	392	262,300	473,398	2
Test	81,745	392	29,066	52,679	2
Overall	817,443	392	291,366	526,077	2

C Evaluation Protocols and Metrics

To ensure fair and consistent evaluation across datasets and tasks, we used the same evaluation metrics for all compared methods under the same train/test splits and downstream protocols. Since our experiments cover classification, spatial domain identification, neighborhood composition prediction, and disease-relevant representation analysis, we report task-specific metrics according to the nature of each downstream task.

C.1 Semantic preservation metrics

For cell-type annotation and cell-wise tissue-region classification, the main paper reports Accuracy and Weighted-F1. Accuracy measures the fraction of correctly predicted samples:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{y}_i = y_i),$$

where N is the number of samples, y_i is the ground-truth label, $\hat{y}_i$ is the predicted label, and $\mathbb{I}(\cdot)$ is the indicator function.

For each class c, precision, recall, and F1-score are

$$\begin{aligned}\text{Precision}_c &= \frac{TP_c}{TP_c + FP_c},\\ \text{Recall}_c &= \frac{TP_c}{TP_c + FN_c},\\ \text{F1}_c &= \frac{2\,\text{Precision}_c\,\text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}.\end{aligned}$$

where TP_c, FP_c, and FN_c denote the numbers of true positives, false positives, and false negatives for class c, respectively. Weighted-F1 is the class-frequency-weighted average

$$\text{Weighted-F1} = \sum_{c=1}^{C} \frac{n_c}{N} \text{F1}_c,$$

where C is the number of classes and n_c is the number of samples belonging to class c.

For the mouse-heart semantic-retention experiment, the representation adapter is trained without slice P14_1_1, and fixed embeddings are then extracted for its 23,786 cells. For each random seed, these cells are divided into stratified 50% reference and 50% query sets. Feature standardization is fitted on the reference embeddings and then applied to the query embeddings. Labels are transferred from the 15 Euclidean nearest reference cells using inverse-distance weighting.

The mouse whole-brain cell-level experiment follows the same standardized, distance-weighted 15-nearest-neighbor protocol on the held-out section C57BL6J-1.099. Cell types represented by fewer than two cells are removed before the stratified 50%/50% split, leaving 39,025 cells and 20 evaluated cell types. For the tissue-region task, all 39,026 cells and 10 classes are retained and evaluated with a stratified 50%/50% linear-probe split. Probe optimization details are provided in Appendix D.4.

C.2 Spatial domain metrics

For spatial domain identification and clustering-based evaluation, we report Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). These metrics compare the predicted cluster assignments with ground-truth domain labels.

Adjusted Rand Index (ARI) measures the agreement between two partitions after correcting for chance. Given a set of samples, ARI compares whether pairs of samples are assigned consistently in the predicted clusters and ground-truth labels. ARI is defined as:

$$\text{ARI} = \frac{\text{RI} - \mathbb{E}[\text{RI}]}{\max(\text{RI}) - \mathbb{E}[\text{RI}]},$$

where RI is the Rand Index and $\mathbb{E}[\text{RI}]$ is its expected value under random labeling. A higher ARI indicates better agreement between predicted spatial domains and ground-truth annotations. ARI ranges from -1 to 1, with 1 indicating perfect agreement and values close to 0 indicating random agreement.

Normalized Mutual Information (NMI) measures the shared information between predicted clusters and ground-truth labels:

$$\text{NMI}(Y, \hat{Y}) = \frac{2I(Y; \hat{Y})}{H(Y) + H(\hat{Y})},$$

where Y denotes the ground-truth labels, $\hat{Y}$ denotes the predicted clusters, $I(Y; \hat{Y})$ is their mutual information, and $H(\cdot)$ is entropy. NMI ranges from 0 to 1, with higher values indicating stronger consistency between predicted clusters and true domain labels.

In the spatial-domain experiments, we construct a 15-nearest-neighbor graph from each test embedding and apply Leiden clustering. The main-paper results use resolution 0.8 for every method and dataset. Appendix E.2 additionally reports the complete resolution sweep from 0.1 to 1.0, so the comparison does not depend on a method-specific test resolution.

C.3 NEIGHBORHOOD COMPOSITION PREDICTION METRIC

For the composition task, we construct radius-based spatial neighborhoods and predict a neighborhood target from fixed cell embeddings. Let $X_l \in \{0,1\}^{N\times K}$ be the one-hot cell-type matrix and let $A^{(m)}$ be a binary radius graph without self-loops. For each target scale $m \in \{100, 200, 500\}$, the radius is selected by binary search so that the mean number of neighbors is within 0.5 of m. The target distribution is

$$p_i^{(m)} = \mathrm{softmax}\left(\left[A^{(m)}X_l\right]_{i,:}\right).$$

A linear layer followed by softmax predicts $\hat{p}_i^{(m)}$ from each fixed embedding. For every seed, cells from the held-out C57BL6J-1.099 section are randomly divided into 80% probe-training and 20% test sets. The probe is optimized for one epoch with mean-squared error, and Mean Absolute Error (MAE) is evaluated on the test cells:

$$\mathrm{MAE}^{(m)} = \frac{1}{N_{\mathrm{test}}K}\sum_{i\in\mathcal{I}_{\mathrm{test}}}\sum_{k=1}^{K}\left|\hat{p}_{i,k}^{(m)} - p_{i,k}^{(m)}\right|.$$

Lower MAE indicates more accurate recovery of the radius-based neighborhood target. For each seed, Mean MAE is the arithmetic mean of $\mathrm{MAE}^{(100)}$, $\mathrm{MAE}^{(200)}$, and $\mathrm{MAE}^{(500)}$. The reported mean and standard deviation are then computed across seeds 0, 42, and 123.

C.4 DISEASE-RELEVANT REPRESENTATION ANALYSIS

For the human lung dataset, we perform disease-relevant representation analysis using the condition labels healthy and cancer. The goal of this experiment is not to construct a standard supervised classification benchmark, but to examine whether the learned representation space reflects disease-related biological variation.

After training, we extract SPARSE embeddings for test cells and visualize them using UMAP. The resulting low-dimensional representation is colored by condition label to assess whether healthy and cancer cells exhibit condition-dependent organization in the embedding space. Since this analysis is primarily qualitative, we do not use it as the main quantitative benchmark; instead, it provides complementary evidence that the learned spatial representations can preserve disease-relevant biological structure.

D MODELS AND IMPLEMENTATION

We compare SPARSE with non-spatial representations, task-specific spatial methods, spatial foundation models, and controlled lightweight adapters. We report the software versions, released pretrained models, and embedding-extraction rules used in the experiments.

D.1 TRANSCRIPTOMIC BACKBONES

Table D1 lists the transcriptomic pretrained models and extraction rules. The same pretrained model and extraction rule are used across datasets. These embeddings are frozen before any spatial adapter is trained.

Table D1: Transcriptomic backbone releases and cell-embedding extraction rules. These representations are computed once, kept frozen during spatial adaptation, and used without evaluation annotations.

Backbone	Version and pretrained model	Embedding extraction
scGPT	Official scGPT version 0.2.4; whole-human pretrained model.	We use the final encoder representation of the classification token followed by per-cell ℓ_2 normalization.
Geneformer	Geneformer version 0.1.0; Geneformer-V2-316M.	We use the final-layer representation of the leading classification token. No token pooling or additional normalization is applied.
scFoundation	Official scFoundation cell-embedding pretrained model with the released target-resolution configuration.	We use the final encoder token representations and apply max pooling over the token axis, matching the official extraction rule.
scBERT	Official Panglao-pretrained scBERT model with the released Panglao reference gene set.	A classification token is appended to the expression-token sequence, and its final encoder representation is used directly, without additional pooling or normalization.

D.2 Spatial foundation model baselines

Table D2 lists the released spatial-model configurations; neither checkpoints nor inference settings are selected using downstream annotations.

Table D2: Spatial foundation model releases and inference configurations for held-out evaluation.

Method	Version and pretrained model	Inference and extraction
Novae	Novae 1.0.3; the released brain model for human brain and the released mouse model for mouse datasets.	We construct the default Delaunay spatial graph and use the released model in its zero-shot inference setting. The per-cell encoder representation is used directly, without custom pooling.
Nicheformer	Nicheformer 0.0.1; released pretrained model.	The spatial-modality token is used for every dataset, with the human or mouse species token as appropriate and assay-context token 11 fixed by the inference configuration. Inputs use a maximum sequence length of 1497 plus three context tokens. We extract the last transformer layer, remove the three context tokens, and mean-pool the remaining token representations.
scGPT-spatial	scGPT-spatial 0.1.0; official V1 pretrained model.	We use a maximum sequence length of 1200 and the final classification-token representation followed by per-cell ℓ_2 normalization. MVC, MVC imputation, and the mixture-of-experts decoder are enabled; pre-normalization, domain-adversarial, batch-label, generative-training, and fast-transformer components are disabled.

D.3 Task-Specific Spatial Baselines

PCA, SpaGCN, and SCAN-IT are fitted independently to each unlabeled evaluation sample. PCA uses only its expression matrix, whereas SpaGCN and SCAN-IT additionally use its spatial coordinates. Evaluation annotations are never used to fit these representations.

PCA. We use scikit-learn 1.3.2 and Scanpy 1.10.3. Each evaluation sample is library-size normalized to 10^4, log-transformed, filtered to retain cells expressing at least 10 genes, and restricted to the top 2,000 Seurat-v3 highly variable genes. PCA is then fitted directly to that sample using the corresponding experimental seed. We do not apply feature scaling or whitening, and the resulting cell scores are used directly without pooling.

SpaGCN. We use SpaGCN 1.2.7 without histology. To accommodate the large evaluation samples, the dense spatial-distance matrix is replaced by a symmetric sparse KNN distance graph; the original Gaussian weighting, graph-convolutional DEC objective, and convergence rule are otherwise retained. The Gaussian length scale is selected with target $p = 0.5$ over $[0.01, 1000]$, tolerance 0.01, and at most 100 search iterations. For human brain, we filter genes observed in fewer than three cells, remove special genes, apply per-cell normalization and the transformation $\log(1 + x)$, select 2,000 Seurat-v3 highly variable genes, and construct a 30-nearest-neighbor spatial graph. Optimization uses Adam with learning rate 0.005, no weight decay, at most 2,000 epochs, and Louvain initialization with 10 neighbors, resolution 0.4, and convergence tolerance 10^{-3}. For mouse embryo,

we remove special genes, apply per-cell normalization and $\log(1+x)$, and use a 10-nearest-neighbor spatial graph without an additional HVG-selection step. Optimization uses Adam with learning rate 0.05, no weight decay, at most 40 epochs, and Louvain initialization with 300 neighbors, resolution 0.1, and convergence tolerance 5×10^{-3}. We use the trained graph-convolution output for each cell without pooling.

SCAN-IT. We use SCAN-IT 0.1 with NetworkX 3.2.1. After selecting 2,000 Seurat-v3 highly variable genes, the spatial graph is constructed with the alpha-shape method using one alpha layer and an automatically estimated cutoff based on 10-nearest-neighbor distances. The DGI representation model is optimized for 1,000 epochs with learning rate 10^{-3} and one consensus run. Python, NumPy, and PyTorch use the corresponding experimental seed. We use the learned per-cell representation directly, without pooling.

D.4 Implementation Details

Hyperparameter specification

We distinguish fixed architectural choices from downstream evaluation parameters. The spatial-neighborhood size and rCTM token-retention ratio were fixed uniformly across datasets, backbones, and random seeds rather than selected separately for each test sample. After fixing this common configuration, we evaluated the alternative values in Table D3 as controlled sensitivity settings. The final values reflect architectural and computational considerations rather than optimization against a labeled downstream metric.

Table D3: Structural hyperparameter specification. K excludes the centered target token, and ρ denotes the fraction of tokens retained by rCTM between adjacent stages.

Hyperparameter	Values examined	Final value	Selection principle
Spatial neighbors K	$\{99, 199, 499, 999\}$	199	Together with the centered target, $K = 199$ forms a 200-token neighborhood, balancing local-context coverage against the quadratic cost of attention.
rCTM retention ratio ρ	$\{0.25, 0.50, 0.75\}$	0.50	A one-half retention ratio produces the regular hierarchy $200 \to 100 \to 50 \to 25$, avoiding both overly aggressive early compression and limited token reduction.

The masking ratio was fixed a priori at 0.20 for every adaptation run and was not swept or selected using downstream labels. Leiden resolution is not a training hyperparameter of SPARSE: the primary tables use a common resolution of 0.8 for all methods, while Appendix E.2 reports the complete range from 0.1 to 1.0. No method-specific resolution is selected for the primary comparison. Figures D1 and D2 provide the complete seed-42 sensitivity curves for the two held-out human-brain regions. The final shared settings are highlighted for reference; the plots are not used for region- or metric-specific model selection.

Taken together, these plots show why K and ρ are treated as shared structural settings rather than tuned downstream parameters. The best curve can vary with slice, metric, or Leiden resolution, whereas $K = 199$ provides a fixed 200-token local field and $\rho = 0.50$ provides a regular four-stage hierarchy. We therefore keep both values unchanged in every reported experiment.

Final SPARSE architecture

Table D4 records the complete configuration used to produce the quantitative SPARSE embeddings. The architecture is shared across datasets and transcriptomic backbones; only the input projection dimension and the number of reconstructed genes depend on the backbone and dataset. The effective input to the adapter is always the centered target together with its 199 nearest spatial neighbors.

Within rCTM, DPC center selection and assignment are non-differentiable routing operations evaluated without gradient tracking. Pairwise Euclidean feature distances are divided by the square root of the projected feature dimension. The density of a token is the negative mean squared distance to its five nearest entries plus its learned raw token score; center priority is the product of this density

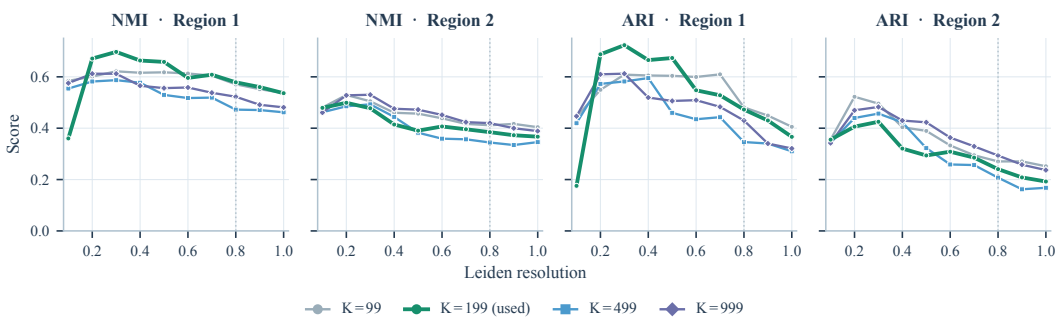

Figure D1: Sensitivity to the spatial-neighborhood size K on the two held-out human-brain regions (seed 42). From left to right, panels show NMI for regions 1 and 2, followed by ARI for regions 1 and 2. Each curve spans Leiden resolutions 0.1–1.0; the dotted vertical line marks the common evaluation resolution 0.8. The highlighted $K = 199$ curve is the fixed setting shared by all main experiments, not a region-specific optimum.

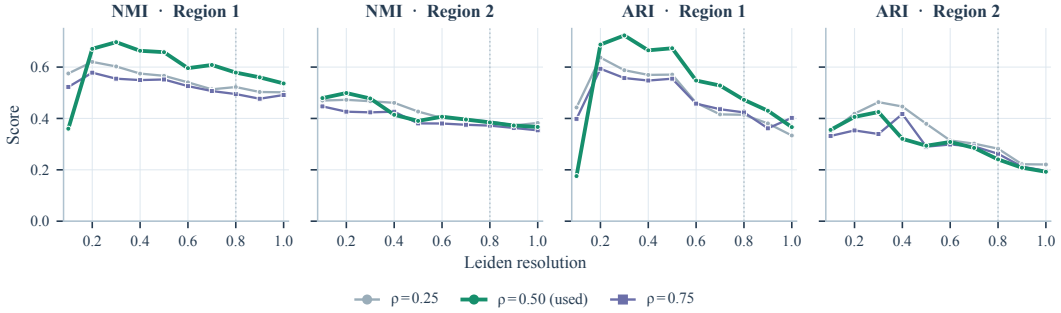

Figure D2: Sensitivity to the rCTM token-retention ratio ρ on the two held-out human-brain regions (seed 42). Panel order, metrics, and resolution range follow Figure D1. The highlighted $\rho = 0.50$ setting yields the fixed $200 \to 100 \to 50 \to 25$ hierarchy used throughout the paper.

Table D4: Final SPARSE architecture. Values in brackets are ordered from the finest to the coarsest stage.

Component	Final configuration
Centered spatial input	One target plus $K = 199$ Euclidean spatial nearest neighbors (200 tokens in total); token counts $[200, 100, 50, 25]$.
Encoder stages	Feature dimensions $[64, 128, 320, 512]$; depths $[1, 1, 1, 1]$; attention heads $[1, 2, 5, 8]$; per-head dimension 64; MLP ratios $[8, 8, 4, 4]$.
Attention	QKV bias enabled; query–key scale $d_{\text{head}}^{-1/2}$; projection/MLP dropout 0; attention dropout 0.
Encoder stochastic depth	Maximum drop-path rate 0.10, linearly assigned across the four encoder blocks as $[0, 0.0333, 0.0667, 0.10]$.
rCTM	Retention ratios $[0.50, 0.50, 0.50]$; next-stage token count $\lceil 0.5 M_s \rceil$; feature-space DPC-KNN with $k = 5$ nearest entries including the token itself. rCTM uses no temperature or distance-cutoff hyperparameter.
Coordinate propagation	Feature and coordinate aggregation use the same normalized weights $\exp(r_j)$; coordinates do not enter DPC center selection or assignment.
Spatial RoPE	Base 100; independent bounding-box normalization of each coordinate axis to $[-1, 1]$ within every centered neighborhood. The original neighborhood bounding box is reused at all encoder stages.
HSF	All lateral projections output 64 dimensions. The three top-down cross-attention blocks use 2 heads, MLP ratio 4, and drop-path rates $[0, 0.05, 0.10]$; the final refinement block uses 2-head self-attention, MLP ratio 4, and zero drop path. HSF uses neither token-score bias nor additional RoPE.
Mask token	Shape $1 \times 1 \times d_0$ and initialization $\mathcal{N}(0, 0.02^2)$, where d_0 is the input backbone dimension. A uniformly sampled 20% of the 200 tokens (40 tokens) is masked without replacement.
Reconstruction head	$\text{Linear}(64, G)$ followed by $\text{LayerNorm}(G)$, with no hidden layer, output activation, or decoder dropout. Here $G = 2000$ for human brain, mouse embryo, and mouse heart; $G = 1120$ for mouse whole brain; and $G = 392$ for human lung.
Adaptation objective	Mean cosine reconstruction loss over masked positions only; no auxiliary consistency loss.

and the distance to the nearest token of higher density. The selected centers are the top $\lceil 0.5M_s \rceil$ priorities, and all remaining tokens are assigned to their nearest center in feature space. The learned score remains trainable through its use in weighted merging and as an additive key-side attention bias.

Optimization and checkpoint protocol

The optimization settings in Table D5 are taken from the configurations stored in the epoch-30 checkpoints rather than from class defaults. In particular, the learning-rate scheduler is updated after every optimizer step.

Table D5: SPARSE optimization and checkpoint configuration used in the quantitative experiments. Every downstream embedding is exported from the prespecified epoch-30 checkpoint without metric-based checkpoint selection.

Setting	Value
Optimizer	AdamW; base learning rate 10^{-4}; minimum learning rate 10^{-6}; weight decay 0.05; $\beta = (0.9, 0.999)$; $\epsilon = 10^{-8}$.
Learning-rate schedule	Linear warmup for 10 epochs from 10^{-7} to 10^{-4}, followed by cosine annealing toward 10^{-6}. The cosine horizon follows the configured run length.
Batch and updates	One training process on a single GPU; batch size 64; no gradient accumulation; incomplete final batches are dropped.
Data loading	Four workers; pinned memory, persistent workers, and prefetch factor 2.
Precision	BF16 automatic mixed precision without dynamic loss scaling.
Gradient clipping	Global gradient-norm clipping at 1.0.
Configured run length	40 epochs for all mouse-heart runs and for seeds 0 and 123 on human brain, mouse embryo, and mouse whole brain; 60 epochs for the corresponding seed-42 runs on the latter three datasets. Every quantitative run therefore contains the prespecified epoch-30 checkpoint.
Checkpointing	Saved every two epochs; no metric-based best-checkpoint saving and no early stopping. Downstream embeddings are always exported from the epoch-30 checkpoint.

Although the configured training horizon is 40 or 60 epochs depending on the run, it is not used for downstream model selection. To avoid downstream-label-driven checkpoint selection, every SPARSE embedding reported in the paper is exported from epoch 30 and then kept fixed, including the qualitative human-lung visualization in Appendix E.1. We do not select checkpoints using NMI, ARI, accuracy, weighted F1, or any evaluation annotations.

Computational cost

Table D6 reports the system cost of the seed-42 scGPT-based SPARSE runs on a single 40 GB NVIDIA A100 GPU. Every measurement ends at the completion of epoch 30, matching the checkpoint used to produce both the quantitative embeddings and the qualitative human-lung visualization. Wall-clock time is measured from the beginning of SwanLab system monitoring to the epoch-30 record. Peak allocated memory is the maximum GPU-memory value observed on the training device, and mean utilization is averaged over the same interval. Offline extraction of frozen scGPT embeddings and all downstream evaluations are excluded.

Table D6: Computational cost of the scGPT-based SPARSE configuration through epoch 30. Each row is a single seed-42 run on one 40 GB NVIDIA A100 GPU. Memory is the peak allocation on the training GPU, and utilization is the temporal mean over the reported training interval.

Dataset	Time to epoch 30 (min)	Peak allocated memory (GB)	Mean GPU utilization (%)
Human brain	259.0	2.624	37.2
Mouse embryo	73.2	2.624	36.3
Mouse heart	54.0	2.624	37.2
Mouse whole brain	59.4	2.402	37.1
Human lung	118.2	2.242	37.3

Across these five runs measured through epoch 30, peak allocated memory ranges from 2.242 to 2.624 GB and mean GPU utilization ranges from 36.3% to 37.3%. Time to epoch 30 varies with the

size and composition of the precomputed neighborhood dataset, from 54.0 minutes for mouse heart to 259.0 minutes for human brain. These values document the resource footprint of SPARSE and are not used as baseline comparisons because the competing methods have different training and inference pipelines.

RANDOMNESS CONTROL AND DETERMINISTIC SETTINGS

For SPARSE and the other trainable adapters, each run identifier $s \in \{0, 42, 123\}$ pairs the seed-specific representation checkpoint with the same downstream seed. Adapter training seeds the Python, NumPy, PyTorch CPU, and CUDA random-number generators. We additionally disable cuDNN benchmarking and enable deterministic cuDNN execution. SPARSE training uses one process on one GPU; the three controlled adapters use four-process distributed training with a global batch size of 64. Random mask selection, data shuffling, and stochastic depth draw from the seeded PyTorch generators.

Checkpoint export is performed in evaluation mode with non-shuffled data loading and without masking or stochastic augmentation. It therefore introduces no additional random sampling for a fixed checkpoint and input. In spatial-domain evaluation, the same s controls Leiden and PCA, while SpaGCN and SCAN-IT seed their Python, NumPy, and PyTorch generators with s. The mouse-heart reference/query splits and the mouse whole-brain probe splits use stratified splitting controlled by s; when a trainable probe is used, s also controls its initialization and shuffled data loading. KNN transfer itself is deterministic after the split. Thus, for trained adapters, variation across seeds includes both adaptation and downstream randomness, whereas for fixed pretrained embeddings it reflects only the downstream split, probe, or clustering seed.

The settings above do not enforce deterministic algorithms globally or fix the CUDA workspace. We therefore control the principal software random-number generators but do not claim bitwise-identical results across different GPU architectures, CUDA versions, or library versions. Results reported as mean $\pm$ standard deviation use the sample standard deviation across seeds 0, 42, and 123. The ablation study is the stated exception: it uses seed 42 and averages the two held-out human-brain regions.

DOWNSTREAM EVALUATION

For the semantic-retention task on mouse heart and the cell-level task on mouse whole brain, we use the distance-weighted 15-nearest-neighbor protocols described in Appendix C.1 and do not fit a parametric classifier. For neighborhood-composition evaluation, the fixed embeddings are not standardized. At each neighborhood scale, a linear layer followed by softmax is optimized for one epoch with mean-squared error, Adam with a learning rate of 0.05, and a batch size of 256. For cell-wise tissue-region evaluation, a standardizer is fitted only on the probe-training split and applied to the test split. A single linear layer is then optimized for 30 epochs using cross-entropy loss, AdamW with a learning rate of 10^{-3} and weight decay of 10^{-4}, and a batch size of 256.

CONTROLLED ADAPTER BASELINES

The Linear, Transformer, and Laplacian Reg controls use the same centered input as SPARSE: one target cell or spot and its 199 precomputed spatial neighbors. During adaptation, 20% of the tokens are sampled uniformly without replacement and their backbone embeddings are replaced by a learned mask token initialized from $\mathcal{N}(0, 0.02^2)$. All three controls reconstruct expression profiles at masked positions using the same cosine reconstruction loss as SPARSE; the Laplacian control adds the regularizer below.

The Linear control is a point-wise residual MLP adapter rather than a single affine layer. It applies an input projection and LayerNorm, adds a two-layer ReLU coordinate projection, and processes every token independently with nine residual MLP blocks. Each block uses pre-normalization, an MLP ratio of 4, GELU, and zero dropout. Because there is no token-mixing operation, this control uses each token's coordinate but cannot aggregate information from neighboring tokens.

The Transformer control uses the same input and coordinate projections followed by six pre-LayerNorm Transformer blocks with eight attention heads and an MLP ratio of 4. Attention, projection, and MLP dropout are zero, and stochastic-depth rates increase linearly from 0 to 0.1. It

performs standard self-attention over the centered neighborhood but contains no coordinate-aware token merging, spatial RoPE, cross-resolution attention, or multi-scale aggregation.

Laplacian Reg uses exactly the same architecture as the Linear control and differs only in its training objective. A binary star graph connects the centered target to each of its 199 spatial neighbors, and the additional loss penalizes the squared distance between their ℓ_2-normalized hidden representations over edges whose endpoints are unmasked:

$$\mathcal{L}_{\text{Lap}} = \frac{1}{|\mathcal{E}|} \sum_{(0,j)\in\mathcal{E}} \|\bar{z}_0 - \bar{z}_j\|_2^2, \qquad \mathcal{L} = \mathcal{L}_{\text{MCM}} + 0.1\,\mathcal{L}_{\text{Lap}}. \tag{24}$$

This control therefore isolates generic local graph smoothness without changing the adapter architecture.

All three controls use AdamW with learning rate 10^{-4}, minimum learning rate 10^{-6}, weight decay 0.05, $\beta = (0.9, 0.999)$, and $\epsilon = 10^{-8}$. We use a global batch size of 64 (16 per GPU on four GPUs), 10 epochs of linear warmup followed by cosine annealing, BF16 automatic mixed precision, and gradient-norm clipping at 1.0. Each run is configured for 31 epochs and evaluated from the epoch-30 checkpoint; no downstream metric is used for checkpoint selection. For every control, the downstream representation is the final adapter LayerNorm output of the centered target token, with no pooling, and the reconstruction head is discarded.

E Additional Experimental Results

This section complements the main results with a qualitative lung analysis, the full Leiden-resolution sweep, region-level ablations, neighborhood-scale errors, and three-seed confidence intervals. The evaluation protocols are unchanged.

E.1 Disease-Relevant Analysis

We visualize fixed epoch-30 scGPT+SPARSE embeddings from the 10% test partition of the merged healthy and cancer lung sections using UMAP; the split is defined in Appendix B.6. Figure E1 shows partially separated but overlapping condition-associated regions: healthy cells lie mainly to the left, while cancer cells extend through the center and right. This is not a supervised classification result, and no quantitative main-paper claim depends on it.

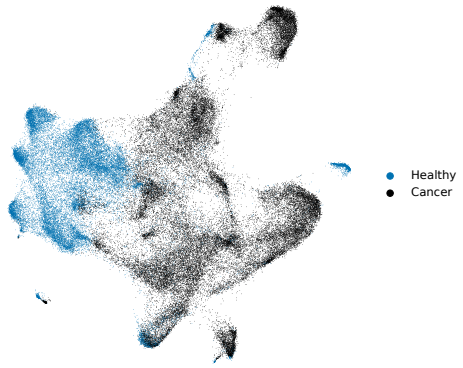

Figure E1: UMAP of fixed epoch-30 scGPT+SPARSE embeddings from the human-lung test partition. Each point is one test cell; color denotes the healthy or cancer section. The overlap limits this view to qualitative evidence.

E.2 Resolution Sensitivity

Spatial-domain identification requires a clustering-resolution parameter. The main paper uses 0.8 as a common primary setting for every method and dataset rather than selecting a method-specific

test resolution. To assess sensitivity to this shared setting, we evaluate every method at Leiden resolutions from 0.1 to 1.0 in increments of 0.1. Figure E2 reports mean NMI and ARI over random seeds 0, 42, and 123.

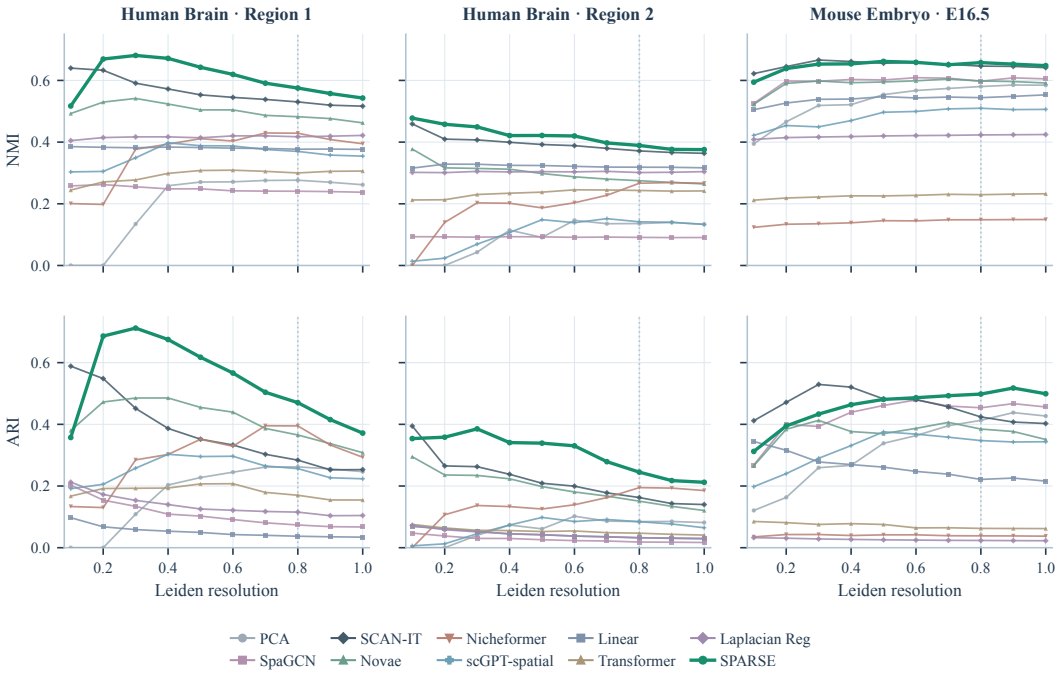

Figure E2: Spatial-domain sensitivity to Leiden resolution on the held-out human-brain regions and mouse-embryo final stage. Columns show Human Brain Region 1, Human Brain Region 2, and Mouse Embryo E16.5; rows show NMI and ARI. Curves are the three-seed means for each method. The dotted vertical line marks the common primary setting of 0.8, used for every method without per-method tuning.

The sweep exposes the expected dependence of clustering scores on Leiden resolution and allows every method to be inspected under the same range of operating points. It is not used to choose a separate test resolution for any method: all primary values remain those obtained at the common resolution 0.8.

E.3 Full Ablation Results

The main-paper ablation table reports NMI and ARI averaged across the two held-out human-brain regions. Table E1 provides the complete seed-42 result for each region and the two-region average. The variant names correspond to the main-paper ablation table.

Table E1: Complete seed-42 ablation on the two held-out human-brain regions at Leiden resolution 0.8. NMI and ARI are reported for each region and as their arithmetic two-region average; higher is better. "w/o Cross" removes cross-scale attention, and the other row names follow the components defined in Section 3 of the main paper.

	Human Brain Region 1		Human Brain Region 2		Two-Region Average	
Model variant	**NMI % ↑**	**ARI % ↑**	**NMI % ↑**	**ARI % ↑**	**Mean NMI % ↑**	**Mean ARI % ↑**
w/o spatial neighborhood	43.64	32.04	31.27	18.99	37.46	25.52
w/o HSF	52.76	48.01	33.32	20.38	43.04	34.20
w/o rCTM	48.48	33.00	39.76	31.35	44.12	32.18
w/o Cross	52.00	41.58	37.32	24.03	44.66	32.81
w/o token score	49.39	41.32	37.23	22.26	43.31	31.79
Full scGPT+SPARSE	**57.87**	**47.26**	**38.46**	**24.05**	**48.16**	**35.65**

The full model gives the highest two-region mean NMI and ARI. Removing centered spatial-neighborhood construction causes the largest joint reduction, while every remaining component removal lowers both two-region averages. The region-level columns additionally show that the magnitude of each ablation effect depends on the evaluated coordinate-defined region, motivating the averaged summary reported in the main paper.

E.4 Neighborhood Scales

The main paper reports Mean MAE for neighborhood-composition prediction. Tables E5, E6, and E7 give the complete results at average neighborhood sizes of 100, 200, and 500 cells for probe learning rates 0.05, 0.01, and 0.005, respectively. At each scale, a linear probe predicts the local cell-type composition from fixed embeddings. Each entry combines the three-seed mean, sample standard deviation, and 95% confidence interval. The Linear, Transformer, and Laplacian Reg rows use the retrained, objective-matched adapters described in Appendix D.4, with each epoch-30 checkpoint evaluated using its corresponding downstream seed.

Under the primary probe setting in Table E5, several models exhibit partially convergent performance, with some entries coinciding at the reported precision. To ensure that the comparison is not driven by a single probe learning rate, we repeated the complete three-seed evaluation with learning rates of 0.01 and 0.005. All other settings, including the embeddings, data splits, neighborhood targets, probe architecture, batch size, and training duration, were held fixed. These supplementary experiments assess learning-rate sensitivity and are not used to select a favorable probe setting.

SPARSE achieves the lowest mean MAE at all three neighborhood scales under each evaluated learning rate. At 0.01, its Mean MAE is 2.751%, compared with 3.066% for the strongest competing mean; at 0.005, the corresponding values are 3.003% and 3.318%. The lower learning rates produce more differentiated results among the compared embeddings while preserving the ordering toward lower SPARSE error. Absolute MAE values nevertheless vary with probe optimization, so the results should be interpreted jointly as a learning-rate sensitivity analysis. The corresponding scale-specific confidence intervals are reported in Appendix E.5.

E.5 Confidence Intervals

The main-paper tables report three-seed means and sample standard deviations, including the semantic-retention results in Table 3. Tables E2–E4 add the corresponding 95% confidence intervals. The three neighborhood-composition tables combine mean ± sample standard deviation with the interval in each cell, avoiding duplicate tables for the same runs. For each metric, we compute a two-sided 95% confidence interval using Student's t distribution:

$$\bar{x} \pm t_{0.975,2}\frac{s}{\sqrt{3}}, \qquad t_{0.975,2} = 4.303, \tag{25}$$

where s is the sample standard deviation across the three seeds. In the neighborhood-composition tables, the first line is mean ± sample standard deviation and the bracketed second line is the 95% confidence interval; the other tables show mean and interval. The intervals are untruncated and

therefore may extend slightly beyond the natural range of a metric when the between-seed variance is large. The single-seed ablation study is not included because a between-run confidence interval cannot be estimated from one seed.

Because each interval is estimated from only three runs, these intervals summarize observed seed variability and should not be interpreted as formal significance tests. They restate the same underlying runs as the mean ± standard deviation tables in the main paper; no additional checkpoint or downstream-model selection is performed.

Table E2: Spatial-domain identification on two held-out human-brain regions and transfer to the held-out final mouse-embryo stage. Each cell reports the percentage mean and bracketed 95% confidence interval over seeds 0, 42, and 123; higher is better. In the "No Adapt" column, ✓ means no task-specific representation fitting on the evaluation region or stage; ✗ means fitting to its unlabeled inputs. The human-brain regions are not known to be independent physical sections.

		Held-out Spatial Evaluation				Held-out Stage Transfer	
		Human Brain Region 1		Human Brain Region 2		E16.5 Stage Evaluation	
Methods	**No Adapt**	**NMI % ↑**	**ARI % ↑**	**NMI % ↑**	**ARI % ↑**	**NMI % ↑**	**ARI % ↑**
Non-spatial							
PCA	✗	27.65 [24.25, 31.05]	26.20 [20.31, 32.08]	13.59 [10.44, 16.73]	8.47 [6.12, 10.83]	58.02 [57.18, 58.87]	41.29 [40.39, 42.18]
Spatial Methods							
SpaGCN	✗	24.07 [23.78, 24.36]	7.37 [6.60, 8.14]	9.11 [8.95, 9.26]	1.83 [1.06, 2.60]	59.74 [51.40, 68.08]	45.37 [35.82, 54.93]
SCAN-IT	✗	53.03 [51.63, 54.43]	28.39 [26.62, 30.15]	37.15 [32.74, 41.55]	16.26 [12.60, 19.91]	64.64 [62.62, 66.65]	42.40 [35.59, 49.20]
Spatial Foundation Models							
Novae	✓	48.23 [42.79, 53.67]	36.55 [31.92, 41.18]	27.49 [24.47, 30.51]	15.09 [12.44, 17.74]	59.81 [58.00, 61.62]	38.47 [31.65, 45.29]
Nicheformer	✓	42.88 [42.31, 43.44]	39.48 [38.77, 40.19]	26.66 [26.04, 27.27]	19.48 [18.65, 20.30]	14.82 [14.77, 14.88]	3.84 [3.54, 4.14]
scGPT-spatial	✓	36.98 [34.87, 39.09]	25.65 [22.11, 29.18]	14.16 [12.88, 15.45]	8.43 [8.22, 8.64]	50.97 [48.64, 53.29]	34.73 [32.38, 37.08]
Lightweight Adapters							
scGPT + Linear	✓	37.10 [36.59, 37.60]	3.65 [3.04, 4.25]	32.48 [32.24, 32.72]	3.42 [3.11, 3.73]	52.95 [49.31, 56.59]	21.25 [13.40, 29.09]
+ Transformer	✓	41.60 [34.12, 49.08]	16.29 [10.70, 21.89]	28.60 [19.42, 37.78]	6.12 [2.65, 9.58]	19.82 [10.76, 28.88]	6.93 [1.48, 12.38]
+ Laplacian Reg	✓	37.85 [27.78, 47.92]	5.81 [4.76, 6.86]	31.44 [30.22, 32.67]	3.33 [2.79, 3.86]	41.04 [38.35, 43.73]	7.79 [6.06, 9.53]
Ours							
scGPT + SPARSE	✓	**57.53 [56.09, 58.98]**	**47.03 [45.39, 48.67]**	**38.92 [31.26, 46.58]**	**24.50 [13.03, 35.96]**	**65.77 [63.36, 68.18]**	**49.81 [46.86, 52.77]**

Across the six spatial-domain metrics in Table E2, SPARSE has the highest sample mean on both human-brain regions and on the held-out final mouse-embryo stage. The intervals are comparatively narrow on Region 1 but wider on Region 2, particularly for ARI, showing that the difficulty of the held-out spatial sample also affects between-seed variability. Several intervals overlap with those of competing methods, especially on Region 2 and the mouse-embryo task; accordingly, the table supports consistency of the observed mean improvements rather than a claim of statistical significance from only three runs.

The strongest competing method differs across datasets and metrics. Relative to the best competing sample mean in each column, SPARSE improves NMI/ARI by 4.50/7.55 percentage points on Human Brain Region 1, by 1.77/5.02 points on Human Brain Region 2, and by 1.13/4.44 points on the final mouse-embryo stage. The larger ARI margins indicate that the gain is particularly visible in agreement between the inferred domains and the reference partition, whereas the smaller NMI margins on Region 2 and the embryo task reflect closer competition under those protocols.

Table E3: Cell-type separability on the held-out mouse-heart slice. Each cell reports the percentage mean and bracketed 95% confidence interval over seeds 0, 42, and 123; higher is better. "Frozen" denotes the unadapted transcriptomic backbone, and both rows use the same within-slice reference/query label-transfer protocol defined in Appendix C.1.

Model		Mouse Heart Annotation	
		Accuracy % ↑	F1 Weighted % ↑
scGPT	*Frozen*	83.00 [82.86, 83.13]	77.63 [77.30, 77.97]
	+SPARSE	**83.91** [**83.42**, **84.41**]	**78.79** [**77.28**, **80.31**]
scFoundation	*Frozen*	84.28 [83.89, 84.67]	79.11 [78.73, 79.49]
	+SPARSE	**84.53** [**84.39**, **84.66**]	**79.14** [**78.72**, **79.57**]
scBERT	*Frozen*	65.73 [65.22, 66.24]	55.80 [55.40, 56.21]
	+SPARSE	**75.40** [**69.88**, **80.92**]	**69.43** [**62.74**, **76.12**]
Geneformer	*Frozen*	83.31 [82.94, 83.68]	77.58 [76.99, 78.16]
	+SPARSE	**84.76** [**84.02**, **85.51**]	**79.85** [**78.38**, **81.31**]

Table E4: Multi-scale utility on the held-out mouse whole-brain section. Each cell reports the mean and bracketed 95% confidence interval over seeds 0, 42, and 123. Classification metrics are percentages and higher is better; Mean MAE is a percentage and lower is better. In the "No Adapt" column, ✓ means no task-specific representation fitting on the evaluation section; ✗ means fitting to its unlabeled inputs. Downstream reference sets and probes are fitted only after the representation is fixed.

		Cross-slice Evaluation on Mouse Whole Brain				
		Cell-level		Niche-level	Tissue-level	
Methods	**No Adapt**	**Accuracy % ↑**	**F1 Weighted % ↑**	**Mean MAE % ↓**	**Accuracy % ↑**	**F1 Weighted % ↑**
Non-spatial						
PCA	✗	20.99 [19.62, 22.37]	15.34 [13.33, 17.34]	3.777 [3.505, 4.050]	80.05 [79.16, 80.95]	79.41 [78.52, 80.31]
Spatial Foundation Models						
Novae	✓	49.23 [48.43, 50.03]	43.38 [42.74, 44.02]	2.628 [2.539, 2.718]	84.35 [83.70, 84.99]	83.30 [82.79, 83.81]
Nicheformer	✓	56.72 [53.84, 59.60]	52.12 [51.07, 53.16]	2.952 [2.826, 3.079]	43.19 [42.96, 43.41]	38.99 [37.38, 40.61]
scGPT-spatial	✓	87.69 [87.38, 88.01]	87.03 [86.69, 87.36]	4.057 [3.909, 4.205]	72.22 [71.67, 72.77]	70.58 [69.95, 71.22]
Lightweight Adapters						
scGPT + Linear	✓	50.26 [47.04, 53.47]	44.01 [40.77, 47.24]	2.952 [2.826, 3.079]	84.86 [82.50, 87.23]	82.45 [78.45, 86.45]
+ Transformer	✓	65.53 [53.87, 77.20]	61.58 [49.49, 73.67]	2.941 [2.772, 3.109]	79.00 [61.50, 96.50]	76.47 [56.50, 96.44]
+ Laplacian Reg	✓	41.79 [41.07, 42.50]	36.43 [35.52, 37.34]	2.952 [2.826, 3.079]	83.82 [82.33, 85.31]	80.16 [77.24, 83.07]
Ours						
scGPT + SPARSE	✓	**87.68** [**84.34**, **91.01**]	**87.09** [**83.57**, **90.61**]	**2.272** [**2.048**, **2.497**]	**88.12** [**87.22**, **89.01**]	**87.35** [**86.36**, **88.34**]

Table E3 provides a complementary view of semantic retention. The adapted embedding has a higher sample mean than its frozen backbone for every evaluated backbone and metric. The largest change occurs for scBERT, whereas scFoundation remains close to its unadapted representation. This pattern indicates that the spatial adapter does not systematically trade away cell-type-discriminative structure, while also showing that the magnitude of the change depends on the expression-derived backbone.

The three task levels in Table E4 evaluate different information in the same fixed representation. SPARSE obtains the lowest neighborhood-composition MAE and the highest tissue-region accuracy and weighted F1. At the cell level, its mean performance is essentially matched by scGPT-spatial,

differing by only 0.01 percentage points in accuracy and 0.06 points in weighted F1. Thus, the principal advantage in this experiment is not an isolated cell-level gain, but the ability of one transferred embedding to remain useful across cell, neighborhood, and tissue-region readouts.

Table E5 resolves the niche-level Mean MAE in Table E4 into the three evaluated neighborhood scales. Its mean, sample standard deviation, and confidence interval all summarize the same three seed-level measurements; the intervals follow Equation 25.

Table E5: Neighborhood-composition prediction at three spatial scales on the held-out mouse whole-brain section using a probe learning rate of 0.05. Each cell reports mean MAE $\pm$ sample standard deviation (%) above the bracketed 95% confidence interval over seeds 0, 42, and 123; lower is better.

	Average Neighborhood Size			
Methods	**MAE@100 % ↓**	**MAE@200 % ↓**	**MAE@500 % ↓**	**Mean MAE % ↓**
Non-spatial				
PCA	3.762 ± 0.114 [3.479, 4.045]	3.774 ± 0.085 [3.563, 3.984]	3.796 ± 0.172 [3.368, 4.223]	3.777 ± 0.110 [3.505, 4.050]
Spatial Foundation Models				
Novae	2.707 ± 0.047 [2.592, 2.823]	2.683 ± 0.044 [2.573, 2.794]	2.495 ± 0.018 [2.450, 2.539]	2.628 ± 0.036 [2.539, 2.718]
Nicheformer	3.385 ± 0.047 [3.268, 3.503]	2.993 ± 0.061 [2.842, 3.145]	2.479 ± 0.052 [2.348, 2.609]	2.952 ± 0.051 [2.826, 3.079]
scGPT-spatial	4.389 ± 0.033 [4.305, 4.472]	4.232 ± 0.178 [3.791, 4.673]	3.551 ± 0.042 [3.446, 3.656]	4.057 ± 0.060 [3.909, 4.205]
Lightweight Adapters				
scGPT + Linear	3.385 ± 0.047 [3.268, 3.503]	2.993 ± 0.061 [2.842, 3.145]	2.479 ± 0.052 [2.348, 2.609]	2.952 ± 0.051 [2.826, 3.079]
+ Transformer	3.385 ± 0.047 [3.268, 3.503]	2.993 ± 0.061 [2.842, 3.145]	2.443 ± 0.113 [2.162, 2.724]	2.941 ± 0.068 [2.772, 3.109]
+ Laplacian Reg	3.385 ± 0.047 [3.268, 3.503]	2.993 ± 0.061 [2.842, 3.145]	2.479 ± 0.052 [2.348, 2.609]	2.952 ± 0.051 [2.826, 3.079]
Ours				
scGPT + SPARSE	**2.620 ± 0.084** **[2.410, 2.830]**	**2.325 ± 0.138** **[1.982, 2.669]**	**1.872 ± 0.055** **[1.734, 2.009]**	**2.272 ± 0.090** **[2.048, 2.497]**

SPARSE has the lowest mean MAE at each evaluated scale: 2.620%, 2.325%, and 1.872% for average neighborhood sizes of 100, 200, and 500 cells, respectively. The reduction relative to the strongest competing mean is modest at 100 cells and becomes larger at the two broader scales. This trend is consistent with the intended role of the hierarchical adapter: local information is retained at fine resolution, while coarser token representations provide additional context for composition prediction over broader neighborhoods.

The confidence intervals should be read together with this mean trend. Some intervals overlap because they are estimated from only three seeds, so Table E5 is not used to assert pairwise statistical significance. Instead, it shows that the same ordering toward lower SPARSE mean error is observed at all three predefined neighborhood scales, without selecting a favorable scale after evaluation.

Tables E6 and E7 give the corresponding mean, standard-deviation, and confidence-interval summaries for the two learning-rate sensitivity runs.

Table E6: Neighborhood-composition MAE at probe learning rate 0.01: mean ± sample standard deviation (%) and 95% confidence interval over seeds 0, 42, and 123; lower is better.

Methods	Average Neighborhood Size MAE@100 % ↓	MAE@200 % ↓	MAE@500 % ↓	Mean MAE % ↓
Non-spatial				
PCA	3.713 ± 0.079 [3.517, 3.909]	3.667 ± 0.129 [3.346, 3.988]	3.557 ± 0.206 [3.046, 4.069]	3.646 ± 0.136 [3.309, 3.983]
Spatial Foundation Models				
Novae	4.093 ± 0.043 [3.986, 4.200]	3.958 ± 0.038 [3.863, 4.053]	3.479 ± 0.026 [3.414, 3.543]	3.843 ± 0.034 [3.758, 3.928]
Nicheformer	4.172 ± 0.434 [3.093, 5.251]	3.386 ± 0.662 [1.741, 5.031]	2.479 ± 0.052 [2.348, 2.609]	3.346 ± 0.368 [2.431, 4.260]
scGPT-spatial	4.819 ± 0.047 [4.701, 4.936]	4.524 ± 0.076 [4.334, 4.713]	3.960 ± 0.172 [3.531, 4.388]	4.434 ± 0.097 [4.192, 4.676]
Lightweight Adapters				
scGPT + Linear	4.289 ± 0.046 [4.175, 4.403]	3.565 ± 0.081 [3.365, 3.766]	2.754 ± 0.226 [2.193, 3.314]	3.536 ± 0.090 [3.312, 3.760]
+ Transformer	3.679 ± 0.536 [2.347, 5.011]	3.074 ± 0.610 [1.558, 4.589]	2.447 ± 0.554 [1.071, 3.823]	3.066 ± 0.562 [1.670, 4.462]
+ Laplacian Reg	4.227 ± 0.197 [3.737, 4.717]	3.649 ± 0.059 [3.501, 3.797]	3.092 ± 0.172 [2.664, 3.520]	3.656 ± 0.079 [3.459, 3.852]
Ours				
scGPT + SPARSE	**3.090 ± 0.072 [2.912, 3.268]**	**2.816 ± 0.084 [2.608, 3.025]**	**2.348 ± 0.050 [2.224, 2.472]**	**2.751 ± 0.065 [2.590, 2.913]**

Table E7: Neighborhood-composition MAE at probe learning rate 0.005: mean ± sample standard deviation (%) and 95% confidence interval over seeds 0, 42, and 123; lower is better.

Methods	Average Neighborhood Size MAE@100 % ↓	MAE@200 % ↓	MAE@500 % ↓	Mean MAE % ↓
Non-spatial				
PCA	4.139 ± 0.105 [3.878, 4.401]	4.183 ± 0.113 [3.903, 4.463]	4.273 ± 0.072 [4.093, 4.452]	4.199 ± 0.074 [4.014, 4.384]
Spatial Foundation Models				
Novae	5.230 ± 0.038 [5.134, 5.325]	5.110 ± 0.038 [5.017, 5.204]	4.830 ± 0.026 [4.765, 4.895]	5.057 ± 0.034 [4.972, 5.141]
Nicheformer	4.551 ± 0.088 [4.332, 4.770]	4.206 ± 0.193 [3.726, 4.686]	3.084 ± 0.516 [1.803, 4.365]	3.947 ± 0.197 [3.458, 4.435]
scGPT-spatial	4.978 ± 0.103 [4.723, 5.232]	4.656 ± 0.120 [4.357, 4.955]	4.071 ± 0.100 [3.822, 4.320]	4.568 ± 0.107 [4.303, 4.833]
Lightweight Adapters				
scGPT + Linear	4.266 ± 0.068 [4.097, 4.434]	3.887 ± 0.147 [3.522, 4.253]	3.056 ± 0.084 [2.847, 3.266]	3.736 ± 0.098 [3.494, 3.979]
+ Transformer	3.867 ± 0.447 [2.757, 4.976]	3.400 ± 0.556 [2.020, 4.781]	2.687 ± 0.567 [1.279, 4.094]	3.318 ± 0.522 [2.022, 4.614]
+ Laplacian Reg	4.374 ± 0.016 [4.335, 4.414]	3.919 ± 0.074 [3.734, 4.104]	3.110 ± 0.038 [3.016, 3.203]	3.801 ± 0.036 [3.713, 3.889]
Ours				
scGPT + SPARSE	**3.344 ± 0.072 [3.167, 3.522]**	**3.063 ± 0.088 [2.846, 3.281]**	**2.602 ± 0.080 [2.404, 2.800]**	**3.003 ± 0.077 [2.813, 3.194]**

SPARSE retains the lowest sample mean at every neighborhood scale for both alternative learning rates. The Mean MAE confidence interval is [2.590, 2.913] at 0.01 and [2.813, 3.194] at 0.005. These intervals overlap those of the strongest competing Transformer adapter, so the three-seed sensitivity analysis supports stability of the mean ordering but is not used to claim pairwise statistical significance.