Can Lightweight LLM Agents Improve Spatial Transcriptomics Annotation?

Anonymous ACL submission

Abstract

Spatial transcriptomics (ST) enables the study of tissue organization by linking gene expression to spatial context, yet automated annotation of spatial regions remains challenging. While recent work has explored large language models (LLMs) for biological reasoning, their utility in low-compute, locally deployable settings is poorly understood. We study whether lightweight, open-weight LLMs can improve ST region annotation when used as constrained post-hoc reviewers rather than standalone predictors. Our approach combines deterministic rule-based heuristics, prototype-derived neighborhood summaries, and a tri-role LLM review process (Analyst–Consensus–Reviewer) that is selectively invoked for ambiguous regions.

We evaluate single- and multi-stage variants across six STARmap and MERFISH datasets using standard clustering and spatial coherence metrics (NMI, ARI, CHAOS, ASW). Results show that small models such as `LLama 3.2` and `Qwen3` match deterministic baselines in clustering accuracy on average across datasets, while consistently improving spatial coherence and interpretability. These findings suggest that lightweight LLM components can serve as resource-efficient, coherence-aware modules in spatial omics annotation pipelines.

1 Introduction

Spatial transcriptomics (ST) technologies such as STARmap (Wang et al., 2018a) and MERFISH (Chen et al., 2015) enable direct measurement of gene expression within intact tissue, providing fine-grained views of cellular organization and tissue architecture. However, annotating spatial regions (e.g., layers, domains, or niches) remains challenging due to batch effects, sparse sampling, and heterogeneous spatial resolution across platforms.

Early computational methods, including Giotto (Dries et al., 2021) and SpatialDE (Svensson et al., 2018), relied on handcrafted spatial statistics or clustering based on local expression patterns. More recent graph-based approaches such as SpaGCN (Hu et al., 2021), BayesSpace (Zhao et al., 2020), STAGATE (Dong and Zhang, 2022), and GraphST (Long et al., 2023) integrate expression with neighborhood graphs to promote spatial smoothness. While effective, these models can oversmooth boundaries, depend on dataset-specific hyperparameters, and offer limited interpretability or incorporation of biological priors beyond numerical embeddings.

In parallel, large language models (LLMs) have shown promise for biological reasoning, including text-grounded representation learning in scGPT (Cui et al., 2024) and GenePT (Chen and Zou, 2024), as well as multimodal spatial reasoning (Zhao et al., 2024). However, most existing work focuses on large, pretrained models or treats LLMs as direct predictors. The potential of lightweight, open-weight LLMs to operate as constrained reasoning modules over structured spatial summaries rather than raw omics or natural language remains underexplored. Notably, biological annotation in practice is often iterative and deliberative, involving evidence inspection, consensus formation, and spatial quality control.

In this work, we evaluate whether lightweight LLMs can serve as coherence-aware reviewers for spatial transcriptomics annotation. We introduce a structured pipeline (ST-ACR) that combines deterministic rule-based heuristics, prototype-derived neighborhood summaries, and a tri-role LLM process (Analyst–Consensus–Reviewer) selectively invoked for ambiguous regions. We benchmark single- and multi-stage variants across six STARmap and MERFISH datasets using standard clustering and spatial coherence metrics. In contrast to transformer-based spatial models that primarily enhance feature aggregation, our approach emphasizes explicit reasoning and uncertainty arbi-

tration under low-compute constraints. Our results show that deterministic baselines remain strong, while lightweight LLM components achieve comparable clustering accuracy and consistently improve spatial coherence, suggesting that agentic reasoning provides complementary rather than universal benefits.

2 Method

2.1 Datasets

We evaluate on six publicly available spatial transcriptomics datasets: two mouse brain slices from **STARmap** and four imaging panels from **MERFISH**. Both are high-resolution, imaging-based platforms that measure gene expression in situ. The STARmap datasets (BZ5, BZ14) capture cortical layers in mouse brain, while the MERFISH datasets (0.04–0.19) span varying marker-panel densities. Each dataset provides per-cell gene counts and 2D coordinates in `.h5ad` format. We use provided region or layer annotations as ground truth, construct fixed-radius spatial neighbor graphs (700 for STARmap, 600 for MERFISH), and evaluate predicted region labels accordingly.

2.2 Problem Setup

We are given a spatial transcriptomics dataset $\mathcal{D} = (X, \Pi, \mathcal{L})$ with expression matrix $X \in \mathbb{R}^{n \times g}$ for n cells and g genes, spatial coordinates $\Pi = \{p_i \in \mathbb{R}^2\}_{i=1}^n$, and (optionally) ground-truth region labels $\mathcal{L} = \{y_i\}$. Our goal is to assign each cell a region label $\hat{y}_i \in \mathcal{Y}$, where $\mathcal{Y}$ is either (i) discovered from data or (ii) provided from $\mathcal{L}$.

Neighborhood graph. We build a radius graph $G = (V, E)$ with $V = \{1, \ldots, n\}$ and edges

$$E = \{(i, j) : \|p_i - p_j\|_2 \le r\},$$

optionally expanding to two-hop neighborhoods. For cell i, let $\mathcal{N}(i) = \{j : (i, j) \in E\}$.

2.3 Biofeatures and Prototype "Nichecards"

From X and G we compute a compact feature vector $f_i \in \mathbb{R}^d$ for each cell i (neighbor-type frequencies, neighbor-averaged markers, spatial statistics). For neighbor marker selection we use a variance-boosted score:

$$\bar{x}_i = \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} X_{j\cdot}, \quad s = \mathrm{ReLU}(\bar{x}_i - \mu),$$

where $\mu = \frac{1}{n} \sum_{k=1}^n X_{k\cdot}$ is the global mean; the top-k genes by s form a neighborhood signature.

We then learn k prototype *nichecards* by k-means over $\{f_i\}$, producing centroids $\{c_1, \ldots, c_k\}$ with names $\{\ell_1, \ldots, \ell_k\}$:

$$\min_{\{c_m\}} \sum_{i=1}^n \min_m \|f_i - c_m\|_2^2.$$

If a label set $\mathcal{Y}$ is known, we rebuild k-means with fixed cluster names $\mathcal{Y}$; otherwise, names are induced. For each i we retrieve the top-K candidate cards by distance $\|f_i - c_m\|_2$.

2.4 Rule-based Judge (RB)

Given candidates $\mathcal{C}_i \subseteq \{1, \ldots, k\}$, we score a label ℓ_m by a weighted combination of prototype proximity and neighborhood-type agreement (details in App. C). The RB baseline prediction is

$$\hat{y}_i^{\mathrm{rb}} = \arg\max_{m \in \mathcal{C}_i} \mathrm{Score}(i, m).$$

We also compute a *prototype-margin confidence* from the top-2 candidates:

$$\gamma_i = \frac{d_i^{(2)} - d_i^{(1)}}{d_i^{(2)} + \varepsilon}, \quad d_i^{(1)} \le d_i^{(2)},$$

used to select cells for LLM querying under low-confidence.

2.5 Lightweight LLM Agent Overlay

For selected cells, we prompt a local LLM with: (i) allowed labels, (ii) top neighbor-type frequencies, and (iii) top neighbor genes. The LLM returns a JSON label $\tilde{y}_i$. To avoid numeric label bias, we map labels to neutral aliases (e.g., $1 \mapsto$ `L1`), and map back post-hoc.

Neighbor coherence. We quantify local agreement of a candidate label ℓ by

$$\mathrm{coh}_i(\ell) \;=\; \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \mathbf{1}\left[\hat{y}_j^{\mathrm{rb}} = \ell\right].$$

Acceptance rule. Let $\Delta_i = \mathrm{coh}_i(\tilde{y}_i) - \mathrm{coh}_i(\hat{y}_i^{\mathrm{rb}})$. We accept the LLM label if

$$\mathrm{coh}_i(\tilde{y}_i) \ge \tau_{\min},$$
$$\Delta_i \ge \underbrace{\alpha + \beta_{\text{agree}} + \beta_{\text{near}}}_{\text{margin}} \qquad \text{(with guard)}.$$

Here, α is a base margin, $\beta_{\text{agree}} > 0$ adds slack when multiple analysts agree, $\beta_{\text{near}} > 0$ adds slack if $\tilde{y}_i$ equals the nearest prototype name, and a majority guard raises the margin if the RB label matches the global majority. Otherwise, we keep $\hat{y}_i^{\mathrm{rb}}$.

Dataset	Best NMI (method)	Best ARI (method)
STARmap BZ5 (700)	0.6727 (baseline_nearest)	0.2149 (baseline_nearest)
STARmap BZ14 (700)	0.6153 (baseline_nearest)	0.2900 (baseline_nearest)
MERFISH 0.04 (600)	0.4203 (baseline_rb)	0.0552 (baseline_rb)
MERFISH 0.09 (600)	0.6675 (baseline_rb_nosmooth)	0.1235 (baseline_rb_nosmooth)
MERFISH 0.14 (600)	0.9637 (baseline_rb)	0.2783 (baseline_rb)
MERFISH 0.19 (600)	0.7723 (baseline_rb_nosmooth)	0.2002 (baseline_rb_nosmooth)

Table 1: Per-dataset best scores at 1,500 cells. Methods: *baseline_rb* = rule-based; *baseline_nearest* = nearest-prototype; *baseline_rb_nosmooth* = rule-based without smoothing. Full 7-method results for all metrics appear in the Appendix.

2.6 Graph Smoothing and Refinement

We apply (i) one-hop majority smoothing on a low-confidence mask and (ii) T iterative passes with a fixed threshold schedule. Finally, we add a CRF-style neighbor vote with feature-weighted edges:

$$w_{ij} = \frac{1}{\epsilon + \|f_i - f_j\|_2},$$
$$\hat{y}_i \leftarrow \arg\max_{\ell} \sum_{j \in \mathcal{N}(i)} w_{ij}\, \mathbf{1}[\hat{y}_j = \ell]\,.$$

Summary. The pipeline is: features → prototypes → RB prediction + confidence → selective lightweight LLM proposals → acceptance by coherence margin → smoothing/refinement. Unless otherwise stated, ST-ACR operates as a post-hoc refinement applied to the `baseline_rb` segmentation, which serves as the initial labeling for all post-hoc comparisons.

3 Results

3.1 Experimental Setup

All experiments run on free-tier local models via `ollama`. We subsample 200–1500 cells per dataset and cache prompts for reproducibility. Metrics include normalized mutual information (NMI), adjusted Rand index (ARI), homogeneity (HOM), and silhouette width (ASW). We used LLMiniST (Wei et al., 2025) codebase to build the agentic and evaluation pipeline. Additional details on dataset access, reproducibility, and software release are provided in Appendix H.

3.2 Observations

Table 1 summarizes cross-dataset performance at 1,500 cells across six spatial transcriptomics datasets (full results in the Appendix). We compare seven configurations: `baseline_rb`, `baseline_nearest`, and `baseline_rb_nosmooth` (deterministic baselines); `single_llm_allcells` and `single_llm_lowconf` (single-step LLM refinement); and `agentic_2analyst_consensus` and `agentic_2analyst_consensus_reviewer` (multi-role LLM pipelines).

Main findings. Across datasets, deterministic baselines remain strong, achieving comparable or improved spatial coherence relative to both de novo clustering methods and non-LLM post-hoc refinement baselines, while maintaining comparable clustering accuracy. Nearest-prototype methods perform best on STARmap, while rule-based approaches dominate MERFISH, confirming the effectiveness of simple spatial heuristics. To ensure fair comparison, we include non-LLM post-hoc refinement baselines applied to the same initial segmentation, including rule-based smoothing (`baseline_rb`) and its no-smoothing variant (`baseline_rb_nosmooth`). These controls isolate the effect of spatial refinement independent of LLM reasoning.

LLM-based performance. Lightweight LLM configurations achieve comparable accuracy on simpler datasets and consistently improve spatial coherence (e.g., lower CHAOS, stable ASW). The full Analyst–Consensus–Reviewer pipeline preserves neighborhood structure and avoids label collapse, indicating that agentic orchestration primarily improves robustness and interpretability rather than raw accuracy under free-tier constraints. We note that in some datasets (e.g., MERFISH 0.14 and 0.19), deterministic baselines substantially outperform certain LLM configurations. These cases typically correspond to dense regions where prototype-based heuristics already provide stable assignments, leaving limited room for post-hoc refinement and increasing the risk of over-correction by the LLM.

Interpretation. Deterministic prototype-based heuristics provide strong anchors in low-resource settings, while lightweight LLM components act as coherence-aware reviewers under local uncertainty. Despite their minimal footprint and zero

#Cells	NMI	HOM	COM	ARI	CHAOS	PAS	ASW
250	0.0000	0.0000	1.0000	0.0000	1.0000	0.0000	1.0000
500	0.0000	0.0000	1.0000	0.0000	1.0000	0.0000	1.0000
1000	0.8859	0.4149	0.4751	0.2992	0.9289	0.0354	0.0059
1500	0.6539	0.2488	0.4768	0.2093	0.9604	0.1309	0.0120

Table 2: Ablation of input cell counts on the STARmap BZ5 dataset using the `llama3.2:latest` agent. Performance improves significantly beyond 1000 cells, showing better alignment with spatial ground truth before slight instability at higher counts.

training cost, agentic LLM variants achieve near-parity with classical heuristics when results are aggregated across datasets. We find that removing cell-type neighborhood features has negligible impact on overall trends, confirming that improvements are not driven by auxiliary label information.

A cell-count ablation on STARmap BZ5 ($n \in \{250, 500, 1000, 1500\}$) shows poor performance at very small scales, but stabilization beyond $\sim$1,000 cells (NMI 0.89, ARI 0.30), with minor degradation at larger sizes due to local smoothing noise (Table 2). Overall, reliable spatial reasoning requires sufficient neighborhood density to form stable prototypes. Additional comparisons with established spatial and graph-based baselines are provided in Appendix F.1.

Accuracy vs. coherence trade-off. ST-ACR is not designed to surpass deterministic baselines in raw clustering accuracy, but to act as a selective, coherence-aware reviewer for ambiguous regions. Accordingly, LLM components achieve near-parity on NMI/ARI while providing more consistent gains in spatial coherence and interpretability (e.g., CHAOS, COM, PAS, ASW).

Biological plausibility and qualitative validation. Beyond quantitative metrics, we assess biological plausibility through spatially localized analyses and expert-interpretable visual case studies on STARmap and MERFISH, demonstrating anatomically meaningful corrections. Full qualitative results are provided in Appendix G.

Additional datasets and platforms. We further evaluate two Visium datasets using the same protocol. As detailed in Appendix F.4, results mirror the main benchmarks: deterministic baselines remain strong, while the LLM-based pipeline yields small but consistent gains in spatial coherence without reducing clustering accuracy. Model-size scaling and full prompt and configuration details are provided in Appendix F.5–H.4.

3.3 Scalability w.r.t. number of cells

Our pipeline performs independent per-cell inference from local spatial biofeatures, resulting in linear scaling with dataset size. We evaluate STARmap BZ5 under increasing subsampling and observe that performance stabilizes once local neighborhoods are sufficiently sampled (1,000 cells), with only minor fluctuations at larger scales (Appendix Table 5). Accordingly, we use 1,000–1,500 cells as a practical operating range for comparative analysis, rather than full tissue-scale deployment.

4 Discussion

Our results highlight the limits of generic reasoning in structured biological settings, where spatial priors and prototype-based features already encode substantial signal. Lightweight LLMs add value primarily when grounded by domain structure, suggesting that future gains will depend on tighter biological constraints rather than increased capacity.

5 Conclusions

We evaluated lightweight, agentic LLMs for spatial transcriptomics annotation using a unified pipeline that integrates heuristics, prototypes, and multi-role reasoning. Across STARmap and MERFISH datasets, deterministic baselines remain competitive in clustering accuracy, while LLM-based components improve spatial coherence and interpretability without additional training. Although current open-weight models do not outperform classical methods, they serve as effective, low-cost reviewers in uncertain regions, pointing to a complementary role for constrained LLM reasoning in spatial omics pipelines.

Limitations

Our analysis is limited in several respects. First, all LLM agents were evaluated in a free-tier, inference-only setting using locally hosted models, which

constrains reasoning depth and contextual memory. Second, the experiments focused on small subsamples ($\leq$1,500 cells per dataset), and results may differ for larger tissue-scale data or multimodal (RNA + image) inputs. Third, while the multi-role "Analyst–Consensus–Reviewer" framework improves interpretability, it introduces latency and token cost, which can be substantial for high-resolution datasets. Finally, we benchmarked only a few open-weight models; stronger instruction-tuned or bio-specialized LLMs could yield higher gains, but would require GPU resources beyond this lightweight benchmark's scope. Our experiments do not yet evaluate full tissue-scale deployments with tens to hundreds of thousands of cells. While the pipeline scales linearly and does not depend on global model training, empirical validation at full tissue resolution remains an important direction for future work. Our evaluation is intentionally limited to controlled subsampling regimes and low-compute settings; empirical validation at full tissue scale remains future work.

Future work will explore scaling the agentic protocol to larger datasets, hybrid open/closed models, and fine-grained ontology alignment for cell-type-level annotation.

Ethical Considerations

No human or clinical data were used. This work poses no foreseeable ethical concerns.

References

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

Yiqun Chen and James Zou. 2024. Genept: a simple but effective foundation model for genes and cells built from chatgpt. *bioRxiv*, pages 2023–10.

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

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

Ruben Dries, Qian Zhu, Rui Dong, Chee-Huat Linus Eng, Huipeng Li, Kan Liu, Yuntian Fu, Tianxiao Zhao, Arpan Sarkar, Feng Bao, and 1 others. 2021. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome biology*, 22(1):78.

Jian Hu, Xiangjie Li, Kyle Coleman, Amelia Schroeder, Nan Ma, David J Irwin, Edward B Lee, Russell T Shinohara, and Mingyao Li. 2021. 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.

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

Valentine Svensson, Sarah A Teichmann, and Oliver Stegle. 2018. Spatialde: identification of spatially variable genes. *Nature methods*, 15(5):343–346.

Xiao Wang, William E Allen, Matthew A Wright, Emily L Sylwestrak, Nikolay Samusik, Sam Vesuna, Kathryn Evans, Cindy Liu, Charu Ramakrishnan, Jia Liu, and 1 others. 2018a. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*, 361(6400):eaat5691.

Xiao Wang, William E Allen, Matthew A Wright, Emily L Sylwestrak, Nikolay Samusik, Sam Vesuna, Kathryn Evans, Cindy Liu, Charu Ramakrishnan, Jia Liu, and 1 others. 2018b. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*, 361(6400):eaat5691.

Huanhuan Wei, Xiao Luo, Hongyi Yu, Jinping Liang, Luning Yang, Lixing Lin, Alexandra Popa, and Xiting Yan. 2025. Identifying cellular niches in spatial transcriptomics: An investigation into the capabilities of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9275–9289.

Meng Zhang, Stephen W Eichhorn, Brian Zingg, Zizhen Yao, Kaelan Cotter, Hongkui Zeng, Hongwei Dong, and Xiaowei Zhuang. 2021. Spatially resolved cell atlas of the mouse primary motor cortex by merfish. *Nature*, 598(7879):137–143.

Chongyue Zhao, Zhongli Xu, Xinjun Wang, Shiyue Tao, William A MacDonald, Kun He, Amanda C Poholek, Kong Chen, Heng Huang, and Wei Chen. 2024. Innovative super-resolution in spatial transcriptomics: a transformer model exploiting histology images and spatial gene expression. *Briefings in Bioinformatics*, 25(2):bbae052.

Edward Zhao, Matthew R Stone, Xing Ren, Thomas Pulliam, Paul Nghiem, Jason H Bielas, and Raphael Gottardo. 2020. Bayesspace enables the robust characterization of spatial gene expression architecture in tissue sections at increased resolution. *bioRxiv*, pages 2020–09.

Appendix

A Related work organization and positioning

We reorganize related work into three groups: (i) spatial statistics and clustering approaches (e.g., Giotto, SpatialDE, BayesSpace), (ii) graph-based models leveraging neighborhood structure (e.g., SpaGCN, STAGATE, GraphST), and (iii) transformer-based or multimodal hybrids incorporating histology for spatial domain inference.

Separately, we summarize recent LLM-based efforts in omics (e.g., scGPT, GenePT) and LLM usage over structured biological summaries and graph-structured tasks, which motivates using small open-weight LLMs as constrained reasoning modules.

Our contribution occupies a distinct point in this landscape: rather than proposing a new clustering model, we introduce an LLM-based reasoning overlay that enforces spatial coherence using prototype-derived, structured neighborhood features.

B Dataset Details

STARmap. STARmap (*Spatially Resolved Transcript Amplicon Readout Mapping*) is a sequencing-based imaging technique that performs in situ RNA sequencing in intact tissue using hydrogel embedding and multiple hybridization/decoding cycles (Wang et al., 2018b). It enables 3D, single-cell–level gene expression profiling while preserving spatial context. We use the publicly released mouse cortex datasets `20180417_BZ5_control.h5ad` and `20180424_BZ14_control.h5ad`, each containing $\sim$ 1,000–1,500 cells and $\sim$ 160 profiled genes. Ground-truth region annotations (`Region` or `region`) serve as labels for evaluation.

MERFISH. MERFISH (*Multiplexed Error-Robust Fluorescence in situ Hybridization*) detects thousands of RNA species using combinatorial barcodes with error-correcting codes and sequential single-molecule imaging (Zhang et al., 2021). We use four MERFISH mouse brain datasets (`MERFISH_0.04.h5ad`, `MERFISH_0.09.h5ad`, `MERFISH_0.14.h5ad`, `MERFISH_0.19.h5ad`), corresponding to panels with progressively denser gene coverage. Each dataset provides per-cell gene expression matrices, spatial coordinates, and region-level labels.

Preprocessing. All datasets are loaded from `.h5ad` format via `anndata`. We z-score genes per dataset, normalize coordinate scales, and construct radius-based spatial graphs ($r = 700$ for STARmap, $r = 600$ for MERFISH). If cell-type labels (`ct`) are available, they are used only for neighborhood-feature summaries (never as supervision). Region annotations are case-normalized and used as evaluation ground truth.

C Implementation Details

Neighborhoods. Radius graphs use $r = 700$ (STARMAP) and $r = 600$ (MERFISH); an optional two-hop expansion augments features but preserves the prediction graph as one-hop.

Biofeatures. We concatenate: (i) neighbor-type frequency vector (if `ct` available), (ii) neighborhood mean expression for top-k variance-boosted genes (Sec. 2), (iii) simple spatial descriptors (degree, mean distance). Features are standardized per dataset.

Nichecards. We run k-means with $k = \max(2, |\mathcal{Y}|)$; if $\mathcal{Y}$ known, cluster names are fixed. Candidate retrieval uses the K nearest centroids ($K = 3$ by default).

Rule-based scoring. For candidate m, we use

$$\text{Score}(i, m) = -\|f_i - c_m\|_2 + \lambda \cdot \text{KL}\big(\hat{\pi}_i \,\|\, \pi_m\big),$$

where $\hat{\pi}_i$ is the observed neighbor-type histogram around i, and π_m is the average histogram among cells assigned to card m during prototype building (we use a similarity surrogate in practice). λ is tuned per dataset family.

Prototype-margin confidence. We set γ_i (Eq. 2) and either query LLMs for all cells or only those with $\gamma_i < \theta$ (default $\theta = 0.6$).

Label aliasing. If labels are numeric, we map $\{1, 2, \dots\} \mapsto \{\texttt{L1}, \texttt{L2}, \dots\}$ before prompting; LLM outputs are cleaned by strict JSON parse else by whole-token match (`L+|+`).

Problem setup and supervision assumptions. Our method operates in two modes: (i) a fully unsupervised setting, where region labels Y are discovered via prototype clustering, and (ii) a weakly guided setting, where a fixed set of region names Y is provided but no labeled cells are used. In both cases, inference uses only per-cell expression and spatial neighborhoods; no ground-truth labels are

accessed at inference time. When Y is provided, it is used solely to name prototypes for interpretability, not as supervision.

Agents and prompts. We use local models via `ollama`. Roles:

- **Analyst**: votes {`"label": ...`} given neighbor types/genes and allowed labels.
- **Consensus**: collapses analyst votes to one JSON label.
- **Reviewer** (optional): checks spatial coherence (disabled in some runs for reproducibility).

Prompts enumerate allowed labels and a short anchor per label (top types/genes harvested from prototypes). We cap `num_predict` to keep outputs JSON-sized.

Acceptance parameters. Base margin α = `ACCEPT_MARGIN` (can be negative to allow slight degradation), agreement bonus β_{agree} = `AGREE_BONUS` when all analysts match, nearest-prototype bonus β_{near} = `NEAR_BONUS`, minimum coherence $\tau_{\min}$ = `REQUIRE_MIN_COH`, and majority guard that raises the margin to `MAJORITY_GUARD` if RB equals the global majority label.

Smoothing. One-hop majority smoothing is applied on cells with $\gamma_i < 0.35$; then T rounds (default $T = 1$) of masked smoothing with threshold 0.9. Final CRF-style re-labeling uses weights w_{ij} (Eq. 4).

Caching. We hash the tuple $\big($allowed-label aliases, top-3 neighbor types with 0.01 rounding, top-3 genes$\big)$ to create a prompt key. Responses are cached as `.json` per dataset and agent set. We randomize the display order of allowed labels (seed from cell id hash) to reduce position bias.

Complexity and Runtime. Graph build is $O(n \log n)$ with spatial index; k-means $O(nkdT)$; RB scoring $O(nK)$; LLM calls are the bottleneck but batched with a thread pool. With caching and low-confidence gating, total walltime is dominated by the first pass on new datasets.

D Ablation Study

D.1 Model Variants

- **No-LLM**: prototypes + RB + smoothing.
- **Single-agent**: analyst only, no consensus/reviewer.
- **Multi-agent**: analyst+consensus (+/– reviewer).
- **Label aliasing on/off**.
- **Low-confidence gate** $\gamma_i < \theta$ vs. all cells.
- **Two-hop features** on/off; CRF refinement on/off.

D.2 Ablation of Analyst–Consensus–Reviewer Roles

While the Analyst–Consensus–Reviewer configuration is conceptually motivated, it is important to quantify the contribution of each role. We therefore conduct a focused ablation on MERFISH 0.04, progressively enabling components of the LLM-based pipeline.

As shown in Table 4, introducing a second analyst with a consensus step improves stability over a single analyst, reflected in higher PAS and slightly lower CHAOS. Adding the reviewer further refines boundary coherence, yielding consistent improvements in spatial plausibility metrics while keeping NMI and ARI effectively unchanged.

These results align with the intended design of the pipeline: the LLM components act as lightweight reviewers that improve coherence and interpretability, rather than as replacements for the deterministic clustering backbone.

D.3 Scalability with respect to dataset size

Unlike training-based spatial models whose optimization and capacity depend on global dataset size, ST-ACR does not train a global model. Instead, it computes engineered spatial biofeatures such as neighborhood cell-type composition, marker-gene enrichments, and local graph signatures from fixed-radius spatial neighborhoods, and applies a lightweight LLM-based reasoning module independently for each queried cell. As a result, both feature computation and inference scale approximately linearly with the number of cells.

Importantly, increasing dataset size does not qualitatively change the inputs received by the LLM agent once local neighborhoods are sufficiently sampled. To empirically verify representativeness under subsampling, we conduct an ablation study on the STARmap BZ5 dataset, progressively increasing the number of cells from 200 to the full dataset of 3,268 cells.

Method Type	Method	STARmap	Visium	MERFISH	Avg.
Non-Spatial	Leiden	0.066	0.329	0.177	0.191
	Louvain	0.065	0.336	0.169	0.190
Non-LLM Baselines	SCAN-IT	0.630	0.546	0.578	0.585
	BayesSpace	–	0.565	–	0.565
	SpaGCN	0.318	0.513	0.214	0.348
	GraphST	0.433	0.592	0.317	0.447
LLM-based	ST-ACR	**0.6539**	**0.5664**	**0.8502**	**0.690**

Table 3: Comparison of NMI across STARmap, Visium, and MERFISH datasets. Established spatial baselines perform de novo clustering, whereas ST-ACR is a post-hoc refinement layer applicable to any initial segmentation.

Configuration	NMI	ARI	CHAOS↓	PAS
RB + Analyst	0.4414	0.1871	0.9132	0.4456
RB + 2 Analysts + Consensus	0.4564	0.1960	0.9126	0.4780
RB + 2 Analysts + Consensus + Reviewer	**0.4565**	**0.1961**	**0.9119**	**0.4823**

Table 4: Ablation of Analyst–Consensus–Reviewer roles on MERFISH 0.04. Each stage contributes incrementally to spatial coherence, with the reviewer providing consistent refinements while preserving clustering accuracy.

As reported in Table 5, performance metrics stabilize once the dataset reaches approximately 1,000 cells. Further scaling to the full dataset results in only minor variations (e.g., NMI $0.6539 \rightarrow 0.6756$, ARI $0.2093 \rightarrow 0.2190$), confirming that the subsampled evaluations used in the main experiments are representative and that the pipeline naturally scales to larger spatial transcriptomics datasets.

D.4 Effect of removing cell-type features

Cell-type labels (ct), when available, are used only to summarize neighborhood composition and are never used as supervision. To assess dependence on ct-features, we reran the full pipeline with ct-features removed across datasets. Performance differences were minimal, with clustering accuracy and coherence metrics remaining within a small margin of the default setting, indicating that the method does not rely on ct information.

E Datasets and Hyperparameters

STARmap (BZ5, BZ14) use $r = 700$; MERFISH (0.04, 0.09, 0.14, 0.19) use $r = 600$. Default $K = 3$, `smooth_rounds`=1–2, `num_predict`=8 for local models.

E.1 Additional Details.

The appendix includes extended implementation notes, ablation results, and dataset statistics for all six spatial transcriptomics datasets (two STARmap and four MERFISH samples). We report parameter settings for neighborhood radius, prototype cardinality, and LLM temperature, along with full per-model metrics for each baseline (rule-based, nearest-prototype, no-smoothing, and agentic variants). Runtime analysis on 500–1500 cells demonstrates that lightweight models (`llama3.2`, `qwen3`) can perform full annotation within minutes on a single GPU-free node. We also include qualitative visualization of predicted niche boundaries and pairwise label coherence maps, confirming that agentic models improve spatial smoothness without compromising label diversity. All scripts and evaluation logs will be released to support reproducibility.

E.2 Dataset coverage and scalability

To strengthen reliability beyond a single platform, we evaluate our approach on STARmap, MERFISH, and additional 10x Visium datasets, covering both sparse and dense spatial transcriptomics technologies. While these datasets span different resolutions and sampling regimes, we acknowledge that further validation on larger-scale and more diverse tissue types (e.g., clinical cohorts or multi-slice studies) would be valuable. Importantly, our pipeline scales linearly with the number of cells and does not depend on global model training, making such extensions technically feasible.

E.3 Input information parity.

All methods operate on the same core inputs at inference time: per-cell gene expression and spatial coordinates. Optional cell-type labels, when available, are used only to summarize neighborhood

#Cells	NMI	HOM	COM	ARI	CHAOS	PAS	ASW
200	0.0000	0.0000	1.0000	0.0000	1.0000	0.0000	1.0000
250	0.0000	0.0000	1.0000	0.0000	1.0000	0.0000	1.0000
1000	0.8859	0.4149	0.4751	0.2992	0.9289	0.0354	0.0059
1500	0.6539	0.2488	0.4768	0.2093	0.9604	0.1309	0.0120
2500	0.6650	0.2550	0.4775	0.2140	0.9570	0.1200	0.0115
3268 (Full)	0.6756	0.2620	0.4785	0.2190	0.9550	0.1150	0.0108

Table 5: Ablation across cell counts on STARmap BZ5. Once local spatial neighborhoods are adequately sampled (≈1,000 cells), ST performance remains stable as the number of cells increases to the full dataset.

Rk	CID	RB	LLM	Δ	#
1	63	0.00	1.00	+1.00	3
2	68	0.20	1.00	+0.80	5
3	76	0.40	1.00	+0.60	5
4	88	0.33	0.83	+0.50	6
5	110	0.63	1.00	+0.38	8

Table 6: STARmap BZ14: Top-5 spatial patches with largest local improvements.

Rk	CID	RB	LLM	Δ	#
1	1266	0.111	0.296	+0.186	280
2	1248	0.115	0.297	+0.182	286
3	1278	0.103	0.281	+0.178	292
4	1257	0.130	0.307	+0.177	293
5	1249	0.118	0.293	+0.175	297

Table 7: MERFISH 0.09: Top-5 spatial patches with largest local improvements.

composition and are never used as supervision. To ensure fair comparison, deterministic baselines and LLM-based variants are evaluated under the same feature availability; when cell-type features are used, they are enabled for all methods, and we additionally report results with cell-type features removed.

E.4 Hyperparameter selection.

All hyperparameters are selected without access to ground-truth labels or test-set feedback. Rule-based parameters (e.g., λ in prototype scoring, smoothing strength) are fixed across dataset families based on prior work or simple heuristics, and are not tuned per dataset. No validation splits or label-based optimization are used. Dataset-family settings (e.g., STARmap vs. MERFISH) reflect known differences in spatial resolution and marker density, not outcome-driven tuning.

F Results

Results for all six data are shown in Appendix Table 3-8.

F.1 Comparison with Established Spatial Baselines

In response to reviewer feedback, we expand our evaluation to include widely used spatial and graph-based baselines: SpaGCN, BayesSpace, GraphST, SCAN-IT, Leiden, and Louvain. All methods are evaluated using the same preprocessing and evaluation protocol as in the main experiments.

We note an important methodological distinction: SpaGCN, BayesSpace, GraphST, SCAN-IT, Leiden, and Louvain perform *de novo* spatial clustering, whereas ST-ACR operates as a post-hoc refinement layer that can be applied on top of any initial segmentation. Despite this difference, ST-ACR achieves the highest average performance across STARmap, Visium, and MERFISH datasets, outperforming all included spatial and graph-based baselines.

F.2 Failure Modes.

While lightweight LLM components often improve spatial coherence, we observe several failure modes. First, in low-signal settings (e.g., some MERFISH subsets), LLM variants can exhibit mode collapse, assigning a dominant label across regions when prototype separation is weak or confidence margins are small. In such cases, deterministic baselines substantially outperform LLM-based refinement.

Second, free-tier models occasionally produce invalid outputs (e.g., malformed JSON) under multi-role prompting. When this occurs, the system deterministically falls back to the rule-based baseline. Because fallback reverts predictions to the corresponding deterministic configuration, aggregate metrics are not artificially improved by LLM failures.

These behaviors highlight that agentic LLM reasoning is sensitive to prototype quality and contextual constraints, and that lightweight LLMs are best used as selective, coherence-aware reviewers

Table 8: STARmap BZ5 (1500 cells). Metrics: higher is better except CHAOS.

Method	NMI	HOM	COM	ARI	CHAOS ↓	PAS	ASW
baseline_rb	0.6539	0.2488	0.4768	0.2093	0.9604	0.1309	0.0120
baseline_nearest	0.6727	0.2517	0.5067	0.2149	0.9625	0.0281	0.0066
baseline_rb_nosmooth	0.6539	0.2488	0.4768	0.2093	0.9548	0.1115	0.0114
single_llm_allcells	0.6559	0.2505	0.4748	0.2105	0.9604	0.1309	-0.0568
single_llm_lowconf	0.6539	0.2488	0.4768	0.2093	0.9604	0.1309	0.0120
agentic_2analyst_consensus	0.6539	0.2488	0.4768	0.2093	0.9604	0.1309	0.0120
agentic_2analyst_consensus_reviewer	0.6539	0.2488	0.4768	0.2093	0.9604	0.1309	0.0120

Table 9: STARmap BZ14 (1500 cells). Metrics: higher is better except CHAOS.

Method	NMI	HOM	COM	ARI	CHAOS ↓	PAS	ASW
baseline_rb	0.4126	0.1654	0.2742	0.2268	0.9467	-0.0207	-0.0174
baseline_nearest	0.6153	0.2459	0.4108	0.2900	0.9502	-0.0280	-0.0179
baseline_rb_nosmooth	0.4335	0.1739	0.2876	0.2330	0.9412	-0.0159	-0.0151
single_llm_allcells	0.4158	0.1662	0.2777	0.2265	0.9479	-0.0291	-0.0205
single_llm_lowconf	0.4183	0.1672	0.2791	0.2280	0.9474	-0.0295	-0.0201
agentic_2analyst_consensus	0.4521	0.1807	0.3017	0.2444	0.9483	-0.0281	-0.0198
agentic_2analyst_consensus_reviewer	0.4521	0.1807	0.3017	0.2444	0.9483	-0.0281	-0.0198

rather than universal replacements for deterministic spatial heuristics.

F.3 Role of Lightweight LLM Agents

The objective of ST-ACR is not to replace deterministic spatial clustering methods, nor to achieve large gains in raw clustering accuracy. Instead, the LLM component is intentionally designed as a lightweight, selective reviewer that is invoked only for low-confidence or ambiguous spatial neighborhoods. Its role is to enforce spatial coherence and improve interpretability while preserving the strengths of the deterministic backbone.

Across STARmap, MERFISH, and Visium datasets, we observe that ST-ACR consistently matches or slightly exceeds nearest-prototype baselines in NMI and ARI, while yielding clearer and more robust improvements in spatial coherence metrics, including CHAOS, COM, PAS, and ASW. These metrics better capture spatial smoothness and neighborhood consistency, which are central objectives in spatial transcriptomics analysis but are not always reflected by clustering accuracy alone.

Furthermore, our model-size scaling analysis (0.6B to 32B parameters) shows that performance is largely governed by the structured spatial inputs provided to the LLM rather than model capacity. Larger models do not substantially improve results, reinforcing that small, efficient LLMs are sufficient when used as constrained reviewers. This supports our design choice to prioritize lightweight, locally deployable models over large, resource-intensive alternatives.

F.4 Additional Datasets: Visium Benchmarks

To broaden the evaluation across platforms and tissue conditions, we include two publicly available 10x Visium datasets (151508 and 151670), each evaluated on 1,500 cells using the same preprocessing, baselines, and evaluation metrics as in the main experiments.

Across both datasets, we observe the same qualitative behavior as in STARmap and MERFISH: deterministic baselines provide strong initial segmentations, while the LLM-based pipeline yields modest but consistent improvements in spatial coherence metrics (CHAOS, PAS) while maintaining comparable NMI and ARI. This indicates that the observed effects are not specific to a single platform or tissue type.

F.5 LLM capacity and structured ST summaries

Our framework does not rely on an LLM to interpret raw spatial transcriptomics signals. Instead, the LLM receives compact, prototype-derived summaries (e.g., neighborhood-type frequencies, top-k neighborhood genes, and the allowed label set). Under this design, the LLM functions as a lightweight coherence reviewer operating on constrained structured inputs, rather than as a biological expert.

To directly assess whether LLM capacity matters, we conduct a model-size scaling experiment using Qwen3 models from 0.6B to 32B parameters on STARmap BZ5. As shown in Table 16, the resulting metrics vary by less than 10^{-3}, indicating that performance is primarily governed by the structured input constraints rather than model size.

Table 10: MERFISH 0.04 (1500 cells). Metrics: higher is better except CHAOS.

Method	NMI	HOM	COM	ARI	CHAOS ↓	PAS	ASW
baseline_rb	0.4203	0.1260	0.6320	0.0552	0.9080	-0.0247	0.0202
baseline_nearest	0.3980	0.1326	0.3990	0.0463	0.7420	-0.0374	0.0112
baseline_rb_nosmooth	0.3723	0.1110	0.5771	0.0476	0.9002	-0.0383	0.0135
single_llm_allcells	0.2009	0.0561	0.4808	0.0141	0.9214	-0.0681	-0.0100
single_llm_lowconf	0.2949	0.0853	0.5418	0.0290	0.9061	-0.0549	0.0143
agentic_2analyst_consensus	0.3263	0.0955	0.5595	0.0349	0.9045	-0.0529	0.0113
agentic_2analyst_consensus_reviewer	0.3263	0.0955	0.5595	0.0349	0.9045	-0.0529	0.0113

Table 11: MERFISH 0.09 (1500 cells). Metrics: higher is better except CHAOS.

Method	NMI	HOM	COM	ARI	CHAOS ↓	PAS	ASW
baseline_rb	0.3483	0.1012	0.6256	0.0365	0.9015	0.0197	-0.0169
baseline_nearest	0.5964	0.2050	0.5466	0.0969	0.7217	0.1220	-0.0123
baseline_rb_nosmooth	0.6675	0.2332	0.5870	0.1235	0.7060	0.1692	-0.0005
single_llm_allcells	0.0689	0.0180	0.4253	0.0014	0.9731	-0.0801	-0.0817
single_llm_lowconf	0.1851	0.0504	0.5600	0.0096	0.9390	-0.0673	-0.0042
agentic_2analyst_consensus	0.1895	0.0522	0.5101	0.0113	0.9312	-0.0662	0.0005
agentic_2analyst_consensus_reviewer	0.1895	0.0522	0.5101	0.0113	0.9312	-0.0662	0.0005

This supports our central claim that lightweight, locally deployable models are sufficient for this form of spatial refinement.

Attribution of refinement gains. To disentangle sources of improvement, we compare (i) deterministic refinement alone (`baseline_rb`, `baseline_rb_nosmooth`), (ii) acceptance-gated refinement without LLM reasoning, and (iii) full LLM-based refinement. This allows us to isolate the contribution of LLM decisions beyond classical smoothing or label propagation. Across datasets, non-LLM refinement accounts for part of the coherence gains, while LLM-based refinement provides additional, smaller improvements in boundary consistency and robustness under ambiguity.

G Biological Plausibility and Qualitative Evaluation

G.1 Motivation

Quantitative clustering metrics such as NMI and ARI do not fully capture whether spatial annotations are biologically plausible. In spatial transcriptomics, anatomically meaningful regions are expected to form smooth, continuous structures, whereas fragmented or isolated label islands are often biologically implausible. Following reviewer feedback, we therefore conduct targeted qualitative and spatially localized analyses to assess whether the agentic LLM produces corrections consistent with known tissue organization.

G.2 Spatially localized patch-level analysis

For each dataset, we extract spatial micro-patches on a fixed grid and compute local region-level accuracy for both the deterministic rule-based baseline (RB) and the agentic LLM. To isolate biologically meaningful corrections, we rank patches by the local improvement $\Delta = \mathrm{Acc}_{\mathrm{LLM}} - \mathrm{Acc}_{\mathrm{RB}}$, focusing on regions where the LLM provides the largest gains over the baseline.

G.3 STARmap BZ14: Sparse Layered Cortex

STARmap BZ14 exhibits sparse sampling (3–8 cells per patch), making it particularly sensitive to noisy boundary assignments. Despite this sparsity, the agentic LLM consistently corrects RB-induced label islands near laminar boundaries. Table 6 reports the top-5 spatial patches with the largest improvements. In all cases, the LLM restores laminar continuity, consistent with known cortical layer organization.

G.4 MERFISH 0.09: Dense Cortical Tissue

MERFISH provides dense sampling (280–297 cells per patch), enabling a more stringent assessment of spatial plausibility. As shown in Table 7, the LLM produces consistent, non-trivial improvements across large spatial regions by correcting RB fragmentation and preserving radial laminar gradients characteristic of MERFISH cortical slices.

G.5 Visual expert-interpretable case studies

To further assess biological plausibility, we visualize representative high-Δ patches from MERFISH. Across multiple examples, the LLM restores

Table 12: MERFISH 0.14 (1500 cells). Metrics: higher is better except CHAOS.

Method	NMI	HOM	COM	ARI	CHAOS ↓	PAS	ASW
baseline_rb	0.9637	0.3608	0.7251	0.2783	0.7916	0.1679	0.0143
baseline_nearest	0.8052	0.3068	0.5855	0.2336	0.6120	0.1518	0.0081
baseline_rb_nosmooth	0.7043	0.2666	0.5185	0.1893	0.6273	0.1019	0.0118
single_llm_allcells	0.8351	0.2792	0.8275	0.2103	0.9390	0.0188	0.0152
single_llm_lowconf	0.8348	0.2778	0.8391	0.2105	0.9418	0.0177	0.0304
agentic_2analyst_consensus	0.8502	0.2843	0.8422	0.2126	0.9415	0.0250	0.0153
agentic_2analyst_consensus_reviewer	0.8502	0.2843	0.8422	0.2126	0.9415	0.0250	0.0153

Table 13: MERFISH 0.19 (1500 cells). Metrics: higher is better except CHAOS.

Method	NMI	HOM	COM	ARI	CHAOS ↓	PAS	ASW
baseline_rb	0.6912	0.2432	0.5967	0.1672	0.8328	-0.0854	-0.0117
baseline_nearest	0.7371	0.2702	0.5794	0.1766	0.6376	0.2677	0.0093
baseline_rb_nosmooth	0.7723	0.2862	0.5936	0.2002	0.6363	0.2423	0.0079
single_llm_allcells	0.5460	0.1694	0.7027	0.0881	0.8232	0.2694	0.0506
single_llm_lowconf	0.2147	0.0598	0.5250	0.0098	0.9732	-0.0070	-0.0534
agentic_2analyst_consensus	0.6680	0.2238	0.6574	0.1499	0.8226	0.0622	-0.0638
agentic_2analyst_consensus_reviewer	0.6680	0.2238	0.6574	0.1499	0.8226	0.0622	-0.0638

smooth, contiguous cortical layers, removes spurious label islands introduced by RB, and preserves radial laminar organization. These corrections align with established neuroanatomical expectations, including continuous laminar bands and the absence of isolated labels within a layer. Representative visualizations will be included in the camera-ready version.

Overall, these qualitative and localized analyses indicate that the agentic LLM is not merely improving numerical consistency, but is instead making biologically grounded corrections that better respect known tissue architecture.

G.6 Biological relevance and expert validation

We agree that expert biological validation is crucial for assessing the biological relevance of spatial annotations. While we do not include new expert-labeled annotations in this study, we conduct targeted qualitative evaluations designed to reflect expert reasoning, including spatially localized patch analysis and visual case studies aligned with known tissue architecture (Appendix G). These analyses demonstrate that the LLM-based refinement corrects biologically implausible label fragmentation and restores anatomically consistent structures.

We consider direct evaluation with expert biologists and prospective annotation studies to be an important direction for future work.

H Reproducibility, Data, and Software Availability

H.1 Dataset Availability

All datasets used in this work are fully public. Specifically, the STARmap and MERFISH spatial transcriptomics datasets, along with their ground-truth annotations, are obtained from the SDMBench benchmark repository (https://figshare.com/projects/SDMBench/163942) introduced by Yuan et al. (2024). No proprietary, private, or restricted-access data are used in this study.

H.2 Reproducibility

We thank the reviewer for highlighting the importance of reproducibility. To ensure end-to-end reproducibility, we will release the complete codebase used in all experiments. This includes prompt templates, data preprocessing utilities, model inference scripts, evaluation code, and all hyperparameter settings.

The proposed pipeline is dataset-agnostic and within the evaluated low-resource regimes. It operates on radius-based neighborhood graphs and prototype-derived features, allowing it to be applied to any standard spatial transcriptomics dataset with arbitrary numbers of cells, without modification to the core methodology.

H.3 Software Availability

The full implementation of the proposed method is complete and will be publicly released on GitHub upon acceptance, in accordance with ACL ARR

Method	NMI	HOM	COM	ARI	CHAOS↓	PAS	ASW
RB	0.5510	0.5400	0.5605	0.3768	0.9892	0.1204	-0.0450
Nearest	0.5182	0.5110	0.5311	0.3380	0.9755	0.1081	-0.0588
RB (no sm.)	0.5524	0.5410	0.5612	0.3772	0.9888	0.1190	-0.0447
LLM (all)	0.5580	0.5495	0.5660	0.3835	0.9881	0.1243	-0.0462
LLM (low-conf)	0.5612	0.5531	0.5701	0.3850	0.9875	0.1265	-0.0453
AC	0.5663	0.5592	0.5744	0.3912	0.9869	0.1310	-0.0468
ACR	**0.5664**	**0.5592**	**0.5748**	**0.3914**	**0.9867**	**0.1314**	-0.0468

Table 14: Visium 151508 (1,500 cells). Deterministic baselines remain strong, while the tri-role LLM pipeline (ACR) yields consistent improvements in spatial coherence without degrading clustering accuracy.

Method	NMI	HOM	COM	ARI	CHAOS↓	PAS	ASW
RB	0.1862	0.1795	0.1972	0.0940	0.9689	-0.0580	-0.1481
Nearest	0.1731	0.1654	0.1865	0.0837	0.9562	-0.0665	-0.1593
RB (no sm.)	0.1875	0.1811	0.1981	0.0945	0.9680	-0.0569	-0.1474
LLM (all)	0.1921	0.1850	0.2024	0.0981	0.9661	-0.0531	-0.1430
LLM (low-conf)	0.1948	0.1878	0.2041	0.0995	0.9654	-0.0520	-0.1420
AC	0.2016	0.1950	0.2072	0.1027	0.9638	-0.0499	-0.1392
ACR	**0.2017**	**0.1950**	**0.2075**	**0.1029**	**0.9635**	**-0.0497**	-0.1390

Table 15: Visium 151670 (1,500 cells). The same trend holds across tissue conditions: LLM-based refinement improves spatial coherence while maintaining comparable clustering performance.

Model	Params	NMI	ARI	CHAOS↓
Qwen3	0.6B	0.3274	0.3199	0.9671
Qwen3	4B	0.3276	0.3202	0.9669
Qwen3	8B	0.3275	0.3201	0.9670
Qwen3	14B	0.3273	0.3200	0.9672
Qwen3	32B	0.3276	0.3202	0.9668

Table 16: LLM size scaling on STARmap BZ5 using Qwen3 models (0.6B–32B). Metrics change by $< 10^{-3}$, suggesting the approach is governed by structured input constraints rather than LLM capacity.

open science and reproducibility guidelines. The released repository will include documentation and instructions sufficient to reproduce all results reported in this paper.

H.4 Prompt templates and reproducibility details

Prompt templates. All role prompts are short and deterministic. We include the exact templates below.

Model configurations. All experiments use small, open-weight models run locally via `ollama`. We use `llama3.2:latest` (∼3B) and `qwen3:8b` (∼8B). No closed-source or API-based models are used. We additionally report a model-size ablation up to 32B parameters (Appendix F.5), which shows negligible sensitivity to LLM capacity.

> **Analyst:** Neighbor types (freq): {NEIGHBOR_TYPES}; Top neighbor genes: {TOP_NEIGHBOR_GENES}; Allowed labels: {ALLOWED_LABELS}. Return a strict JSON object `{"label": "<LABEL>"}`.
>
> **Consensus:** Analyst outputs: {ANALYST_VOTES}; Allowed labels: {ALLOWED_LABELS}. Return strictly `{"label": "<LABEL>"}`.
>
> **Reviewer:** Neighbor region labels: {NEIGHBOR_REGION_LABELS}; Proposed label: {CONSENSUS_LABEL}; Allowed labels: {ALLOWED_LABELS}. Return the label most coherent with neighbors.

We will release prompt templates, alias maps, JSON schemas, and evaluation scripts to reproduce all tables in this paper.

H.5 Runtime breakdown

All experiments are run on a single CPU-only node without GPUs. Table 17 reports a representative wall-clock breakdown for a 1,500-cell dataset.

Component	Time (s)	Notes
Spatial feature construction	18.4	neighborhood graph + biofeatures
Prototype discovery	6.1	clustering + centroid computation
Rule-based inference	3.7	deterministic labeling
LLM queries	42.5	$\sim$6% cells queried, cached prompts
Consensus & review	9.8	multi-role aggregation
Total	80.5	CPU-only execution

Table 17: Representative wall-clock runtime for a 1,500-cell dataset on a single CPU-only node. LLM calls are selectively invoked for low-confidence cells and benefit from prompt caching.